Andrea Bonaccorsi *Editor*

The Evaluation of Research in Social Sciences and Humanities

Lessons from the Italian Experience

Springer

The Evaluation of Research in Social Sciences and Humanities

Andrea Bonaccorsi
Editor

The Evaluation of Research in Social Sciences and Humanities

Lessons from the Italian Experience

Springer

Editor
Andrea Bonaccorsi
DESTEC
University of Pisa
Pisa, Italy

IRVAPP-FBK
Trento, Italy

ISBN 978-3-319-68553-3 ISBN 978-3-319-68554-0 (eBook)
https://doi.org/10.1007/978-3-319-68554-0

Library of Congress Control Number: 2017959886

© Springer International Publishing AG 2018
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

What Does the Italian Experience in Research Evaluation Tell to the International Debate on SSH?

The widespread diffusion of research evaluation practices is part of the new higher education and research landscape at all levels. While scholars active in those disciplines that are conventionally labelled science, technology, engineering and mathematics (STEM) find themselves relatively at ease with this reality, their colleagues in social sciences and humanities (SSH) raise a number of theoretical and practical issues.

Research evaluation is an activity that has two, not just one, sources of legitimation. On the one hand, the parliament and the government in many countries have created dedicated structures to carry out research evaluation and actively make use of their results. This is a legal and institutional form of legitimation. It comes from the legitimate democratic authority of modern states. On the other hand, however, scientific and academic communities create and manage a different source of legitimation, based on scientific recognition and expertise.

To borrow the influential theory of power in organizations developed in 1959 by French and Raven, there are five forms of power: legitimate, reward, coercive, expert and referent. The first three come from organizational sources, the latter two from personal sources. The evaluation may use legitimate power from the parliament and the government and may also have reward and, to a certain extent, coercive power if its results are used for the allocation of results, but its legitimation in the academic community is mostly based on expert and referent power—in other words, whether evaluation is perceived as based on sound principles of knowledge and whether it is respected and esteemed.

For this reason, it is mandatory for those involved in research evaluation to open a never-ending dialogue with the scientific communities, in order to gain legitimation from a bottom-up and trust-based process, not only from the (inevitably top down) institutional procedures. Communities in SSH raise serious problems that deserve respectful listening, close scrutiny, a lot of discussion and conversation, some cre-

ativity in finding solutions and persistence. The best way is to establish a dialogue on those issues that are close to the heart of the people: how to do better research.

This book addresses the difficult and controversial issue of evaluation of research in SSH with the support of experiences and data from the recent activities carried out in Italy.

Why is the case of Italy relevant for the debate on this issue?

First, this country has experienced perhaps the single largest and most comprehensive assessment process in large European countries, in a few years. After a long legislative process, started in 2006, the National Agency for the Evaluation of Universities and Research Institutes (ANVUR) was created in 2011. There was large recognition that the Italian system had been lagging behind the best European experiences in the field. The agency started immediately to implement several initiatives, on which the chapters in this book give large details. Here, it is important to recall the main features of these initiatives, in order to offer the overall framework.

- A research assessment exercise (VQR 2004–2010, or Valutazione della Qualità della Ricerca) was launched in 2011 and published in July 2013. *All* researchers submitted to the evaluation: researchers affiliated to universities had to submit three products, while those affiliated to public research organizations (PROs) were requested to submit six products. Research products were evaluated by 14 panels (GEV, or Gruppi di Esperti della Valutazione), composed by 450 members overall. Panels in STEM disciplines adopted bibliometric methods, with minor role for peer review, while panels in SSH exclusively used peer review (with the exception of economics and statistics). Overall, more than 14,000 referees were mobilized and over 180,000 products were evaluated. The results have been used by the ministry to allocate a share of the performance-based funding to universities.
- A second research assessment exercise has been launched in 2015 and completed by the end of 2016. Several improvements were introduced after the first VQR. As in the 2004–2010 exercise, individual researchers have received their own personal scores in a confidential way, while the scores of all disciplines (with at least four researchers per department) have been published in a highly disaggregated way.
- A major change in the legislation for academic recruitment and promotion was put in force in 2012. A National Scientific Habilitation for Associate and Full Professors (ASN, Abilitazione Scientifica Nazionale) system was created, with some similarities to the ones in place in Spain and France. The legislation introduced a system of quantitative indicators to be satisfied not only by candidates but also by full professors wishing to enter the national habilitation committees. This was a major departure from academic traditions. ANVUR was requested by the ministry to calculate the national distributions of indicators on scientific publications and to publish the median value as a threshold to be satisfied. In SSH sectors, given the absence or lack of validity of bibliometric data, three indicators

were selected: number of books, number of book chapters and articles and number of articles in A-rated journals.
- As a result of the implementation of ASN, ANVUR was requested to classify journals, in order to separate scientific from non-scientific journals, and to identify a set of A-class scientific journals. ANVUR asked learned societies to give an opinion and appointed an expert group. The starting point was offered by the self-administered database in which all Italian researchers report their publications. It turned out that there were more than 60,000 titles of journals. Out of these titles, the agency classified around 24,000 as scientific journals and 3000 as A-class journals. The lists were published in 2012 and revised periodically. By end 2017, a general revision is expected.
- In the context of quality assurance principles applied to teaching activities and to departments, a data collection exercise was undertaken, in which departments filled a database on their research activities, including metadata on publications. The resulting database (SUA-RD, Scheda Unica Annuale della Ricerca Dipartimentale, or University-level sheet on departmental research) has not yet been published at the time of this book. However, several authors in this book asked access to the SUA-RD documentation in order to compare data. This source will be important in particular in SSH fields, in which no external or bibliometric source is considered valid.
- Finally, in 2016, after a preparation process, the evaluation of the third mission of universities and PROs was undertaken, on the basis of a data collection process initiated experimentally in 2014. The process is based on the informed peer review methodology. In parallel to the VQR 2011–2014, the results have been made public in 2017.

From this list one can figure out how many issues and controversies have been addressed in a short time frame. In just a few years, universities and PROs, rectors and administrators, academicians and early researchers have been involved in completely new processes, arguments and metrics. They had to learn about citations, h-index and median values. A lot of controversies have been generated and are still alive. A huge data collection is now available, most of which in the public domain (although, due to resource limitations, mainly in Italian language). Several papers have been published in the last few years on this experience.

A second reason of interest is that SSH disciplines have a strong tradition in Italy, due to historical reasons. Entire branches of humanities, for example, from philology to history of art or archaeology, have a long tradition in this country, which originated from the Middle Ages and the Renaissance. Communities in SSH are lively, create discussions and take action.

Third, in the recognition that research on evaluation is a crucial element of the quality of evaluation, ANVUR started a major initiative aimed at funding, with small-scale grants, original research projects. This allowed scholars in a variety of fields to engage into theoretical or empirical studies that might contribute to the improvement of evaluation practices. A Call for Proposals was issued, and, following a competitive ex ante selection procedure, a series of projects were funded.

Some of the papers that originated from the call are included in this book and acknowledged directly in the text. The authors cover a large range of expertise, from computer science and computational linguistics to library sciences, from sociology to law, in addition to many branches of humanities. This is a remarkable enrichment of the community of those that study research evaluation in a systematic way.

The book is opened by a chapter in which **Andrea Bonaccorsi** calls for an epistemic approach to evaluation. By epistemic approach it means the systematic analysis of the ways in which communities in SSH produce *valid* knowledge. Evaluation must rest on the solid foundations of the criteria by which epistemic communities validate and value the knowledge they produce. There is a remarkable theoretical gap here.

Despite the work of sociologists such as Collins, Abbott, Ben-David, Becher and, more recently, Michèle Lamont, we lack a full-scale theory of the epistemic practices of scholars in SSH. By and large, most philosophy and sociology of science in the twentieth century have been dealing with hard sciences, not with SSH. We do have regional theories, particularly in social sciences (e.g. methodology of social sciences, philosophy of economics), but the linkage between epistemic issues and research evaluation is missing. The chapter offers a preliminary sketch of such a linkage and calls for further research.

Part I of the book deals with the controversial issue of criteria of research quality in SSH. Do scholars in SSH define research quality in a consistent way? Do they converge on common definitions?

The part suggests the idea that research quality criteria *do exist* in the practice of scientific communities in SSH, but they require a long process of conversation and elicitation to be formulated in linguistic terms, made reflexive and explicit and put in the public space. Scholars put these criteria in practice innumerable ways in evaluating papers and books submitted for publication, in making recruitment and promotion decisions and in the daily activity of reading others' works. These criteria are typically held in an implicit, non-declarative or tacit form. They form a crucial part of the *personal knowledge* of scholars, following the concept introduced by Michael Polanyi. The evaluation asks scientific communities to articulate these criteria, in order to make them declarative and subject to intersubjective discussion. Claims about what is valid and what is not valid in the knowledge produced in a community, and about the degrees of quality of valid knowledge, require rational justification. In this perspective, the evaluation is a kind of democratic process in which decisions are put under the light of public conversation and hence require justification.

The chapter by **Chiara Faggiolani** and **Giovanni Solimine** examines the recent Italian experience of research evaluation in the VQR 2004–2010, comparing it with the international experience. It is confirmed that the pattern of publication of scholars in SSH is structurally different from the one in STEM. The central role of books creates a difficult issue, because the only way to evaluate books is by reading them from cover to cover. They examine some of the most common solutions adopted in various countries, namely, the weighting of coefficients for monographs by type and size, the rating of publishers and series, the use of reviews in evaluation and Library Catalogue

Analysis (LCA). They do not recommend the uniform weighting of books, given the extreme heterogeneity of their type and content. Also, while publisher rating and the use of reviews might contribute to approximate elements of quality, the authors find that the evidence of validity and reliability of these methods are still not sufficiently demonstrated. Finally, they show caution with respect to LCA.

Andrea Capaccioni and **Giovanna Spina** examine international experiences in the drafting of guidelines for peer review in SSH. These initiatives are important in a variety of ways. First, they call for a continuous activity of self-reflection of communities on the way in which research quality criteria should be formulated linguistically. It is one thing to hold a notion of criteria, another thing to formulate it in verbal terms in such a way that other people, sharing the same notion at conceptual level, apply it in a consistent way. Second, peer review is a learning process for both those that carry it out and those who must accept it. The guidelines help to put in place solutions that improve the learning generated in the process, for example, in terms of training of referees involved in research assessment exercise or calibration of subjective judgements. Finally, there is a delicate issue of conflict of interest. It does not only refer to the problem of direct and personal conflict given by personal relations but to the more subtle issue of whether referees evaluate products for which they have little expertise. This problem can be managed, to a certain extent, by constructing databases of referees based on their past performance in peer review, an activity that the Italian Agency has implemented. However, there is room for self-constraint based on forms of ethical codes. Referees should not accept the evaluation of works for which they have no scientific legitimation. In the SSH field, this might be a controversial issue, given the extreme specialization and articulation of disciplines in subfields.

The chapter by **Andrea Bonaccorsi** reports about an initiative taken by the Italian Agency to open a dialogue with scientific communities in SSH about the improvement of the peer review process. A number of expert groups were formed, which were given full autonomy and freedom to formulate suggestions, in several SSH areas, from architecture to humanities and arts, from social and political sciences to law, to business studies. The outputs of these groups were transmitted to the evaluation groups under the new research assessment exercise (VQR 2011–2014) and were used to modify the evaluation sheets to be used in the peer review and to enlarge the range of admissible products. As an example, there was a large debate on the meaning of "internationalization" as a research quality criterion after the first research assessment exercise (VQR 2004–2010). One major concern was that referees could have interpreted this broad notion in a mechanistic way, by giving automatically higher scores to articles or books in English. The expert groups elaborated on the meaning of this notion for their respective disciplines and suggested solutions, most of which have been implemented in the new wording of evaluation. This is an example of the way in which engaging into conversation with the SSH communities may improve the validity of the research assessment. At the same time, it shows that communities should be challenged to formulate their own criteria in a declarative way, so that they can be implemented by other experts (in this case, the members of the evaluation groups at the national level).

The chapter by **Ginevra Peruginelli** and **Sebastiano Faro** enters into the troubled waters of research evaluation in legal studies, or law. This is an open issue at the international level, and several initiatives are under way to study the problems and formulate suggestions. In this field the adoption of national language is mandated by the object of research, which is mainly the positive law at the national level (with some exceptions, such as philosophy of law and history of law or, obviously, the various branches of international legal studies).

At the same time in these fields, there is a remarkable trend to incorporate judicial decisions produced by international courts, particularly at the European level, into the decision of national judges, using the original English formulation. This means that the epistemic process of production of valid knowledge in law will endogenously become more international in the near future. A second problem that has been hotly debated in the Italian context has been the legitimation of the agency to classify scientific journals. The initial position of learned societies was negative. A legal action was taken in 2012 by a prominent scientific society, based on the argument that only scientific communities, and learned societies on behalf of these communities, had the legitimation to produce research quality criteria. The administrative judge rejected the claim, arguing that there is no legally defined role for learned societies, while an agency created by the law, acting on the principles of transparency stated in the law, may directly refer to scientific communities in order to define research quality criteria. This may happen in a variety of ways, by appointing experts, consulting documents, examining data and combining different sources of judgement. As a matter of fact, the classification of journals has been carried out also in law. The chapter suggests that, after the initial conflict, there is now a reflexive and constructive attitude by the academic legal community.

Part II of the volume deals with the role of books and monographs in the evaluation of SSH research and the appropriate methods.

Geoffrey Williams, **Antonella Basso**, **Ioana Galleron** and **Tiziana Lippiello** first examine the treatment of books in research assessment in several European countries, characterized by largely different methodological approaches: qualitative systems (UK and Netherlands), quantitative systems (Czech Republic and Poland) and database systems (Norway and Flanders). In addition, they mention the Spanish experience of rating publishers. They submitted a questionnaire to a sample of Italian SSH scholars, using a version of an instrument first developed in France, and received 578 usable answers. Among the results, two are in my view particularly interesting. The first is that the respondents mention, among the communication channels they use for their research work beyond traditional academic channels, the Internet as the first option. This opens interesting directions for further research. The second finding is that the choice of communication channels by scholars seems to follow a well-defined strategy. On the one hand, junior researchers consider research articles much more than in the past due to their speed of publication but also because journals offer, differently from books, a formal peer review process. It does not seem that researchers choose journal articles because they are easier to publish, as some critics often argue, but quite the contrary, because they submit their results to a rigorous process of validation. On the other hand, respondents are clear

in answering that, in their departments, books are evaluated more than articles. My reading of their results is that a transition is in place in the way in which SSH scholars communicate their research results. This transition does not follow the path illustrated by the alarmistic wisdom according to which research evaluation will eliminate books and monographs, because researchers turn their attention exclusively to easy-to-write, easy-to-publish journal articles. Rather, a more complex and sophisticated strategy is emerging, mixing journal articles on refereed journals (particularly in the very early stages of career) and books (in a more mature stage). This academic strategy seems to be associated with the enlargement of communication channels towards the Internet.

Carla Basili and **Luca Lanzillo** offer a survey of guidelines for the evaluation of book and monographs, combining sources from international bodies (ISO standards), national experiences (Spain, Australia and New Zealand) and university-level manuals (University of Turin and University of Bologna).

The analysis shows a rich variety of objects covered under the notion of book and monograph, almost all of which tend to be included in research assessment exercises at the national level and university level. The authors observe a trend of inclusiveness of new categories over time. These are, almost universally, subject to peer review. This creates a tension, however, with the need to include books and monographs in simple counts of publications, based on administrative data and/or self-reported data. Existing classification systems, in fact, have a hard time distinguishing, to make just an example, between those edited books that are the result of complex, multi-year, interdisciplinary research projects and those edited books that are just a collection of disparate contributions. The tension between full reading books in peer review and just counting books is inevitable. The authors confirm the recommendation to make use of peer review.

The need to introduce methods for the evaluation of books without the need to read them from cover to cover was at the origin of the proposal of Library Catalogue Analysis, the object of the chapter by **Maria Teresa Biagetti**, **Antonella Iacono** and **Antonella Trombone**. The authors examine in great detail the internal working of digital catalogues in academic and nonacademic libraries and the evolution of standards and digital technologies. The evaluation question here is simple: can we rely on metrics based on the diffusion of books in qualified libraries at the international and national level, in order to build up a proxy of their quality? Are more diffused books, or books present in a qualified subset of libraries, also better books? The chapter examines several apparently technical issues that, however, are crucial for addressing this issue, such as the distinction between acquisition and donation of books, the commercial policies of publishers in the acquisition process or the distinction between the original edition, new edition and republishing of previous editions. According to the authors, existing digital catalogues fail to offer reliable data for evaluation purposes.

They also put under scrutiny new development in the digital technology applied to libraries, namely, Advanced Discovery Tools and Linked Data Technologies. The former are commercial tools offered to researchers for the fast and comprehensive search of metadata on publications of interest, using advanced data mining tech-

niques. The authors carry out a simulation, comparing the output of some Advanced Discovery Tool with first-hand information from library catalogues, and conclude that there are large gaps in information that prevent its use in the evaluation procedure. Linked Data Technologies are the answer to the challenge of opening library information systems to the new environment created by the web and open data.

Ginevra Peruginelli, **Sebastiano Faro** and **Tommaso Agnoloni** return to the field of legal studies with an original survey on Italian scholars. They obtained an unusually high rate of return to an online questionnaire (26% of the population, or 1241 usable responses), a result that indicates the interest (and controversies) raised by research evaluation in the legal disciplines. 809 respondents, or 76.4% of the total, are members of editorial boards of journals, and 796, or 75.5% of total, declare they are engaged in ex ante evaluation of journal articles.

With respect to journals, an interesting finding is that a minority of respondents believe peer review is carried out systematically (23%). The bulk of respondents (55%) believe it is carried out only in most cases, while 22% believe it is not carried out at all. This finding seems to be in contrast with the large diffusion (76,4% of the total) of editorial roles in the legal community. This discrepance can be explained by the organization of the ex ante process in most Italian journals in law. Peer review is not systematically managed with the double-blind, or single-blind, approach, by recruiting a large pool of anonymous referees and establishing formal procedures. Rather, it is directly managed by the editor or the editorial board. The presence of almost all scholars in at least one of the journals, demonstrated by the data, is a way to protect oneself, and one's students and collaborators, against the uncertainties of a fully competitive process. The chapter calls for more transparency in the description of the peer review procedures followed by legal journals.

A second interesting finding is that, once asked which research quality criteria they mention with respect to books, there is a consistent ranking across legal disciplines in which the top criteria include the inclusion in a particular series with special characteristics and the publisher. In other words, respondents admit that they use information on the book series and the publisher as an indicator of quality. Yet, they vigorously reject the notion of rating or ranking series and publishers and probably, as it is demonstrated in other chapters, rightly so.

Part III deals with one common methodology used in the evaluation of SSH research, namely, the rating of journals. As it is well known, it is a largely adopted methodology but also one subject to some controversy, as the experiences of Australia and France show.

In her chapter, **Domenica Fioredistella Iezzi** investigates the notion of "publication strategy", by examining the mix of types of publications submitted to the research assessment exercise (VQR 2004–2010) and the resulting average scores. It turns out that large heterogeneity can be detected. To summarize the argument, there are two broad areas (Area 10, or antiquities, philology, literary studies and art history, and Area 11, or history, philosophy, pedagogy and psychology), in which there is little difference in the average score across the main categories of products (journal article, monograph, edited book, book chapter and proceedings). The scores are in the range 0.63–0.72 in Area 10 and 0.55–0.64 in Area 11. In these fields, it

seems that scholars may submit whatever product they have produced and receive similar scores. Also, monographs receive the highest score (0.72 in Area 10 and 0.64 in Area 11). Furthermore, in these areas proceedings are highly valued (0.66 and 0.59 average score, respectively). The remaining three fields share a common negative evaluation of proceedings: their score is as low as 0.38 in social and political sciences, 0.26 in law and a mere 0.07 in economics and business. In these fields, paper presentation to conferences is not considered equivalent to articles or book chapters. They differ, however, under other respects. In the field of law, chapters, monographs and journal articles are valued high (range 0.56–0.63), while edited books are penalized (0.33). It seems that the editorial work of collecting chapters is not considered scientifically valid. Edited books are more valued in political and social sciences (0.51), while they are extremely penalized in economics and statistics (0.04). The area of economics and statistics is also interesting because the scores assigned to monographs, all evaluated in peer review, are extremely low (0.11 on average). In this area, the only product that is highly valued is the journal article (0.51). From these data, it is clear that there is a transition in the pattern of scientific communication and that scholars should adopt a well-defined strategy. The chapter also investigates the role of language and finds that, particularly in humanities and arts (Area 10), there is a large role for products in languages other than the national one. Interestingly, these products are not in English but in (many) other non-national languages.

The paper by **Andrea Bonaccorsi**, **Tindaro Cicero**, **Antonio Ferrara** and **Marco Malgarini** addresses the contested issue of journal rating. In the SSH field, there is almost a universal agreement on the impossibility to use indexed journals as the basis upon which to build indicators, since they represent a tiny fraction of journals used. Therefore, all journal rating exercises make use of expert-based assessments or combine quantitative indicators with expert judgements. After an extensive review of international experiences and of the academic literature, the chapter addresses the issue of whether the classification of journals (or the container) is a good predictor of the evaluation of individual articles published in journals (or the content). The Italian context offers an unprecedented opportunity to examine this issue on a large scale, due to the presence of two (relatively) independent evaluation processes, one carried out on individual articles (VQR 2004–2010) by thousands of anonymous referees and another carried out on journals, by a small-size expert group, which used also evaluations from learned disciplinary societies. While there might be debate about the degree of independence between the two measures (in particular, a small subset of journals was rated within the VQR exercise and communicated to referees, following the notion of *informed peer review*), it is true that there is no evidence that referees did not read the articles (the content) and followed the suggested classification of the journal (the container). In this case the variables would clearly be non-independent. Therefore, a classical multiple-regression approach seems to be justified. The authors find strong support for the validity of journal rating in predicting the average quality (score) of individual articles, although there remains (and should remain) variability around the predicted values.

Part IV examines a range of frontier problems in the evaluation of SSH research: new indicators based on Google Scholar (GS), library holdings and the issue of impact and/or third mission of SSH scholars.

There is a large literature comparing indicators from Google Scholars (GS) to traditional indicators based on Scopus and Web of Science. These studies invariably deal with STEM disciplines. The chapters in this book are highly original because they deal with Google Scholar indicators in SSH and compare these data with information derived independently from research assessment exercises.

Alfio Ferrara, **Stefano Montanelli** and **Stefano Verzillo** have developed a novel methodology for the large-scale disambiguation of GS data. They utilize the large database created within the National Scientific Habilitation procedure, opened in 2012. In this procedure candidates submitted their lists of publications and published them on the official ministry website. Since candidates declared which scientific discipline they were applying for, it is possible to establish a systematic linkage between the sets of words appearing on the metadata of their publication lists (e.g. words in titles) and their disciplines. The authors have applied state-of-the-art computational linguistics tools to train the algorithms for searching GS data, obtaining excellent results in terms of recall and precision. The authors conclude that GS can indeed be used within evaluation activities. It is important to underline that this approach does *not* require that authors build up their own profile on GS: the algorithms were able to identify precisely whatever author was in the list of Italian scholars *without* using profiles. I believe this methodology might be extended to other disciplines and other countries and open a new research direction for the use of GS as a complement to, or alternative to, other bibliometric sources.

In a companion chapter, **Ferruccio Biolcati-Rinaldi**, **Francesco Molteni** and **Silvia Salini** make use of Ferrara et al.'s methodology, and of the related dataset, to investigate the publication behaviour of scholars in social sciences. They compare the data from GS to data on publications from a newly created archive of publications of Italian scholars, called SUA-RD, which is yet unpublished. This archive comes from self-declared publications collected at the departmental level by ANVUR. They confirm that GS data can be used for evaluation purposes, although self-declared publications are a much larger number. At the same time, they show that many authors do not receive any citation, not only from indexed journals (which might be a result of the adoption of national language) but also in GS. Since Google Scholar reflects the largest available collection of documents, it is noteworthy that part of the academic community shows such a small impact.

Maria Teresa Biagetti, Antonella Iacono and **Antonella Trombone** address the issue of whether the diffusion of books in libraries can be used as a valid and reliable information for evaluation purposes. They run a simulation exercise, by comparing the data obtained by consulting digital catalogues of a sample of high-status libraries. They find large variability across catalogues, a result that casts doubts about the possibility to normalize this kind of information. Summing up, they do not recommend data on the diffusion of books in library holdings as a viable alternative to peer review or a useful complement to it.

These three chapters give us interesting recommendations for the use of new tools in evaluation: green light to Google Scholar, but only conditional on the adoption of a rigorous disambiguation system, and red light to Library Catalogue Analysis, until large improvements are introduced in the treatment of some critical issues.

The last two chapters of this part deal with a novel dimension of research evaluation in SSH, that is, the impact of research or the third mission.

Luca Lanzillo offers an articulated discussion about the concept of social impact assessment, in the light of the orientations of the European Commission and the experience of the REF in the UK. He defends the view that scholars in SSH do have a large impact with their research, although the definition and measurement of this impact require much research and implementation work.

Brigida Blasi, **Sandra Romagnosi** and **Andrea Bonaccorsi** report about the first effort to evaluate the third mission of universities and PROs undertaken by ANVUR, after a preparation process. In particular they examine the differences between STEM and SSH scholars in the activities labelled under the Public Engagement heading, a large and comprehensive category created in order to take into account the variety of interaction modes between research and society. At the time of closing this volume, the data on third mission are still under evaluation. The chapter offers a very preliminary snapshot of data, in the form of a frequency distribution of activities carried out by scholars. It turns out that scholars in SSH are no less active in public engagement than their colleagues in STEM.

Taken together, these last chapters call for a rigorous discussion of the notion of impact of SSH research. On the one hand, as the authors remark clearly, SSH researchers are very active at the interface with society and carry out a variety of initiatives. They are less visible because their products are more intangible and immaterial than those produced by STEM research (e.g. patents, spinoff companies, science and technology parks, etc.). They are also less visible simply because the notion of third mission has been initially considered equivalent to the valourization and commercialization of research, which is only part of it. There is a need to place the impact and third mission of SSH higher in the agenda, for both SSH researchers and policymakers and the public opinion.

On the other hand, we should avoid the trap of imitating STEM research in trying to demonstrate the economic, or short-term, impact of SSH research. This approach would never deliver the kind of immediate, tangible, measurable results that are often the request of policymakers. Rather, it is important to build up a full-scale and compelling argument, in which the epistemic dimensions are associated to ethical and political arguments, in order to show the constitutive role of SSH research for the cohesion and advancement of contemporary societies. In studying languages and texts, in preserving the material, iconic and symbolic memory of civilizations, in supporting the creation and interpretation of the rules by which people regulate their rights and duties, produce and exchange goods and services and make collective decisions, SSH research builds up and maintains “the cement of society”, to use the celebrated expression of Jon Elster. On the basis of such an argument, it will be possible to spell out the relevant dimensions of evaluation for SSH communities.

The book is closed by a contribution by a non-Italian scholar, a leading expert in the evaluation of research in SSH, **Alesia Zuccala**, from the Royal School of Library and Information Science of the University of Copenhagen. This is an important contribution to the book, which would otherwise be subject to self-referentiality. The author offers first a fascinating theoretical discussion of the role of language and of national and cultural specificities in the SSH fields. She then surveys the chapters of the book in an effort to examine the contributions they give to the international debate on research evaluation, either at the conceptual level (epistemic issues, classifications, research quality criteria, impact and third mission) or the methodological and technical level (peer review, LCA, journal rating, publisher rating, book reviews, Google Scholar). The position taken by the author is that the Italian experience stands in front of the elder and more developed Anglo-Saxon approach: "The plural evaluation culture that I describe in this chapter – a European-Anglo culture at this point – possesses greater opportunities for uniting when there is a subtle respect for differences, rather than uniting due to forced ideals". Plurality of evaluation cultures is, perhaps, the name of the game for the future of research evaluation.

I thank the authors of the chapters for long discussions on the content and on the overall project. ANVUR not only provided the financial support to some of the projects whose results are reported in chapters but also authorized the publication of this book and, most importantly, ensured complete intellectual freedom to all researchers involved.

At the end of the day, the legitimation of research evaluation will be achieved when people will recognize that it is an integral part of the academic profession. We evaluate and we are evaluated. We see and we are seen, all the time. And since working in the academy is perhaps the most rewarding profession in the world, one might also expect that evaluating oneself and the others is a source of professional satisfaction, while being evaluated (yes, sometimes negatively) is part of life.

In a poem that all Italian children learn at school, *The solitary bird*, the great poet Giacomo Leopardi offers a vivid picture of the joyful company of young people in the village: "Dressed for the festival/ young people here/ leave the houses, fill the streets, /to see and be seen, with happy hearts" (tr. A.S. Kline). *To see and be seen*, this is part of the happy hearts. But the poet himself does not share the joy, he stays solitary: "I go out, alone,/ into the distant country,/ postpone all delight and joy/ to some other day".

The solitary attitude of the poet is certainly more noble and more inspired than the common attitude of the young company in the village. To see and to be seen is a social activity, and like all social activities, it can be considered by someone with a certain detachment. Authoritative scholars in SSH are perhaps right in pointing to the inevitable lack of depth of evaluation, when compared with the uniqueness, richness and erudition of the knowledge produced.

This book is an invitation for all who stay "in the distant country" to join, at least in part, the company of those who accept that "to see and be seen" is part of life.

Pisa, Italy — Andrea Bonaccorsi
November 2016

Acknowledgment

The book has been realized under the auspices and with the financial support of the Italian Agency for the Evaluation of Universities and Research institutes (ANVUR).

Contents

Towards an Epistemic Approach to Evaluation in SSH

Andrea Bonaccorsi

1 Introduction

The reaction of scientific communities against evaluation of research is almost entirely concentrated in SSH disciplines. STEM disciplines seem to have accepted that research evaluation, as it is often stated, "is here to stay". Some exceptions are sometimes raised in Mathematics or Clinical medicine, in particular against some of the practices in bibliometrics (e.g., Impact Factor), but overall these arguments constitute a minority opinion.

Why is this the case? I suggest that the reasons cannot be purely sociological, or related to the way in which scientific communities organise their work and communicate their result. Nor can it be political or ideological: the evidence that political or ideological opinions significantly differ by discipline is scattered and not robust. Something different, or deeper, must be at work. Since scholars are motivated more by the intrinsic logic of their scientific work than by external incentives (although incentives matter a lot), the explanation must be found at the epistemic level.

By "epistemic" I mean the way in which scientific communities produce valid knowledge, or the procedures, criteria, practices by which they recognise inter-subjectively the value and validity of the knowledge produced by others, and by this way submit themselves to the same rules. In this perspective, the inter-subjective dynamics of communication and validation are not separated from the internal dynamics of knowledge, or the intrinsic persuasiveness of the knowledge exchanged (Ziman 1978, 2000).

In this sense epistemic is not the same as epistemological, since the latter requires a second-level abstract and professional reflection on the rules of scientific work.

A. Bonaccorsi (✉)
DESTEC, University of Pisa, Pisa, Italy

IRVAPP-FBK, Trento, Italy
e-mail: a.bonaccorsi@gmail.com

© Springer International Publishing AG 2018

A. Bonaccorsi (ed.), *The Evaluation of Research in Social Sciences and Humanities*, https://doi.org/10.1007/978-3-319-68554-0_1

Not all scientists are also philosophers of science (indeed, only very few), but all good scientists have a solid mastery of a series of rules that are used to discriminate knowledge claims according to their purported validity.

Nor is it the same as sociological, since at this level the main interest is the way in which communities build up their agreement (or disagreement), irrespective of the content of knowledge. Sociological studies of science are mainly interested in the way in which socially defined actors, like scientists, set the boundaries of scientific vs non-scientific knowledge (Gieryn 1983, 1995, 1999; Taylor 1996), define scientific disciplines (Lenoir 1997; Abbott 2001), reach agreement or disagreement about claims (Knorr Cetina 1999), use material infrastructure and laboratory facilities to build up shared meanings (Latour and Woolgar 1979), create the conditions for repeatability of experiments (Collins 1975, 1985, 1999) or balance scientific power relations (Frickel and Moore 2006). For programmatic reasons, sociological studies do not deal directly with the epistemic content of knowledge, as separated (or separable) from the social interactions associated with it (Barnes and Edge 1982; Mulkay 1991; Barnes et al. 1996; Yearley 2005).

While I will use materials from epistemology as well as sociology of science, the main focus will be on the epistemic level, as elaborated by authors such as Ziman (1978, 2000).

In this chapter I address the following questions:

- What are the epistemic differences between STEM and SSH that may explain the differences in the orientation towards research evaluation?
- Are there epistemic differences across disciplines in SSH that may explain intra-SSH differences in the orientation towards research evaluation?
- Are there research quality criteria on which communities in SSH may converge? Or, is it possible to address epistemic differences with procedural fairness?

2 Epistemic Differences Between STEM and SSH

An influential stream of literature, inspired by logical positivism, argued that the difference between STEM and SSH is very simple: the former are scientific disciplines, the latter are not (Steinmetz 2005).

By scientific discipline meant a body of knowledge that could, at least in principle, formulate causal propositions. The formulation of causal relations requires a number of conditions that are found in natural sciences, but not fully in social sciences, even less in disciplines that deal with language. In natural sciences it is possible to assume the invariance of the object, so that controlled experiments can be carried out.

This view has dominated the scientific literature for decades after the Second World War. It is still maintained by some authors.

However, it is no longer assumed as the dominant theory, particularly after the developments of philosophy of science and social studies of science in the 1960s. The impact of Kuhn (1962) has been crucial here: the reason why scientists may formulate causal propositions is not that they control *each* of them in isolation,

but because these propositions are consistent with an overall paradigm, whose foundations do not have the same level of controllability. In addition, scientists produce a large variety of propositions, not only causal ones, referring to their experimental apparatus, the concrete rules of operation in the laboratory setting, or the practices of exchanging results.

Post-positivistic accounts of modern science admit a larger range of propositions as scientifically valid. This opens the way for asking to what extent disciplines in SSH may be defined scientific as well.

It is possible to summarise this issue separately for Social Sciences and Humanities.

In Social Sciences the issue of scientific validity of propositions has a long history, starting with the foundations of classical political economy and sociology in the eighteenth and nineteenth centuries. In the thought of classical authors such as Weber and Durkheim, knowledge produced in Social Sciences may well be defined as scientific, but not in the sense of producing invariant causal propositions (explanation) but rather propositions that make the behaviours of social actors intelligible by referring to their motivations (interpretation). Social Sciences are no less scientific than natural sciences, to the extent to which they submit their propositions to the same kind of rigorous control, but not through the use of experiments (which cannot be done by definition) but by establishing some level of stability of the relation between reasons for action (motivation) and observed action.

To what extent these disciplines can be defined "scientific" and what are the differences with respect to STEM? In the following I reject the notion that SSH disciplines are not scientific and investigate rather which epistemic differences can be identified.

First, researchers in STEM aim at discoveries, while researchers in SSH have only occasional discoveries (a new archaeological site, document, text, manuscript…) but most often aim at new interpretations of existing texts. The focus on discoveries means that scientists are in competition amongst themselves. Science is competitive because researchers fight to be the first to publish discoveries and receive the credit.

Second, research in STEM is cumulative, because scientists build upon the contributions of others, either in the past or from current competition. Science is a collective undertaking, not an individual enterprise. Science is both competitive and collaborative at the same time. There is a sharp difference here with respect to SSH, in which cumulativeness is much lower (Walliser 2009a, b).

Third, because of competition for discoveries and cumulativeness, the appropriate communication channel is the journal article (Lindsay 1978; Bazerman 1988; Cronin 1984, 2005). The scientific journal is serial or periodical, it offers researchers all over the world the opportunity to be updated regularly on discoveries, the format of the article is suitable for communicating discoveries, and the peer review system is efficient in solving issues of information asymmetries on the attribution of priority. The scientific journal system follows the competitive structure of science (Dasgupta and David 1994). Over time, the competitive dynamics generate a hierarchical system based on a cumulative process of reputation building: scientific journals that have published important discoveries are read more frequently; consequently authors

compete to be published in them; the increase in the number of submissions makes it possible to raise the rejection rate, making the quality of journals even higher and attracting more readers, and so on. It is this structure of scientific activity that makes it possible to build quantitative measures of research quality. In particular, once the role of citations is clarified in an unambiguous way, and the set of journals for which scientists compete is sufficiently large, then the very competitive dynamics generate a system in which the underlying quality is reflected in the relative measures of citations applied to papers, authors, institutions, while the impact factor of journals is considered a reliable measure of their average quality. Therefore there is a strong connection between the nature of scientific activity in discovery-driven fields, the overall system of academic publishing, and the reliability of quantitative measures of research quality based on citations.

On the contrary, the journal article is not the suitable medium for SSH, because new interpretations require long explanations best suited for the book format (Baldi 1998; Brooks 1985, 1986).

Fourth, because of competition for priority, cumulativeness, and the workings of the scientific journal system, citations are an essential element of scientific communication. In STEM citations have unambiguous meaning of credit assigned to authors that made the previous discoveries. As it has been originally discussed by Merton, Garfield and De Solla Price, and more recently formalised by Dasgupta and David, citations to the previous literature are a necessary ingredient of scientific publishing (Bornmann and Daniel 2008). This necessity is neither ethical nor practical, it is functional. By functional is meant, according to Merton's sociological approach, that individuals are forced to use a citation system that complies with the collective rules of the scientific community, irrespective of the individual willingness.

In order to be credited for a discovery, the authors must demonstrate their contribution is new with respect to the state of the art. In the absence of citations, it would be on the shoulders of the readers to check carefully whether there is anything new, clearly a very inefficient solution. Thus the overall system of scientific journals is based on referees who directly check the credibility of the authors' statements, acting as agents on behalf of the scientific community. In doing so, they force authors to list all relevant citations. Furthermore, due to the cumulativeness of scientific discoveries there is no need to cite authors from the distant past, but only the papers published in the last few years, which include all the relevant knowledge. This is a striking feature of scientific papers: only a few scientific authorities of the past are cited, not because they are ignored, but because their contribution is embedded in the citations of more recent authors.

In SSH, on the contrary, researchers quote authors from a distant past, very often classical authors, and produce works that are not cumulative but complementary, segmented or even alternative to each other. While the segmentation into scientific fields and sub-fields is largely agreed in STEM, and is usually not the outcome of individual decisions, in SSH part of the activity of most creative authors is the definition of new fields or new segmentations. The existence of progress, i.e., that some works are not worth being cited because their contribution has been subsumed

into others' contributions, is usually recognized very late, often after the authors cease their activity or die. Consequently, there is a need for a different theory of citations in SSH. Citations serve different purposes and should be classified accordingly.

Finally, there is a different role of paradigmatic pluralism. In STEM there is most often a dominant paradigm, sometimes with one or a few minority positions. Due to the cumulative nature of science and limited paradigmatic diversity, competition is open. On the one hand, within disciplinary boundaries all researchers compete fiercely for discoveries, without internal segmentations that may protect against competitors. On the other hand, since peer review is (generally speaking) a blind process, the past reputation and academic status of authors are not relevant to the probability of being published. This means that incumbents, or people with a recognised academic status, do not enjoy monopolistic positions in the long run. New entrants like junior researchers and authors with unorthodox views are easily recognised. Under these conditions, it is not possible for a single author or group of authors to monopolise the citations or to manipulate the reputational indicators.

This is not the case in SSH, in which paradigmatic pluralism is not the exception but the rule. On the one hand, there are internal segmentations that are not due to disciplinary differences but rather to paradigmatic options (rooted in the choice of object, methodologies and techniques) but also to value-laden positions (academic schools and traditions, ideological positions, political affiliations and attitudes, cultural orientations).

Competition is not completely open but segmented. Scholars have sometimes a two-layered choice: first, with which paradigm they want to be affiliated; second, how to compete within the paradigm chosen. In some important sense, there is competition among paradigms, but each of them is organised into its own scientific and academic structure (often with dedicated journals, conferences, scientific societies). Competition within the paradigm is not open but controlled by the leaders who contributed its creation. The relationship between paradigms is a matter of academic power, or maybe of paradigmatic change in the long run. On the other hand, in SSH peer review is not universally adopted. The identity of authors is generally known by those who make editorial decisions in journals and book series. Since books are the most important source, the control of editorial decisions is more easily controlled than in journals. This makes competition among researchers even more restricted.

As it appears from the above discussion, there are clear counterparts of this situation in the field of industrial organisation in economics. The kind of competition experienced in science is similar to the situation of competitive markets, in which entry is open, incumbents never get a monopoly position, and it is not possible for an incumbent to manipulate strategic variables to its own advantage.[1] This is why, in my

[1] That is why I find most arguments about the possibility of manipulation of bibliometric information, such as the Impact factor, rather pointless. Science itself is manipulable. There are many examples of fake discoveries or misbehaviour of scientists. The truly interesting question is not why these things happen, but why they happen so infrequently and how it happens that they are almost invariably discovered and punished.

opinion, scientists are not Foucauldian (see below). They find that the representation of commensuration as a form of hidden power is not appropriate for the way in which science works in their fields. It is not a matter of lack of reflexiveness, or pragmatic orientation, as opposed to the kind of critical work advocated in social sciences. Even scientists acutely aware of the social implications of their activity never subscribe to a Foucauldian argument. Simply put, competition is so harsh and the rate of knowledge production so overwhelming that no power coalition is stable.

This is not necessarily the case in SSH, where the competition is more of a monopolistic type, or even collusive oligopolistic. In other words, due to the fragmentation of disciplines and paradigmatic pluralism, the possibility of controlling a discipline for long periods is not negligible.

3 Epistemic Differences Within SSH

Yet this picture is still incomplete. On the one hand there are disciplines in SSH that have historically emulated STEM disciplines. On the other hand there are internal differences within SSH that also have implications on the orientation towards research evaluation. Thus we are faced with the challenge to examine differences *within* SSH disciplines.

In recent years the methodological foundations of social sciences (Sayer 1992; King et al. 1994; Goertz 2006; Moses and Knutsen 2007; Della Porta and Keating 2008; Brady and Collier 2010; Goertz and Mahoney 2012) and the position of social sciences with respect to general issues raised in the philosophy of science (Sayer 2000; Delanty 2005; Delanty and Strydom 2003; Benton and Craib 2011; Steele and Guala 2011) have been investigated thoroughly. A few cross studies (Steinmetz 2005; Walliser 2009a, b; Camic et al. 2011) have examined the differences across disciplines, while some other studies deal with the impact of social sciences in society (Flyvbjerg 2001; Brewer 2013; Bastow et al. 2014).

In parallel, a similar process started in Humanities, though somewhat less articulated, and partly as a response to the academic decline of these disciplines (Kernan 1997; Bate 2011; Belfiore and Upchurch 2013; Small 2013; Brooks 2014). Here a few historical comparative studies are also available (Bod 2013).

From this methodological and comparative literature, associated with related disciplinary studies, I have obtained a clear picture of the main epistemic problems addressed by various disciplines in SSH. In a recently published book (Bonaccorsi 2015) I presented a quite detailed reconstruction of the epistemic debate on four disciplines in SSH (history, political science, anthropology and English literature), combining historical material on the process of institutionalisation of the discipline in the academic system with an analysis of the main theoretical and methodological controversies. I strongly believe that this is a promising direction for research. Comparative studies that combine epistemic issues with institutional details will illuminate the way in which valid scientific knowledge is created. By taking disci-

plines as object of analysis I recognise that there are also internal distinctions within disciplines (Becher 1989; Abbott 2001) and try to take them into account.

This approach is not only useful to address the controversial issue of evaluation. It is my contention that entering into the epistemic black box of disciplines in SSH is also the only way to build up rigorous arguments to defend them vis-à-vis other disciplines, funding agencies and policy makers. There is a need to build up an argument about the *scientific* nature of SSH, based on a thorough recognition of the way in which they build up valid knowledge, though with epistemic processes that are completely different from the ones used in STEM. The damage to SSH generated by the wave of theorising that has suggested that they are just another way of producing texts instead of a truly scientific endeavour is currently underestimated. It is not enough to underline the pragmatic value of SSH in society. What is needed is a demonstration of the *intrinsic* validity of the knowledge produced by SSH scholars.

In this section I will sketch the main results of the detailed analysis carried out in Bonaccorsi (2015) and add other prominent disciplines in SSH. I review the four disciplines discussed at length in the book (history, political science, anthropology and English literature) and add other large disciplines in Humanities (philology, art history, psychology) and Social sciences (economics). The discussion below will be very concise. Interested readers are referred to references quoted in Table 1 and to the extended discussion and long reference list in the book.

I suggest that the orientation towards research evaluation is a function of four constructs, which combine historical factors with epistemic dimensions:

(a) History of the academic institutionalisation of the discipline
(b) Main methodological orientation
(c) Position with respect to neo-positivism after Second World War
(d) Position with respect to post-structuralism in the 1960s and 1970s.

Let me explain the building blocks of the model. By academic institutionalisation I mean the way in which a discipline comes to be separate from others, receive an academic label, is taught at universities in a separate way, academic positions are created and hence learned societies are formed. The institutionalisation process may be very long, taking decades (Becher 1989; Abbott 2001; Hyland 2012). A discriminant factor is whether a discipline is recognised from the beginning or is separated from previously existing disciplines. In the former case, newly created disciplines maintain the "memory" of their institutionalisation by keeping existing disciplines at a distance. They even challenge existing disciplines, either methodologically or substantively.

Thus for example English literature in the US academic system is the outcome of a separation process which took most of the nineteenth century to be completed. Literary studies were initially compressed between philology, which was the dominant discipline in US universities that followed the German educational model, and low level literature reading courses (Baldick 1983; Court 1992). The institutional separation took place in the early twentieth century and was associated with a deep

Table 1 Epistemic differences across disciplines in SSH

Discipline	Institutionalization process	Methodological orientation	Position with respect to neopositivism	Position with respect to post-structuralism	Selected references
Economics	Historical origins in Moral philosophy First re-definition with marginalist revolution (XIX century) Second re-definition in XX century with the creation of the *Econometrica* journal (adoption of formal modelling in mathematical language)	Mathematical language Axiomatic foundation of behavioral assumptions Law-like generalizations (nomothetic) Non-mainstream economics committed to modelling	Practical acceptance Axiomatic foundations linked to high level debate in logic and philosophy Empirical issues regarding behavioral aspects addressed via the "as-if" approach (Friedman) Criteria for validity of causal propositions observed via econometric techniques	No impact Use of mathematical language shielded economics from the impact of post-structuralism, mostly based on the critique of texts	Lerner and Lasswell (1951) Friedman (1952) Bell (1982) Hausman (1992) Blaug (1992) Guala (2006) Walliser (2009a, b) Szenberg and Ramrattan (2014)

Psychology	Historical origins in Medicine and/or Philosophy Initial re-definition in late XIX century through Physiology and Anatomy Disciplinary autonomy gained through experimental protocols Definition of experimental protocols in the US context after Second World War	Experimental and quasi-experimental approach Contextual generalizations (nomothetic)	Practical acceptance of the regularity principle Complex psychological constructs built up by aggregation of more observable variables Laboratory setting	No impact Dissolution of subject (Derrida, Foucault) not accepted Strong distance from Psychoanalysis helped to maintain disciplinary integrity Acceptance only in minority posititions (mainly in Psychotherapy)	Hilgard (1987) Danziger (1990) Ross (1991) Platt (1996) Haskell (2000)
History	Initial conceptualization of the work of historians strongly rooted in the philological analysis of written sources No need to differentiate from related disciplines	Broad convergence around the archive model Enrichment of the methodology via: - microhistory (Ginzburg, Levi) - collective memory (Halbwachs) - debate on revisionism (vs. Shoah) as lacking scientific foundations	Rejection of the neopositivism (historical truth not subject to causal confirmation) but strong emphasis on the scientific nature of historical work	Rejection of the post-structuralist thesis of no difference between fiction and historical texts (White) Fierce reaction based on methodological criteria of archival research	Carr (1961) Momigliano (1984) Novick (1988) Farge (1989) Appleby, Hunt and Jacob (1994) Eley (1996) Iggers (1997) Evans (1997, 2001) Chartier (2009) Tosh (2010) Jenkins (2003) Potin (2013)

(continued)

Table 1 (continued)

Discipline	Institutionalization process	Methodological orientation	Position with respect to neopositivism	Position with respect to post-structuralism	Selected references
Anthro-pology	Historical origins in the study of civilizations without written records Progressive separation from Sociology	In-depth case studies with strong theoretical orientation (Boas) Thick description (Geertz) Participant observation and ethnographic techniques Idiographic orientation	Initial orientation towards positivistic approach (Boas), then rejected in favour of the study of unique, non-inferential objects of analysis Rejection of nomothetic approach by using the case study and thick description methodology	Some limited impact Some acceptance in minority positions of hermeneutics and critical thinking Overall orientation aimed at preserving the material constitution of societies (against relativism)	Geertz (1973, 1983) Thoresen (1975) Fowler (1975) Clifford and Marcus (1986) D'Andrade (1995) Barnard (2000) Eriksen and Nielsen (2001) Darnell (2001, 2008) Barth et al. (2005) Clifford (2005) Kuklick (2008)

Political science	Historical origins in post-Second World War theory of democracies and their constituent elements (party systems, elections, government etc.) Need to sustain the comparison with hard sciences in US departments Need to differentiate from Sociology (on one side) and Economics (on the other side)	Three main methodological orientations - empirical/quantitative approach - rational choice - historical/comparative	Challenge by positivism strong in the US academic environment after Second World War Initial acceptance of quantitative approach in US political science Some reaction from European continental tradition rooted in history and political philosophy Strong controversies around rational choice and game theory modelling	Minor impact Minority positions use hermeneutics to study political orientations of social movements	Lerner and Lasswell (1951) Bernstein (1976) Lieberson (1985) Easton, Gunnell and Graziano (1990) Gunnell (1990) Ross (1991) Platt (1996) Lindblom (1997) Smith (1997) Haskell (2000) Apter (2001) Scott and Keates (2001)

(continued)

Table 1 (continued)

Discipline	Institutionalization process	Methodological orientation	Position with respect to neopositivism	Position with respect to post-structuralism	Selected references
Philology	Birth after the Renaissance Dominates the European academic landscape	Strong technical background Professionalism	Considered itself as a scientific discipline (science of texts)	Defense of text as an object of scientific inquiry, not subjective interpretation Strong rejection of post-structuralism	Scholes (1985) Grafton (1991, 1999, 2009) Fish (1995) Gumbrecht (2003) Bod (2013) Turner (2014)
History of art	Clearly separate object from other disciplines in Humanities	Strong technical background and methodology Iconology opened new ways to link visual arts and written sources	Central role of attribution as high level professional activity	Shared notion of authorship of pieces of art prevents the acceptance of post-structuralist claims	Preziosi (1989) Mansfield (2002)
Archaeology	Clearly separate object from other disciplines in Humanities	Strong material foundation Deep reflection on the notion of "proof" from excavations	Progressive assimilation of contributions from STEM in the excavation activity (eg chemistry, physics)	Defense of material findings as object of scientific inquiry	Renfrew and Bahn (2008) Gallay and Gardin (2009)

English literature	Dominance of Philology in US departments of Humanities in XIX century Institutionalization of Literary studies by opposition to Philology (full texts against chunks of text) Birth of Criticism as methodology to select valuable literary texts. Canon formation	Tension between Philology (strong technical background) and hermeneutics Recent status of the discipline tilted towards subjective interpretation approaches	n.a.	Broad acceptance of post-structuralism Critique of the literary canon as a form of intellectual hegemony and domination Cultural studies Minority studies	Fish (1980, 1989) Baldick (1983) von Hallberg (1983) McMurtry (1985) Scholes (1985, 1998) Graff (1987) Parrinder (1991) Court (1992) Guillory (1993) Bloom (1994) Shumway (1994) Gallagher (1997) Abrams (1997) Hollinger (1997, 2006) Renker (2007) Donoghue (2008) Bérubé and Ruth (2015)

reflection on the epistemic status of the discipline, as opposed to philology (Abrams 1997). Criticism was the solution to this differentiation process, but a consequence of this process was the need to discriminate between authentic literary works and other works. The formation of a canon was not initially a discriminatory practice but rather an epistemic necessity.

At the other extreme, history came to be institutionalised in universities shortly after its epistemic rationalisation. Its methodological foundations have been discussed for decades but within strongly held disciplinary boundaries, with limited need to differentiate from other disciplines.

The second building block is methodological orientation. A classical distinction is between a nomothetic orientation, or the formulation of law-like generalisations, subsuming a large empirical reality under general laws and an idiographic orientation, or the tendency to offer a detailed qualitative description of unique cases or situations. The methodological orientation is not unique in most SSH, although in some of them one can recognise a dominant paradigm and a few minority positions. In most cases there is a coexistence of a pluralism of methodologies, usually associated with higher level choices like the choice of the level of analysis or the value-laden presuppositions.

To make again two opposite examples, in economics the orientation taken by the mainstream, neoclassical school is that propositions take the form of law-like generalisations. The power to infer general conclusions from the examination of individual observations comes from the adoption of a modelling methodology, framed into a mathematical language, associated with rigorous inferential techniques from statistics and econometrics. At the other extreme, anthropologists refuse altogether the notion of law-like generalisations and instead offer long descriptions (Geertz 1973, 1983), or rich, extended, articulated descriptions of unique empirical settings, from which theoretical conclusions can be originated (but without the support of inferential techniques) (Eriksen and Nielsen 2001).

The methodological spectrum, between nomothetic and idiographic approaches, does not render justice to the methodological issue. Another dimension is whether scientific disciplines have developed a core of methodological rules that are commonly accepted and socialised in the discipline from the earliest days of careers (typically, at the doctoral level). History, for example, has a well-developed syllabus of methodological texts and exercises that are built around the discipline of archival sources (Farge 1989; Potin 2013). Surprisingly, art history has a common core of methodological rules centred around the notion of attribution. Attribution of a piece of art to an author, or a period, or a style: a strict discipline that summarises a large variety of technical skills and methodological rules. Another discipline with a strong core of methodological rules is archaeology. Here we find a quasi-discovery kind of human science: archaeological excavations are similar to scientific discoveries, and the explanation of findings mobilises a large set of logical (mainly abductive) rules of reasoning.

The third and fourth building blocks examine the way in which the discipline has addressed two major philosophical and epistemological challenges of the twentieth century: the rise of neo-positivism in the 1930s and its diffusion in the academic

environment after World War II, and the turbulent advent of post-structuralism, associated with the French school of Foucault, Derrida, Barthes, Lyotard, Baudrillard and with authors like Hayden White and Stanley Fish.

The neo-positivist position challenged non-STEM disciplines to demonstrate their scientific status (Bryant 1985). This pressure was felt strongly in the post-Second World War academic environment in the US, and found wide acceptance in economics, political science and psychology (see for example Lerner and Lasswell 1951; Bell 1982; Hilgard 1987), and, to a lesser extent, sociology (Bernstein 1983; Ross 1991; Platt 1996).

First let's examine economics. On the one hand, in economics the issue of scientific validity has been addressed (solved, many would say) by combining two elements: the language of mathematics and the axiomatic approach to motivation, or to the reasons for action of reflexive agents. The formalisation of economic variables in mathematical language ensures the controllability of propositions, as derived logically from formal premises. The adoption of a mathematical language has made economics a somewhat separate social science. The axiomatic approach addresses the fundamental problem posed by Max Weber: social agents are not like inanimate objects, whose behaviour can be examined objectively. Agents have reasons for action (motivations) and representation of reality (beliefs). To interpret their behaviour we need a theory of their reasons for action and their representation. But by definition, reflexive agents may modify their motivations and/or representation in response to the modelling exercise by social scientists. This creates a circle that cannot be closed in the same way as in natural sciences. Modern economic theory does not address these issues, but relies on the *assumption* that agents behave according to a set of abstract criteria described by the theory of rational choice. The axiomatic foundation rests on a powerful philosophical and logical base, which gives plausibility and prestige to the assumptions.

On the other hand, psychology has made somewhat the opposite move. Motivations and cognitions are not assumed at an axiomatic level but are made themselves observable. By developing a powerful experimental apparatus, modern psychological research has reduced all issues of human behaviour to the observable level. In this sense psychology reacted to the neo-positivist challenge in a different way, by accepting the experimental method and/or qualifying quasi-experimental or naturalistic methods in a truly causal perspective. As Hilgard (1987) qualifies it, psychology is committed to "find regularities within limited domains" (p. 803). Economics and psychology are, however, two exceptions. Other disciplines in Humanities, such as history or anthropology, literature or art history, and in Social Sciences, such as sociology or (to a certain extent) political science, have followed a different path. In these disciplines the neo-positivist challenge has been, generally speaking, plainly rejected.

In turn, the post-structuralist challenge originated in the French tradition of human sciences as a reaction against Levi Strauss's structuralism and linguistics (Gellner 1985, 1992; Lamont 1987). This tradition dissolved the distinction between scientific knowledge and folk knowledge by deconstructing the texts in which all kinds of knowledge are embedded. Contrary to the old tradition of philology, which

aimed at reconstructing the true meaning of the texts, post-structuralists emphasise the radical indetermination of the meaning of texts and the interaction between the text and the readers (following the lesson of hermeneutics after Gadamer and Ricoeur, but also after the theory of reception of literature by Iser 2012 and Jauss 1978). Consequently, there is no distinction between texts of various nature: for example, texts written by historians could not claim superior validity to texts of fiction (Rosenau 1992; Sarup 1993).

Two opposite reactions are worth examining. In history, the influential work of Hayden White (White 1973, 1978, 1987) tried to demonstrate that the notion of "historical truth", which was the backbone of the profession since the eighteenth century (Carr 1961), was void of content and ideological. In a number of brilliant books he made the point that historical reconstructions cannot claim any validity in addition to what can be obtained rhetorically for any kind of text (Nelson et al. 1987). Several authors refined his arguments and developed a post-structuralist theory of writing history (Appleby et al. 1994; Megill and McCloskey 1987; Jenkins 2003). Interestingly, the scientific community reacted negatively. Momigliano (1984) defended vigorously the peculiar notion of truth that is the normative goal of professional historians. Several books were written to reject this theory (Windschuttle 1996; Evans 1997, 2001; Iggers 1997), and its arguments are in practice no longer discussed in the community of historians. A reaction against this approach was also developed in philosophy (Boghossian 2006).

Several steps in the methodological evolution of the discipline in the twentieth century help to understand this fierce reaction. First, the discipline had already addressed the issue of the subjective role of historians in selecting archival sources, after the seminal works of Maurice Halbwachs (1925, 1950). Furthermore, it also addressed the epistemological issue of the nature of historical "proof", after the ambitious formulation of the "paradigma indiziario" by Carlo Ginzburg (1986) and the methodological programme of microhistory. Second, historians rejected the defence that authors like Derrida offered for Paul de Man, a Belgian author who migrated to the US, whose early work was found guilty of supporting Hitler's theses. Derrida and others argued that authorship is a collective enterprise and that responsibility was to be assigned to the context, not to the author. Professional historians strongly rejected this line of argumentation. Finally, historians in Europe and the US had to face the wave of revisionist writers, who claimed to be legitimate academic historians but denied the Holocaust. In this occasion, academic historians prohibited these authors from giving seminars in university departments, with the argument that revisionism has no scientific grounds. Summing up, history has developed a deep and articulated epistemic approach by building up a sophisticated methodological toolbox and by reacting vigorously to challenges, either from within the discipline or from outside.

The story is different for English literature. In this discipline the post-structuralist call for breaking the authority of authorship and deconstructing texts was embraced with enthusiasm. An influential argument was that all value judgments are contingent (Herrnstein-Smith 1988). The diffusion of new curricula in US universities about minority literature is a consequence of the critical approach to the formation

of the canon (Bloom 1994), or the list of academically admissible texts, to be studied by students and read in the classroom (von Hallberg 1983; Bérubé and Ruth 2015). An entire new disciplinary field was created, labelled Cultural Studies, in which the methodological tools do not come from philology or literary criticism but from human and social sciences.

Summing up, there are visible differences in the way in which disciplines in SSH have developed, institutionally and epistemically, since their foundation in the academic context. A comparative historical and epistemic analysis sheds light on interesting differences. It is my contention that these differences may explain the approaches taken by these disciplines with respect to research evaluation. I am now providing a sketch of this model based on considerable international literature (reviewed extensively in Bonaccorsi 2015), official documents of learned societies in Europe and elsewhere, and recent Italian experience. Table 2 summarises the main argument.

The definition and measurement of orientation towards the evaluation of research is an interesting issue that deserves further research. For the time being I summarise the evidence by illustrating a spectrum of positions, as follows.

(a) Bibliometric orientation: acceptance of quantitative evaluation based on indexed journals and citation measures.
(b) Positive orientation with extensive consensus on research quality criteria to be used in peer review.
(c) Positive orientation with controversies on research quality criteria.
(d) Negative orientation.

Economics and psychology are among the few disciplines in SSH that have accepted bibliometric evaluation, since their epistemic evolution has led to the acceptance of journals as the main communication channel. Their nomothetic orientation makes the comparability of results easier. They basically accepted the neo-positivist challenge to demonstrate their scientific status (although with different solutions) and rejected altogether the challenge of post-structuralism.

I predict a positive orientation towards research evaluation when disciplines have built up strong epistemic foundations and a shared methodological core. This feature can be found in history (archival work), philology (textual analysis), art history (attribution), archaeology (excavation) or anthropology (extensive description). All these disciplines rejected both neo-positivism and post-structuralism. A distinction can be made among them on the basis of the importance of epistemic pluralism. In disciplines such as history, philology and art history there is a large consensus on a number of criteria of research quality, even across differences in approaches (e.g., for historians of different political orientation).

There is a strong belief in the possibility of evaluating the quality of research work even if it is carried out by authors with an opposite overall orientation. In anthropology or political sciences the epistemic differences are somewhat more problematic. Political scientists taking the rational choice orientation are in conflict with those assuming the historical-comparative approach and vice versa (Apter 2001). In anthropology there is a strong core of methodological criteria, but there is

Table 2 Explanatory model of orientation towards research evaluation in selected SSH disciplines

Orientation towards research evaluation	Examples of discipline	Explanatory elements			
		Early institutionalisation process	Methodological orientation	Neo-positivism	Post-structuralism
Bibliometric	Economics Psychology	Autonomy	Nomothetic	Accepted	Rejected
Positive orientation-consensus	History	Autonomy	Idiographic but with strong methodological foundations	Rejected	Rejected
	Philology				
	Art history				
	Archaeology				
Positive orientation-pluralism	Anthropology Political science	Some conflict	Idiographic; strong methodological foundations but epistemic pluralism	Addressed the challenge, finally rejected	Rejected (accepted by minority positions)
Negative orientation	English literature (Anthropology)	Conflict	Hermeneutics	Rejected	Fully accepted

also a tradition of critical thinking that emphasises the importance of dissent, minority positions and political activism. In addition, a school of study adopted hermeneutics as the main methodology (Clifford 2005; Clifford and Marcus 1986), not without conflicts (D'Andrade 1995). One would classify anthropology somewhat between positive and negative orientation (Barnard 2000). These issues do not lead these disciplines to reject the evaluation altogether (although some authors in anthropology are among the most active against it: see Power 1997; Strathern 2000; Amit 2000; Dahler-Larsen 2012), but raise a series of fundamental questions about the preservation of pluralism and procedural fairness.

Finally, I predict a negative orientation if a discipline was born in conflict with other, more established and "scientific" ones, rejects the notion of a core of methodological foundations and rather accepts the post-structuralist claims. My analysis here is limited to English literature, for which the historical track record is extensive. In this discipline most authors subscribe to the notion that research evaluation is just another way for establishing domination in the academic world, limiting academic freedom. The linkage between evaluation, with its request for a set of agreed research criteria, and the formation of the literary canon, according to these authors, is too strong to stay unnoticed. Since the rejection of the canon (as defended by Bloom 1994) is one of the most recent foundational steps of the evolution of the discipline, it is no surprise to observe a rejection of evaluation as well.

There are other possible candidates for this negative position, including some schools in anthropology, sociology, philosophy (mainly Continental), or literature in various European countries (in addition to Italy, I would mention in particular France and Germany) (see for example Citton 2010 or Eagleton 2015). But to conclude in this direction would require a large scale comparative and historical analysis, similar to the one already carried out in the book.

My results are broadly consistent with the findings of Lamont (2009), one of the few comparative studies on differences in evaluation criteria of SSH disciplines. While Lamont is more interested in observations on procedural fairness in the evaluation process, I converge with her observations moving from an epistemic perspective.

4 Taking into Account Epistemic Differences and Evaluating with Fairness

The discovery of large differences within SSH disciplines, of course, creates a major problem for researching evaluation. As noted by the historian of architecture Carlo Olmo (in Bonaccorsi 2015) what is needed is a theory of *reception* of evaluation, a theory that takes into account the epistemic issues of various scientific communities.

Is such an effort feasible? Before answering this questions it is necessary to review the arguments, put forward in the international debate, that cast doubt on the feasibility and merit of evaluation.

In the last decades there has been a critical movement that has issued arguments based on the works of Michel Foucault (Foucault 1978a, b). His analysis of the modern institutions in medicine, public health, psychiatry, sexuality, prisons had an enormous influence. He has in fact realised the Holy Grail of the explanatory power of social sciences: to show that, underlying the observable reality and against all easily available evidence, there lies an order that is constructed by social actors following implicit rules of behaviour. These rules are not explicit to actors but are instead hidden behind apparently neutral and objective devices, like classifications, categories, numbers, and standards. These social devices shape reality in such a way that they make compliance the only rational behaviour, and deviance from the rules an anomaly.

More recently, in parallel with the surge of evaluation systems and the construction of indicators, the arguments from Foucault have been applied to the fields of higher education and science. According to a number of authors, evaluation realises the kind of surveillance identified by Foucault (Foucault 1966, 1975; Dean 1990) as the dominant trait of modern societies. The activities of universities, which have traditionally been entrusted with autonomy and academic freedom, are increasingly subject to inspection, measurement and evaluation by bodies external to academia (Power 1997; Strathern 2000; Amit 2000). These bodies incorporate instrumental rationality, by asking universities to behave as producers of identifiable objects, like publications, and not as critical social actors. Their rationality is inevitably associated with technical instrumentation, which is ideologically dangerous because it hides the manipulation behind apparently neutral technical indicators, which are presumed to be value-free and objective. The social acceptance that is given to authorities that use numbers is just a form of subordination to a new form or power. Thus quantification of social reality is just a form of power, dressed with the clothes of objectivity.

These arguments are fascinating. The interpretive power of Foucault's work is large, particularly in the studies based on meticulous philological work. Yet, I believe we have to resist this fascination. Like Ulysses with the sirens, we should listen to these arguments while being tightly tied to the mast. In my understanding, the mast is empirical research. Good social sciences should open the way to replicability of empirical findings in different settings and contexts, in space and time. And the findings should be subject to the kind of control that is required to give claims a scientific status.

It is true that commensuration *is* a form of power. Following Bourdieu (1984) universities are producers of cultural capital, a scarce resource that is distributed unevenly in society and consolidates power asymmetries. But the interesting question is whether this power is compatible with modern democracy, or is inevitably associated with manipulation and control.

To start with, the application of numbers to social reality is not the product of industrial capitalism, but goes back to the seventeenth century. It is in this period

that Pascal and Bernoulli lay the foundations for what is now the theory of probability, by giving a mathematical foundation to the notion of uncertainty. Their ideas were soon applied to social events, such as insurance against damages, or mortality tables of the population of large cities. Over time, the need of governments to raise taxes and to create welfare institutions required a standardisation of collection of data across regions and countries, making increasing use of social statistics (Anderson 1988; Anderson and Fienberg 1999; Bulmer et al. 1991).

Patriarca (1996) has shown that in nineteenth century Italian kingdoms the application of statistics to social reality was just an extension of ideas from the Enlightenment, against traditional sources of power. Indeed, statistics were a powerful tool to extend the domain of controllable knowledge, against the claim that social phenomena could only be examined using tacit, non-articulated, comprehensive kind of knowledge, as it was in traditional societies.

Classical studies in the history and sociology of statistics, such as Desrosières (1993, 2008a, b, 2014), Porter (1995) and Stigler (1999), have done a wonderful job in showing how social statistics are not the kind of objective type of knowledge that the public believes, but are inextricably linked to political power and its goals.

I argue that these contributions do not necessarily lead to reject commensuration (Espeland and Stevens 1998). The transformation of social reality into numbers is a fundamental way to extend social control over reality itself (Dudley Duncan 1984; Crosby 1997). It is a profoundly democratic process, although one that, contrary to other institutions of modern democracies, is more difficult to understand and is bound with more technical details. The fact that the public tends to trust numbers without questioning their origin and meaning, and therefore is subject to manipulation, is indeed true but does not detract from the importance of using numbers. Other institutions of modern societies, such as the media, are highly vulnerable to manipulation, but no serious scholar of modernity would deny their role. The fact is, commensuration requires a scientific approach unlike communication, entertainment, or journalism. The scientific approach is not intuitive and does not conform to common sense. Rather, it is highly counterintuitive and requires hard discipline and control of the reasoning process. People are not intuitive statisticians, as they are not intuitive scientists. Therefore commensuration tends to be cultivated only in small circles of experts whose mission is to devise ways to collect data, transform them into information, process information with the use of indicators and other tools in order to produce meaningful knowledge.

Commensuration is therefore an intrinsic part of modernity. The critical attitude towards commensuration, based on Foucault and Bourdieu, and more recently on the psychoanalytic and psychotherapeutic movement (see for example Gori and co-authors: Gori 2011, 2013; Abelhauser et al. 2011), does not serve democracy in society well.

Having stated this general point, let us turn to the question of whether SSH research may be subject to commensuration. SSH scholars have their own quality criteria. When they read a book, they are able to formulate a qualitative judgment about the merit of the underlying research. These judgements are robust with regard to ideological, methodological and political differences, even of a strong nature.

The challenge is whether qualitative judgments can be reliably transformed into quantitative measures, and whether these measures are comparable.

New results in social sciences support the view that this is indeed possible. On one hand, Michèle Lamont (2009) has persuasively shown that scientific communities in SSH have their own research quality criteria, expressed in qualitative terms but firmly held by the members. She also shows that communities that do not commit to the elaboration of such criteria due to the fragmentation of the discipline, ideological conflicts, or the weakness of methodological bases are communities that suffer from loss in reputation, cohesiveness and attractiveness for students.

On the other hand, recent developments in decision theory (Balinski and Laraki 2010) make it possible to conclude that qualitative judgements can be reliably transformed into measurements, even without imposing the unrealistic assumptions associated with rational choice theory. Furthermore, the aggregation and comparability of measurements does not necessarily require the even more unrealistic assumptions required in order to avoid the famous Arrow impossibility theorem (List and Pettit 2011). What is needed is not commensurability (or the existence of a common measurement), which is too demanding, but comparability, which is, according to philosopher Ruth Chang, always possible (Chang 1997, 2002). What is needed is just the moral and political willingness to compare, for the purpose of achieving socially beneficial goals (Bagnoli 2006).

Summing up, an epistemic approach to research evaluation is promising. It recognises the large differences in the epistemic procedures of disciplines in SSH (an argument extensively discussed by Ochsner, Hug and Daniel 2016) and opens the way to understanding their reception of evaluation. At the same time it firmly confirms that SSH disciplines may converge on a core set of discipline-specific quality criteria without violating epistemic pluralism, academic freedom, and the right to dissent. Once these core criteria are discursively established, people will produce qualitative judgements, not quantitative ones, on pieces of research produced by peers. Recent results in decision theory and social psychology confirm that it is possible to transform these qualitative judgements into an ordering (not necessarily a ranking), without violating personal preferences of the evaluators.

A necessary condition for evaluation is procedural fairness. This requires the adoption of a mix of procedural devices, such as transparency in the selection of experts, self-candidatures, rotation of experts, duplication of roles in presence of severe antagonism among schools, short periods in charge. A permanent dialogue with scientific communities should be kept open. Quality criteria must be published and regularly updated. Detailed and continuous work on the drafting and wording of questions to be adopted during the peer review process is also needed. All these solutions (and others) are important to generate trust on the side of evaluated researchers. It is a long process.

Under these conditions, research evaluation in SSH will not be accepted as a necessary evil, but as an occasion to re-open, or sometimes to establish from scratch, a self-reflexive exercise on research quality criteria. It is my contention that this exercise is valuable not only for SSH disciplines, but also for society.

References

Abbott, A. (2001). *The chaos of disciplines*. Chicago: The University of Chicago Press.

Abelhauser, A., Gori, R., & Sauret, M. J. (2011). *La folie èvaluation*. Paris: Mille et une nuits.

Abrams, M. H. (1997). The transformation of English Studies: 1930–1995. In T. Bender & C. E. Schorske (Eds.), *American academic culture in transformation*. Princeton: Princeton University Press.

Amit, V. (2000). The university as panopticon. Moral claims and attacks on academic freedom. In M. Strathern (Ed.), *Audit cultures. Anthropological studies in accountability, ethics and the academy*. London: Routledge.

Anderson, M. J. (1988). *The American census. A social history*. New Haven: Yale University Press.

Anderson, M. J., & Fienberg, S. E. (1999). *Who counts? The politics of census-taking in contemporary America*. New York: Russel Sage Foundation.

Appleby, J., Hunt, L., & Jacob, M. (1994). *Telling the truth about history*. New York: W.W. Norton and Company.

Apter, D. E. (2001). Structure, contingency, and choice: A comparison of trends and tendencies in political science. In J. W. Scott & D. Keates (Eds.), *Schools of thought. Twenty-five years of interpretive social science*. Princeton: Princeton University Press.

Bagnoli, C. (2006). Deliberare, confrontare, misurare. *Ragion Pratica, 26*, 65–80.

Baldi, S. (1998). Normative versus social constructivist processes in the allocation of citations: A network-analytic model. *American Sociological Review, 63*, 829–846.

Baldick, C. (1983). *The social mission of English criticism, 1848–1932*. Oxford: Clarendon Press.

Balinski, M., & Laraki, R. (2010). *Majority judgment. Measuring, ranking, and electing*. Cambridge, MA: The MIT Press.

Barnard, A. (2000). *History and theory in anthropology*. Cambridge: Cambridge University Press.

Barnes, B., & Edge, D. (Eds.). (1982). *Science in context. Readings in the sociology of science*. Milton Keynes: Open University Press.

Barnes, B., Bloor, D., & Henry, J. (1996). *Scientific knowledge. A sociological analysis*. Athlone: Chicago University Press.

Barth, F., Gingrich, A., Parkin, R., & Silverman, S. (2005). *One discipline, four ways: British, German, French, and American Anthropology*. Chicago: The University of Chicago Press.

Bastow, S., Dunleavy, P., & Tinkler, J. (2014). *The impact of the social sciences. How academics and their research make a difference*. Thousand Oaks: Sage Publications.

Bate, J. (Ed.). (2011). *The public value of humanities*. London: Bloomsbury Academic.

Bazerman, C. (1988). *Shaping written knowledge. The genre and activity of the experimental article in science*. Madison: The University of Wisconsin Press.

Becher, T. (1989). *Academic tribes and territories. Intellectual inquiry and the cultures of disciplines*. Milton Keynes: Open University Press.

Belfiore, E., & Upchurch, A. (2013). *Humanities in the twenty-first century. Beyond utility and markets*. Houndmills: Palgrave Macmillan.

Bell, D. (1982). *The social sciences since the Second World War*. New Brunswick: Transaction Books.

Benton, T., & Craib, I. (2011). *Philosophy of social science. The philosophical foundations of social thought* (2nd ed.). Basingstoke: Palgrave Macmillan.

Bernstein, R. J. (1976). *The restructuring of social and political theory*. New York: Harcourt Brace Jovanovich.

Bernstein, R. J. (1983). *Beyond objectivism and relativism. Science, hermeneutics and praxis*. Philadelphia: University of Pennsylvania Press.

Bérubé, M., & Ruth, J. (2015). *The humanities, higher education, and academic freedom*. Houndmills: Palgrave Macmillan.

Blaug, M. (1992). *The methodology of economics*. Cambridge: Cambridge University Press.

Bloom, H. (1994). *The Western canon*. New York: Harcourt Brace & Co.. (Italian translation Milan, Rizzoli, 1996).

Bod, R. (2013). *A new history of humanities. The search for principles and patterns from antiquity to the present*. Oxford: Oxford University Press.
Boghossian, P. (2006). *Fear of knowledge. Against relativism and constructivism*. Oxford: Clarendon Press.
Bonaccorsi, A. (2015). *La valutazione possibile. Teoria e pratica nel mondo della ricerca*. Bologna: Il Mulino.
Bornmann, L., & Daniel, H. D. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation, 64*(1), 45–80.
Bourdieu, P. (1984). *Homo academicus*. Paris: Les Editions de Minuit.
Brady, H. E., & Collier, D. (2010). *Rethinking social inquiry. Diverse tools, shared standards* (2nd ed.). Lanham: Rowman & Littlefield Publishers.
Brewer, J. D. (2013). *The public value of the social sciences*. London: Bloomsbury Academic.
Brooks, T. A. (1985). Private acts and public objects: An investigation of citer motivations. *Journal of the American Society for Information Science, 36*(4), 223–229.
Brooks, T. A. (1986). Evidence of complex citer motivations. *Journal of the American Society for Information Science, 37*(1), 34–36.
Brooks, P. (Ed.). (2014). *The humanities and public life*. New York: Fordham University Press.
Bryant, C. A. G. (1985). *Positivism in social theory and research*. Macmillan: Houndmills.
Bulmer, M., Bales, K., & Kish Sklar, K. (Eds.). (1991). *The social survey in historical perspective*. Cambridge: Cambridge University Press.
Camic, C., Gross, N., & Lamont, M. (2011). *Social knowledge in the making*. Chicago: The University of Chicago Press.
Carr, E. H. (1961). *What is history?* Houndmills: Palgrave Macmillan. (reprint London, Penguin Books, 1990).
Chang, R. (Ed.). (1997). *Incommensurability, incomparability, and practical reason*. Cambridge, MA: Harvard University Press.
Chang, R. (2002). *Making comparisons count*. New York: Routledge.
Chartier, R. (2009). *Au bord de la falaise. L'historie entre certitudes et inquiétude*. Paris: Albin Michel. (first edition, 1998).
Citton, Y. (2010). *L'avenir des Humanités. Èconomie de la connaissance ou cultures de l'interpretation?* Paris: La Découverte.
Clifford, J. (2005). Rearticulating anthropology. In D. A. Segal & S. J. Yanagisako (Eds.), *Unwrapping the sacred bundle: Reflections on the disciplining of anthropology*. Durham: Duke University Press.
Clifford, J., & Marcus, G. E. (Eds.). (1986). *Writing culture. The poetics and politics of ethnography*. Berkeley: University of California Press.
Collins, H. M. (1975). The seven sexes: A study in the sociology of a phenomenon, or the replication of experiment in physics. *Sociology, 9*, 205–224.
Collins, H. M. (1985). *Changing order. Replication and induction in scientific practice*. Chicago: The University of Chicago Press.
Collins, H. M. (1999). Tantalus and the aliens: Publications, audiences and the search for gravitational waves. *Social Studies of Science, 29*, 163–197.
Court, F. E. (1992). *Institutionalizing English literature. The culture and politics of literary study, 1750–1900*. Stanford: Stanford University Press.
Cronin, B. (1984). *The citation process. The role and significance of citations in scientific communication*. Oxford: Taylor Graham.
Cronin, B. (2005). *The hand of science. Academic writing and its rewards*. Lanham: The Scarecrow Press.
Crosby, A. W. (1997). *The measure of reality. Quantification and Western society, 1250–1600*. Cambridge: Cambridge University Press.
Dahler-Larsen, P. (2012). *The evaluation society*. Stanford: Stanford Business Books.
D'Andrade, R. (1995). *The development of cognitive anthropology*. Cambridge: Cambridge University Press.

Danziger, K. (1990). *Constructing the subject. Historical origins of psychological research*. Cambridge: Cambridge University Press.
Darnell, R. (2001). *Invisible genealogies. A history of Americanist anthropology*. Lincoln: University of Nebraska Press.
Darnell, R. (2008). North American traditions in anthropology: The historiographic baseline. In H. Kuklick (Ed.), *A new history of anthropology* (pp. 35–51). Malden: Blackwell Publishing.
Dasgupta, P., & David, P. (1994). Toward a new economics of science. *Research Policy, 23*, 487–521.
Dean, M. (1990). *Governmentality. Power and rule in modern society*. Los Angeles: Sage Publications.
Delanty, G. (2005). *Social science* (2nd ed.). Maidenhead: Open University Press.
Delanty, G., & Strydom, P. (2003). *Philosophies of social science*. Maidenhead: Open University Press.
Della Porta, D., & Keating, M. (Eds.). (2008). *Approaches and methodologies in the social sciences. A pluralist perspective*. Cambridge: Cambridge University Press.
Desrosières, A. (1993). *La politique des grands nombres. Historie de la raison statistique*. Paris: La Découverte.
Desrosières, A. (2008a). *Pour une sociologie historique de la quantification. L'argument statistique I*. Paris: Presses de l'École des Mines.
Desrosières, A. (2008b). *Gouverner par les nombres. L'argument statistique II*. Paris: Presses de l'École des Mines.
Desrosières, A. (2014). *Prouver et gouverner. Une analyse politique des statistiques publiques*. Paris: La Découverte.
Donoghue, F. (2008). *The last professors. The corporate university and the fate of the humanities*. New York: Fordham University Press.
Dudley Duncan, O. (1984). *Notes on social measurement*. New York: Russell Sage Foundation.
Eagleton, T. (2015). The slow death of the university, in *The Chronicle of Higher Education*, 19 August, in http://chronicle.com/article/The-Slow-Death-of-the/228991/. Last access 19 Aug 2015.
Easton, D., Gunnell, J. G., & Graziano, L. (1990). *The development of political science. A comparative survey*. London: Routledge.
Eley, G. (1996). Is all the world a text? From social history to the history of society two decades later. In T. H. McDonald (Ed.), *The historic turn in the human sciences*. Ann Arbor: The University of Michigan Press.
Eriksen, T. H., & Nielsen, F. S. (2001). *A history of anthropology*. London: Pluto Press.
Espeland, W. N., & Stevens, M. L. (1998). Commensuration as a social process. *Annual Review of Sociology, 24*, 313–343.
Evans, R. J. (1997). *In defense of history*. London: Granta Books.
Evans, R. J. (2001). Introduction. In E. H. Carr (Ed.), *What is history?* Houndmills: Palgrave Macmillan.
Farge, A. (1989). *Le goût de l'archive*. Paris: Èditions de Seuil.
Fish, S. (1980). *Is there a text in this class? The authority of interpretive communities*. Cambridge, MA: Harvard University Press.
Fish, S. (1989). *Doing what comes naturally. Change, rhetoric and the practice of theory in literary and legal studies*. Oxford: Clarendon Press.
Fish, S. (1995). *Professional correctness. Literary studies and political change*. Oxford: Clarendon Press.
Flyvbjerg, B. (2001). *Making social science matter. Why social inquiry fails and how it can succeed again*. Cambridge: Cambridge University Press.
Foucault, M. (1966). *Les mots et les choses*. Paris: Editons Gallimard.
Foucault, M. (1975). *Surveiller et punir. Naissance de la prison*. Paris: Editions Gallimard.

Foucault, M. (1978a). La «gouvernamentalitè» : Cours du Collège de France, Annéè 1977–1978 : Securitè, territoire et population, 4e le*çon,* 1er février 1978. In M. Foucault, *Dits et écrits II. 1976–1988.* Paris, Gallimard, 2001, pp. 635–657.

Foucault, M. (1978b). Introduction by Michel Foucault. In G. Canguilhem, *On the normal and the pathological*. Boston, D. Reidel, p. ix–xx. In M. Foucault, *Dits et écrits II. 1976–1988.* Paris, Gallimard, 2001, pp. 429–442.

Fowler, D. D. (1975). Notes on inquiries in anthropology: A bibliographic essay. In H. H. Thoresen (Ed.), *Toward a science of man. Essays in the history of anthropology*. The Hague: Mouton Publishers.

Frickel, S., & Moore, K. (Eds.). (2006). *The new political sociology of science. Institutions, networks, and power*. Madison: The University of Wisconsin Press.

Friedman, M. (1952). The methodology of positive economics. In *Essays in positive economics*. Chicago: University of Chicago Press.

Gallagher, C. (1997). The history of literary criticism. In T. Bender & C. E. Schorske (Eds.), *American academic culture in transformation*. Princeton: Princeton University Press.

Gallay, A., & Gardin, J. C. (2009). Les méthodes logicistes en archéologie. Perspectives et limites. In B. Walliser (Ed.), *La cumulativité du savoir en sciences sociales*. Paris: Èditions de l'École des Hautes Ètudes en Sciences Sociales.

Geertz, C. (1973). *The interpretation of cultures*. New York: Basic Books.

Geertz, C. (1983). *Local knowledge*. Hammersmith: Fontana Press.

Gellner, E. (1985). *Relativism and the social sciences*. Cambridge: Cambridge University Press.

Gellner, E. (1992). *Postmodernism, reason and religion*. London: Routledge.

Gieryn, T. F. (1983). Boundary-work and the demarcation of science from non-science: Strains and interests in professional ideologies of scientists. *American Sociological Review, 48*, 781–795.

Gieryn, T. F. (1995). Boundaries of science. In S. Jasanoff, G. E. Markle, J. C. Petersen, & T. Pinch (Eds.), *Handbook of science and technology studies*. Thousand Oaks: Sage Publications.

Gieryn, T. F. (1999). *Cultural boundaries of science. Credibility on the line*. Chicago: The University of Chicago Press.

Ginzburg, C. (1986). *Miti emblemi spie. Morfologia e storia*. Turin: Einaudi.

Goertz, G. (2006). *Social science concepts. A user's guide*. Princeton: Princeton University Press.

Goertz, G., & Mahoney, J. (2012). *A tale of two cultures. Qualitative and quantitative research in the social sciences*. Princeton: Princeton University Press.

Gori, R. (2011). *La dignité de penser*. Paris: Les Liens qui Libèrent.

Gori, R. (2013). *La societè des imposteurs*. Paris: Les Liens qui Libèrent.

Graff, G. (1987). *Professing literature. An institutional history*. Chicago: The University of Chicago Press.

Grafton, A. (1991). Defenders of the text. In *The traditions of scholarship in an age of science, 1450–1800*. Cambridge, MA: Harvard University Press.

Grafton, A. (1999). *The footnote. A curious history*. Cambridge, MA: Harvard University Press. (ed. or. Paris, Èditions de Seuil, 1997).

Grafton, A. (2009). Worlds made by words. In *Scholarship and community in the Modern West*. Cambridge, MA: Harvard University Press.

Guala, F. (2006). *Filosofia dell'economia. Modelli, causalità, previsione*. Bologna: Il Mulino.

Guillory, J. (1993). *Cultural capital. The problem of literary canon formation*. Chicago: The University of Chicago Press.

Gumbrecht, H. U. (2003). *The powers of philology. Dynamics of textual scholarship*. Urbana/ Chicago: University of Illinois Press.

Gunnell, J. G. (1990). The historiography of American political science. In D. Easton, J. G. Gunnell, & L. Graziano (Eds.), *The development of political science*. London: Routledge.

Halbwachs, M. (1925). *Les cadres sociaux de la mémoire*. Paris: Albin Michel. (new edition, 1994).

Halbwachs, M. (1950). *La mémoire collective*. Paris: Albin Michel. (new edition, 1997).

von Hallberg, R. (1983). *Canons*. Chicago: The University of Chicago Press.

Haskell, T. L. (2000). *The emergence of professional social science. The American Social Science Association and the Nineteenth-Century crisis of authority*. Baltimore: The Johns Hopkins University Press. (ed. or. University of Illinois Press, 1977).
Hausman, D. (1992). *The inexact and separate science of economics*. Cambridge: Cambridge University Press.
Herrnstein-Smith, H. (1988). *Contingencies of value*. Cambridge, MA: Harvard University Press.
Hilgard, E. R. (1987). *Psychology in America. A historical survey*. San Fiego: Harcourt Brace Jovanovich.
Hollinger, D. A. (1997). The disciplines and the identity debates, 1970–1995. In T. Bender & C. E. Schorske (Eds.), *American academic culture in transformation*. Princeton: Princeton University Press.
Hollinger, D. A. (Ed.). (2006). *The humanities and the dynamics of inclusion since World War II*. Baltimore: The Johns Hopkins University Press.
Hyland, K. (2012). *Disciplinary identities. Individuality and community in academic discourse*. Cambridge: Cambridge University Press.
Iggers, G. G. (1997). *Historiography in the twentieth century*. Middletown: Wesleyan University Press.
Iser, W. (2012). *L'appel du texte*. Paris: Èditions Allia.
Jauss, H. R. (1978). *Pour une esthétique de la réception*. Paris: Èditions Gallimard.
Jenkins, K. (2003). *Re-thinking history*. London: Routledge. (ed. or. 1991).
Kernan, A. (Ed.). (1997). *What's happened to the humanities?* Princeton: Princeton University Press.
King, G., Keohane, R. O., & Verba, S. (1994). *Designing social inquiry*. Princeton: Princeton University Press.
Knorr Cetina, K. (1999). *Epistemic cultures. How the science makes knowledge*. Cambridge, MA: Harvard University Press.
Kuklick, H. (Ed.). (2008). *A new history of anthropology*. Malden: Blackwell Publishing.
Lamont, M. (1987). How to become a dominant French philosopher: The case of Jacques Derrida. *American Journal of Sociology, 93*(3), 584–622.
Lamont, M. (2009). *How professors think. Inside the curious world of academic judgment*. Cambridge, MA: Harvard University Press.
Latour, B., & Woolgar, S. (1979). *La vie de laboratoire. La production des faits scientifiques*. Paris: La Découverte.
Lenoir, T. (1997). *Instituting science. The cultural production of scientific disciplines*. Stanford: Stanford University Press.
Lerner, D., & Lasswell, H. D. (1951). *The policy sciences. Recent developments in scope and method*. Stanford: Stanford University Press.
Lieberson, S. (1985). *Making it count. The improvement of social research and theory*. Berkeley: University of California Press.
Lindblom, C. E. (1997). Political science in the 1940s and 1950s. In T. Bender & C. E. Schorske (Eds.), *American academic culture in transformation*. Princeton: Princeton University Press.
Lindsay, D. (1978). *The scientific publication system in social science. A study of the operation of leading professional journals in Psychology, Sociology, and Social Work*. San Francisco: Jossey-Bass.
List, C., & Pettit, P. (2011). *Group agency. The possibility, design, and status of corporate agents*. Oxford: Oxford University Press.
Mansfield, E. (2002). *Art history and its institutions. Foundations of a discipline*. London: Routledge.
McMurtry, J. (1985). *English language, English literature. The creation of an academic discipline*. Hamden: Archon Books.
Megill, A., & McCloskey, D. N. (1987). The rhetoric of history. In J. S. Nelson, A. Megill, & D. N. McCloskey (Eds.), *The rhetoric of the human sciences. Language and argument in scholarship and public affairs*. Madison: The University of Wisconsin Press.

Momigliano, A. (1984). The rhetoric of history and the history of rhetoric: On Hayden White's Tropes. In *Settimo contributo alla storia degli studi classici del mondo antico* (pp. 49–59). Roma: Edizioni di Storia e Letteratura.
Moses, J. W., & Knutsen, T. L. (2007). *Ways of knowing. Competing methodologies in social and political research*. Basingstoke: Palgrave MacMillan.
Mulkay, M. (1991). *Sociology of science. A sociological pilgrimage*. Milton Keynes: Open University Press.
Nelson, J. S., Megill, A., & McCloskey, D. N. (Eds.). (1987). *The rhetoric of the human sciences. Language and argument in scholarship and public affairs*. Madison: The University of Wisconsin Press.
Novick, P. (1988). *That noble dream: The "Objectivity Question" and the American historical profession*. Cambridge, MA: Cambridge University Press.
Ochsner, M., Hug, S. E., & Daniel, H. D. (Eds.). (2016). *Research assessment in the humanities. Towards criteria and procedures*. New York: Springer.
Parrinder, P. (1991). *Authors and authority. English and American criticism, 1750–1990*. Houndmills: Macmillan.
Patriarca, S. (1996). *Numbers and nationhood. Writing statistics in nineteenth-century Italy*. Cambridge: Cambridge University Press.
Platt, J. (1996). *A history of sociological research methods in America, 1920–1960*. Cambridge: Cambridge University Press.
Porter, T. M. (1995). *Trust in numbers. The pursuit of objectivity in science and public life*. Princeton: Princeton University Press.
Potin, Y. (2013). L'historien en "ses" archives. In C. Granger (Ed.), *À quoi pensent les historiens? Faire de l'historie au XXI^e siècle*. Paris: Èditions Autrement.
Power, M. (1997). *The audit society. Rituals of verification*. Oxford: Oxford University Press.
Preziosi, D. (1989). *Rethinking art history. Meditations on a coy science*. New Haven: Yale University Press.
Renfrew, C., & Bahn, P. (2008). *Archaeology: Theories, methods and practice* (5th ed.). London: Thames and Hudson.
Renker, E. (2007). *The origins of American literature studies. An institutional history*. Cambridge: Cambridge University Press.
Rosenau, P. M. (1992). *Post-modernism and the social sciences. Insights, inroads, and intrusions*. Princeton: Princeton University Press.
Ross, D. (1991). *The origins of American social science*. Cambridge: Cambridge University Press.
Sarup, M. (1993). *An introductory guide to post-structuralism and postmodernism*. New York: Harvester Wheatsheaf.
Sayer, A. (1992). *Method in social science. A realist approach* (2nd ed.). New York: Routledge.
Sayer, A. (2000). *Realism and social science*. Thousand Oaks: Sage Publications.
Scholes, R. (1985). *Textual power. Literary theory and the teaching of English*. New Haven: Yale University Press.
Scholes, R. (1998). *The rise and fall of English*. New Haven: Yale University Press.
Scott, J. W., & Keates, D. (Eds.). (2001). *Schools of thought. Twenty-five years of interpretive social science*. Princeton: Princeton University Press.
Shumway, D. R. (1994). *Creating American civilization. A genealogy of American literature as an academic discipline*. Minneapolis: University of Minnesota Press.
Small, H. (2013). *The value of the humanities*. Oxford: Oxford University Press.
Smith, R. M. (1997). Still blowing in the wind: The American quest for a democratic, scientific political science. In T. Bender & C. E. Schorske (Eds.), *American academic culture in transformation*. Princeton: Princeton University Press.
Steele, D., & Guala, F. (2011). *The philosophy of social science reader*. London: Routledge.
Steinmetz, G. (Ed.). (2005). *The politics of method in the human sciences. Positivism and its epistemological others*. Durham: Duke University Press.

Stigler, S. M. (1999). Statistics on the table. In *The history of statistical concepts and methods*. Cambridge, MA: Harvard University Press.
Strathern, M. (Ed.). (2000). *Audit cultures. Anthropological studies in accountability, ethics and the academy*. London: Routledge.
Szenberg, M., & Ramrattan, L. (2014). *Secrets of economics editors*. Cambridge, MA: The MIT Press.
Taylor, C. A. (1996). *Defining science. A rhetoric of demarcation*. Madison: The University of Wisconsin Press.
Thoresen, T. H. H. (Ed.). (1975). *Toward a science of man. Essays in the history of anthropology*. The Hague: Mouton Publishers.
Tosh, J. (2010). *The pursuit of history. Aims, methods and new directions in the study of modern history* (5th ed.). London: Longman.
Turner, J. (2014). *Philology. The forgotten origins of the modern humanities*. Princeton: Princeton University Press.
Walliser, B. (2009a). Les trois sources de la cumulativité en économie. In B. Walliser (Ed.), *La cumulativité du savoir en sciences sociales*. Paris: Èditions de l'École des Hautes Ètudes en Sciences Sociales.
Walliser, B. (Ed.). (2009b). *La cumulativité du savoir en sciences sociales*. Paris: Èditions de l'École des hautes études en sciences sociales.
White, H. (1973). *Metahistory. The historical imagination in nineteenth-century Europe*. Baltimore: The Johns Hopkins University Press.
White, H. (1978). *Tropics of discourse. Essays in cultural criticism*. Baltimore: The Johns Hopkins Press.
White, H. (1987). *The content of the form. Narrative discourse and historical representation*. Baltimore: Johns Hopkins University Press.
Windschuttle, K. (1996). *The killing of history. How literary critics and social theorists are murdering our past*. New York: Encounter Books.
Yearley, S. (2005). *Making sense of science. Understanding the social studies of science*. Thousand Oaks: Sage Publications.
Ziman, J. (1978). *Reliable knowledge. An exploration of the grounds for belief in science*. Cambridge: Cambridge University Press.
Ziman, J. (2000). *Real science. What it is, and what it means*. Cambridge: Cambridge University Press.

Part I
Research Quality Criteria in SSH

Mapping the Role of the Book in Evaluation at the Individual and Department Level in Italian SSH. A Multisource Analysis

Chiara Faggiolani and Giovanni Solimine

1 The Centrality of the Book: Quantitative Tools for an Informed Peer Review

Fashion exerts an influence also in the sphere of research evaluation, as in so many others, and the last few years have unquestionably seen an alternation in Italy of positive and negative attitudes towards bibliometrics among scholars of social sciences and humanities (SSH), as shown by the growth of the associated national literature (Baccini 2010; De Bellis 2014; Faggiolani 2015).

Against this upsurge in interest, part of the SSH community has displayed increasing scepticism[1] with respect not only to bibliometrics but also to the evaluation of research in general (Pinto 2012; Banfi et al. 2014).

It is often claimed that bibliometrics is not applicable to these disciplines due to the inadequate coverage of the databases used to calculate the most widely used indicators, such as the journal impact factor (JIF) of the Thomson Reuters Web of Science (WoS).

In fact, the problem arises further back, as the poor coverage of the SSH literature by Scopus and WoS is itself due to the intrinsic characteristics of the disciplines involved, which require a different tool capable of doing justice to their specificity and channels of dissemination.

This contribution presents some of the findings of the LI.B.RO research project developed by the authors together with Carla Basili, Andrea Capaccioni, Stella Iezzi, Luca Lanzillo, Mario Mastrangelo, Giovanni Paoloni and Giovanna Spina and supported by ANVUR.

[1] See the website ROARS – Return on Academic ReSearch http://www.roars.it/online/

C. Faggiolani (✉) • G. Solimine
Dipartimento di Scienze documentarie, linguistico-filologiche e geografiche,
Università di Roma La Sapienza, Rome, Italy
e-mail: chiara.faggiolani@uniroma1.it; giovanni.solimine@uniroma1.it

© Springer International Publishing AG 2018

A. Bonaccorsi (ed.), *The Evaluation of Research in Social Sciences and Humanities*, https://doi.org/10.1007/978-3-319-68554-0_2

The specificity of these areas of knowledge arises not only from the different ways in which the results of research are circulated with respect to the "hard sciences" – as is known, monographs, books in general and the national literature play a primary role rather than articles in journals (Hicks 2004) – but also from the variety and peculiarity of the epistemic nature of the individual disciplines and the various methodological approaches adopted (see the chapter by Bonaccorsi in this book).

Suffice it to consider how the modern social sciences are organised into a range of theoretical, historical and normative disciplines comprising economic theory, sociology, anthropology, political science, social psychology and law.

With respect to the use of bibliometrics in the evaluation of SSH research, the common approach is an attempt to identify one or more automatic, quantitative tools (Hammarfelt 2016) to serve as support. While peer review indisputably constitutes the key tool for the evaluation of research quality, bibliometrics can help to counteract its potentially subjective and arbitrary aspects, not least in view of a certain degree of academic malpractice, sometimes coloured by bias, capable of justifying everything and its opposite.

The evaluation of the quality of articles in journals in the hard sciences rests on citation counts and therefore on bibliometric indicators of impact, which are effective tools and almost universally accepted (Figà Talamanca 2000). As is known, however, the SSH journals are only minimally covered by the major international databases and it is therefore impossible to employ the traditional bibliometric tools. For this reason, a proxy of the concept of quality has been identified in *reputation*, thus adopting the logic that also underpins the commercial databases, whereby it is the container that confers value on the content.

This recalls one of the three basic mathematical laws of modern bibliometrics formulated by Samuel C. Bradford (1878–1948), a librarian at the Patent Office in London, who demonstrated through his studies that a limited number of journals produced 90% of the fundamental literature in a specific area of research (Bradford 1934). This law subsequently provided the basis for the Science Citation Index.

In Italy, with the support of national associations, ANVUR (Agenzia nazionale di valutazione dell'università e della ricerca) has ranked the journals publishing work by Italian scholars in three classes: A-rated journals, recognized as excellent at the international level for the rigour of the review procedures and for circulation, esteem and ascertained impact within the academic community of the sector; scientific journals, which enjoy a good reputation in the academic community of reference and have at least national circulation; non-scientific journals, which include all the others. Nothing appears to have been established as regard monographs, however, and the question of whether it is possible to imagine an application of bibliometrics today, with the tools currently available,[2] still awaits a precise answer.

[2] We refer to the Thompson Reuters Book Citation Index (BKCI) and the Elsevier Books Expansion Project, for which readers are referred to the dedicated webites: http://wokinfo.com/products_tools/multidisciplinary/bookcitationindex/ and http://blog.scopus.com/topics/books

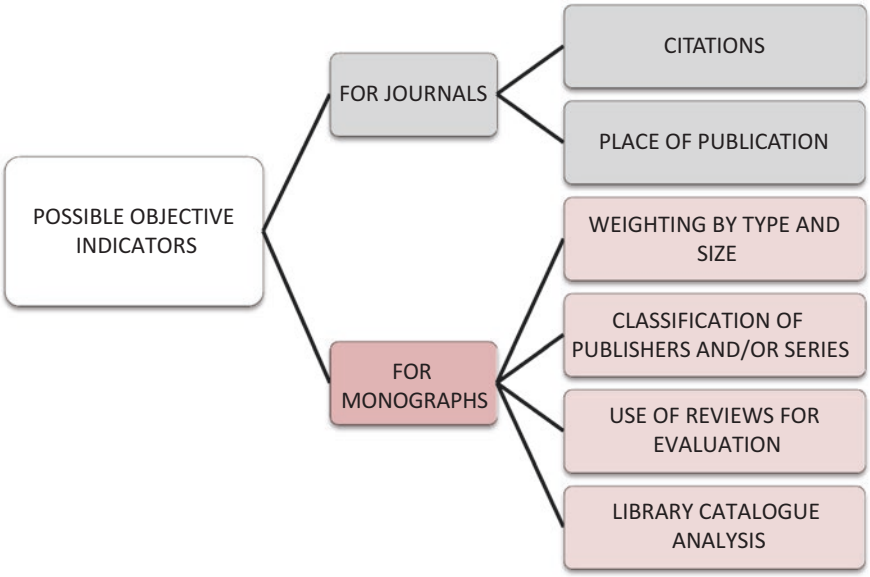


Fig. 1 Quantitative tools for the evaluation of monographs in comparison with journal articles

An attempt has been made by the present authors within the framework of the LI.B.RO project to ascertain the applicability and effectiveness of certain tools (Fig. 1) – in which the international literature has shown growing interest – with respect to the specificity of Italian contributions in the SSH:

- Weighting coefficients for monographs by type and size
- Possible criteria for the evaluation of reputation and classification of publishers and series
- Systems for the use of reviews in evaluation
- Library catalogue analysis

Two mappings were carried out for this purpose. One examines the complex world underlying the publication of the monograph and the book in general, as discussed in this volume by Carla Basili and Luca Lanzillo. The other regards the circulation of the SSH monograph in Italy and its role in the primary activities of research evaluation, also in relation to the criteria employed, as discussed here by Stella Iezzi.

Against this background, some considerations are presented based on data regarding the primary Italian activities of evaluation in recent years, specifically the analysis of publications submitted for VQR (Valutazione della qualità della ricerca, 2004–2010) and the criteria and requirements for the national university teaching qualification (ASN).

It is indeed our contention that a clear understanding of the conduct of scholars as both evaluatees and evalutators is essential to any objective and responsible

definition of the contribution that bibliometrics can make to SSH evaluation, not as the only tool but as an aid to support peer review to counteract any potentially subjective and arbitrary aspects.

As Eugene Garfield stated in 1979:

> The only responsible claim made for citation counts as an aid in evaluating individuals is that they provide an objective measure of the utility or impact of scientific work. They say nothing about the nature of the work, nothing about the reason for its utility or impact. Those factors can be dealt with only by content analysis of the cited material and the exercise of knowledgeable peer judgment. Citation analysis is not meant to replace judgment, but to make it more objective and astute (Garfield 1979, p. 364).

It should be recalled that this thinking underlies the Leiden Manifesto, which calls for a more limited role of bibliometrics in evaluation (not only for SSH) and points out that imposing the models of other disciplinary spheres on SSH could prove dangerous and result over time in a deterioration of research due to a change in strategy (Hicks et al. 2015).

2 Some Italian Specificities

Before entering into the analysis it is important to introduce some background information on the Italian context.

This specificity essentially concerns the birth of research evaluation in Italy and the classification of scientific disciplines.

2.1 A Brief History of the Evaluation of Research in Italy

All the aspects of the Italian university system have been subjected to evaluation processes for about 20 years now, perhaps longer and more intensively than any other sectors of public administration.

This is an obvious consequence of the laws that govern the normative, organisational, financial, scholarly and educational autonomy of Italian universities, which have yet to produce all their potential effects: from Law 168 of 1989, which instituted the new Ministry of Universities and Research (Ministero dell'Università e della Ricerca, hereafter MIUR) and granted broad powers of educational and statutory autonomy, to the recent Law 240 of 2010.

Needless to say, while the need for accountability of the university system arises for the reasons outlined above, it is also connected with the frequent media campaigns against the phenomena of nepotism and "demeritocracy" that so often taint the life of universities.

The evaluation of research began in Italy with Law 537 of 1993, which instituted Nuclei di valutazione and the Osservatorio per la valutazione del sistema universitario, replaced in 1999 with the Comitato nazionale per la valutazione del sistema universitario (CNVSU).

Instituted by Legislative Decree 204 of 5 June 1998, the Comitato di indirizzo per la valutazione della ricerca (CIVR) came into operation in 2004 and launched the first three-year evaluation programme (VTR) for the period 2001–2003 (Reale 2008), some 13 years after the RAE (Research Assessment Exercise) launched by the Thatcher government in the UK in 1986 (now the Research Excellence Framework or REF).

The first VTR (2001–2003) involved the evaluation of 17,329 research products in 20 disciplinary sectors submitted by 102 bodies (77 universities and 25 public research centres) over a period of 18 months between 2004 and 2005. Its completion led to marked acceleration in this sphere by bringing the need for effective links between the results of research and the allocation of resources to the attention of the authorities and the general public.

The method used was the peer review, with experts appointed by the relevant panel to evaluate publications on the criteria of quality, importance, originality/innovation and internationalisation and/or international competitive potential.[3]

The best research must be rewarded and evaluation makes it possible to show how public funds are being spent and ascertain whether resources are being used efficiently. This was the major contribution of the first VTR, the results of which were also taken into account in the allocation of funds for the extraordinary recruitment of researchers for the three-year period 2007–2009.

This need for public accountability led to the requirement for university departments to identify the areas of excellence and strengths of their research so as to attract funding and compete with one another.

This is the background to the activities carried out in Italy over the last few years by the ANVUR,[4] comprising VQR for the period 2004–2010, the national university teaching qualification (Abilitazione Scientifica Nazionale, hereafter ASN) for the years 2012 and 2013, and the introduction of the Scheda Unica Annuale per la ricerca dipartimentale (SUA-RD).

The SUA-RD is a yearly dossier containing the pertinent information and data for the evaluation of research within the system of periodic accreditation (AVA) and evaluation of the university quality control system as well the annual monitoring of research results for the purposes of periodic evaluation. The latter aspect will require provision to integrate the results of the latest VQR with those made available annually in the SUA-RD.[5]

[3] The final report is available online: http://vtr2006.cineca.it/php5/relazione_civr/output/civr_prima_parte.pdf

[4] http://www.anvur.org/

[5] http://www.anvur.org/index.php?option=com_content&view=article&id=394&Itemid=439&lang=it

These activities have prompted lively debate on the need to identify criteria capable of providing an objective yardstick in both quantitative and qualitative terms of the production of professors and researchers and of the universities and departments in which they operate.

The third VQR (2011–2014)[6] was launched in 2015 after VTR 2001–2003 and VQR 2004–2010.

The aim is essentially to monitor the health of the university sphere, the relationship between resources invested and results produced, and adequacy with respect to the requirements of Italian society. All this has not always been well received by the Italian academic world, which has repeatedly manifested distrust and unwillingness to cooperate and to engage in dialogue with the primary bodies involved in evaluation.

2.2 *The Classification of Disciplines in Italy: Disciplinary Sectors*

As stated by the Consiglio Universitario Nazionale (CUN),[7] an elected body representing the Italian university system, correct classification of the branches of knowledge constitutes one of the key elements capable of fostering the sciences with major cultural repercussions.

The CUN serves the MIUR as an independent source of advice and recommendations on matters considered relevant to the university system, such as national programmes, policies and administrative practices affecting higher education, the classification and definition of academic sectors and disciplines for the purposes of recruitment, teaching and research, funding issues and the approval of university teaching regulations.

The present framework of reference for Italian universities as regards both the selection of university professors and researchers at the various levels and the structuring of the degree system comprises 14 areas divided by the CUN into disciplinary sectors (*settori scientifico-disciplinari*, hereafter SSD).

The areas are as follows:

Area 1 – Mathematics and computer science
Area 2 – Physics
Area 3 – Chemistry
Area 4 – Earth sciences
Area 5 – Biology
Area 6 – Medicine
Area 7 – Agricultural and veterinary sciences

[6] See the call for submissions: http://www.anvur.org/attachments/article/825/Bando%20VQR%202011-2014_secon~.pdf

[7] https://www.cun.it/homepage/#null

Area 8 – Civil engineering and architecture
Area 9 – Industrial and information engineering
Area 10 – Antiquities, philology, literary studies, art history
Area 11 – History, philosophy, pedagogy and psychology
Area 12 – Law
Area 13 – Economics and statistics
Area 14 – Political and social sciences

The term *settore scientifico-disciplinare* was used for the first time in Law 341 of 19 November 1990 and the classification of the above 14 areas is regulated by the MIUR Ministerial decree of 23 December 1999 with subsequent modifications.

Law 240 of 30 December 2010 introduced a number of innovations, the most important of which as regards research evaluation include the *settore concorsuale* (competition sector, hereafter SC) and the *macro-settore concorsuale* (macro-competition sector, hereafter MSC).

The new procedures for the recruitment of faculty members, such as the ASN, are based on SC, each of which includes one or more SSD.

The SSD remain relevant for the appointment of professors, assignment of research funds, teaching contracts, hiring of researchers on limited contracts and university syllabuses.

The present sectors established by Ministerial decree 855 of 30 October 2015 and in force since 20 November 2015 comprise 367 SSD organised into 188 SC, 88 MSC and 14 areas.

It should be noted briefly that this structure based on the SSD does not correspond to that of the European Research Council (ERC), which is the framework of reference for international research. As we shall see, this has a marked impact on issues of evaluation.

The ERC model is in fact based on division into the three disciplinary areas of Social sciences and Humanities, Life Sciences and Physical sciences and Engineering, for each of which there are respectively 6, 9 and 10 panels with 79, 101 and 165 descriptors or keywords.[8] This is the reverse of the situation in Italy, where the SSH constitute the most complex and fragmented area.

SSH research in Italy is in fact covered by CUN areas 10–14 with a total of 34 MSC, 74 SC and 171 SSD, as shown in Fig. 2. Areas 10 and 11 can be seen to present a far larger number of SSD than the average of all the 14 areas (26.4 SSD per area).

This administratively effective but rigid framework has appeared on more than one occasion incapable of representing the complexity of sciences in a constant state of change and development.

As shown below, particularly with respect to the ASN, the danger is that this rigidity may obstruct the growth and transmission of knowledge because this sector-based logic works to the advantage of scholars who adhere strictly to the methodological canons of their discipline and the disadvantage of those who address interdisciplinary questions (Pascuzzi 2014).

[8] https://www.researchitaly.it/uploads/14/ERC_Panel_descriptors_2012.pdf

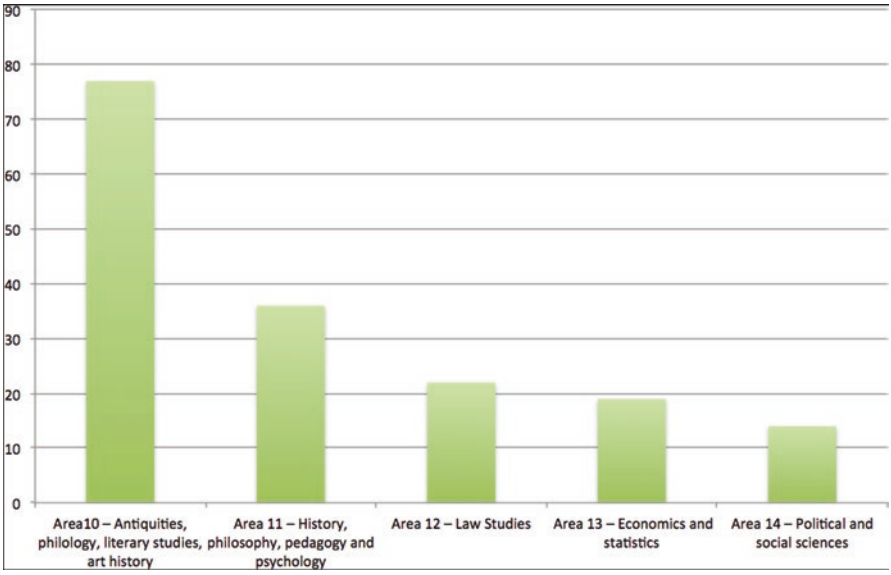


Fig. 2 The breakdown of CUN areas 10–14 (SSH) into SSD

3 Levels of Evaluation and the Point on Which All the Lines Converge

Evaluation has made a marked difference not only to research but also to the everyday activities of universities and faculty. Various examples could be added regarding the monitoring of courses, doctorates and department performance, which are also subjected to specific evaluation procedures.

While it would unquestionably be wrong to say that evaluation is the hub around which university life revolves, as this would mean mistaking the ends for the means employed to measure their attainment, it is certainly true that evaluation constitutes the primary method of assessment for university management today and the key to any in-depth understanding and analysis of how it operates: the point on which all the lines converge.

Failure to distinguish the level at which evaluation is to be applied is one of the major errors to be avoided, and great importance is therefore attached to identification of the object of examination and hence to selection of the right tools for the purpose. As Anthony van Raan, director of the Centre for Science and Technology Studies at Leiden University in the Netherlands, points out, “If there is one thing every bibliometrician agrees, it is that you should never use the journal impact factor to evaluate research performance for an article or for an individual – that is a mortal sin.” (Van Noorden 2010).

It is incorrect to use the same tools to evaluate both departments and individual researchers, just as it is incorrect to apply the same parameters in evaluating both departments at the national level and individual universities.

The different dimensions (university, departments, SSD and individuals) are not susceptible to direct scaling or aggregation in that they respond to different criteria and methods of analysis, which should take into account, among other things, the fact that the phenomena under examination are highly dynamic and often refer to different temporal ranges (Moed and Plume 2011).

One final point to be addressed in this connection before presenting our analysis in detail regards the relationship between individual scholars and the structures to which they belong.

While duties are personal, also as regards the quality of teaching and research, it is hard to evaluate individual responsibilities and merits or demerits without taking into consideration also the operational circumstances in which researchers work. And then there is also the problem of institutional policy in some specific spheres, such as third mission activities, as discussed below.

The assertion that every sphere must be evaluated on appropriate criteria does not refer solely to teaching, research or third mission activities but also to the evaluation of individuals and structures as well as the peculiarities of the various disciplinary areas.

Even though the MIUR and ANVUR have always offered assurances that the VQR activities carried out so far, while based on the evaluation of projects submitted by faculty members, regard departments, the marked impression that the results are not infrequently put to improper use is detrimental both to the evaluatees' peace of mind and to the evalutators' credibility.

The above observations are essential to any understanding of the various objectives of the evaluation activities carried out in Italy over the last few years, the results of which are examined here with respect to the aim of this contribution, namely to assess the applicability of the quantitative tools identified in the previous section in relation to the role still played by the monograph in SSH in Italy.

3.1 The Role of the Monograph for the Subjects Evaluated in VQR 2004–2010

The quality of the research of Italian universities and research centres under the authority of the MIUR is evaluated by the ANVUR through the VQR system. The first VQR regarded the period 2004–2010 and the second the period 2011–2014. VQR involves the evaluation of publications submitted by individuals for the purpose of assessing the departments concerned.

Key importance obviously attaches here to the concept of quality, the criteria for which are identified as *relevance*, *originality* and *internationalisation* in VQR 2004–2010 and *originality*, *methodological rigour* and *expected or potential impact* in VQR 2011–2014. Table 1 presents the definitions of these criteria.

Table 1 Quality criteria of VQR 2004–2010 and VQR 2011–2014

VQR (2004–2010)[a]	VQR (2011–2014)[b]
Relevance	**Originality**
Contributing to the advancement of knowledge in the sector and science in general also in terms of congruity, effectiveness, timeliness and duration of repercussions	Intended as the level of originality with which a product introduces a new way of thinking in relation to the scientific object of research, therefore distinguishing itself from the previous approaches on the same subject
Originality/innovation	**Methodological rigour**
Contributing to the advancement of knowledge or new acquisitions in the sector of reference	Intended as the level with which a product can clearly present the research objectives and the state of the art in literature, adopts a methodology that is appropriate to the research object, and proves that the aims have been achieved
Internationalisation	**Proven or potential impact in the respective international scientific community**
Positioning on the international scene in terms of relevance, competitiveness, circulation and appreciation of the scientific community, including explicit collaboration with researchers and research groups of other countries	Intended as the level to which the product exerts, or is likely to exert, a theoretical and/or practical influence on that same community, based also on its ability to comply with international standards on research quality

[a]ANVUR (2011a), *Valutazione della qualità della ricerca 2004–2010 (VQR 2004–2010): bando di partecipazione*, 7 novembre 2011, http://www.anvur.org/attachments/article/122/bando_vqr_def_07_11.pdf

[b]ANVUR (2015), *Valutazione della qualità della ricerca 2011–2014 (VQR 2011–2014): bando di partecipazione*, 11 novembre 2015, http://www.anvur.org/attachments/article/825/Bando%20VQR%202011-2014_secon~.pdf

The method of evaluation used in VQR 2004–2010 was the informed peer review. A group of experts consisting of scholars of established standing and experience was appointed for each of the 14 CUN areas to evaluate the research projects submitted (3 for each subject) on the basis of the above criteria, combining the peer review wherever possible with bibliometric indicators in order to formulate their judgements.

These concepts are discussed in other chapters in this volume and will therefore not be examined here, also for reasons of space. Suffice it to point out that while there is apparently nothing new in the concept of research quality, it actually takes on a precise meaning with respect to the practical and cognitive objective of the evaluation. The two VQR exercises are in fact based on different concepts of quality.

There was, however, no large consultation between the ANVUR and the scientific community, due to time constraints (but see on this issue the chapter on research quality criteria in this volume, by Bonaccorsi), unlike what happened in the UK with the REF. As the criteria of evaluation had been among the most challenged

aspects of the previous RAE, a series of consultations with panels and the scientific community were held in order to plan the REF (McNay 2003).

VQR (2004–2010) evaluated over 180,000 research products, thus generating a vast amount of data on scientific research in Italy. Specific reports on the quality of research in relation to the structures concerned and their departments were produced for each of the 14 areas and then gathered together in the overall final report.[9]

The greatest legacy of the VQR, and one to which attention should be drawn, is the creation of a new geography of Italian scientific research based on data.

This has been harnessed here to produce a detailed mapping of SSH in Italy through quantitative analysis of the products submitted for VQR 2004–2010 in CUN areas 10–14.[10] The database is called ANVUR 2014 (97%) and it has been anonymised in order to protect confidentiality. Since in 3% of cases the number of researchers in scientific fields in a given department or university is lower than four (the threshold for statistical secrecy), these cases have been omitted and the resulting database represent 97% of subjects. The total number of 49,712 products is distributed as follows:

Area 10 – Antiquities, philology, literary studies, art history: 13,737
Area 11 – History, philosophy, pedagogy and psychology: 9185
Area 12 – Law: 11,565
Area 13 – Economics and statistics: 11,272
Area 14 – Political and social sciences: 3953

As the products were selected and submitted by the scholars themselves, they can be regarded as representing the best of research.

Figure 3 shows that in all the areas examined except for 13, where articles in journals dominate (61.6%), the products primarily took the three forms of articles in journals, monographs or treatises and contributions in books. Of residual importance are conference papers, editing and other items.[11]

In area 10, contributions in books account for most of the products (33.2%), followed by articles in journals (26.3%) and monographs or treatises (22.6%), with conference papers accounting for just 14.30%.

In area 11, contributions in books and monographs or treatises account for practically the same proportion, respectively 33.2% and 32.9%, followed by articles in journals (about 10%) and conference papers (about 7%).

Monographs received the highest rankings on average both in area 10 and in area 11.

[9] http://www.anvur.org/rapporto/

[10] See the contribution by Stella Iezzi in this volume for further details.

[11] The other items include commentary, composition, design, critical esposition, esposition, software, exhibition, aftefact, comment on a verdict, preface, postface, artistic prototype and associated drawings, publication of previously unpublished sources, translation, entry in a dictionary or encyclopaedia, concordance and performance.

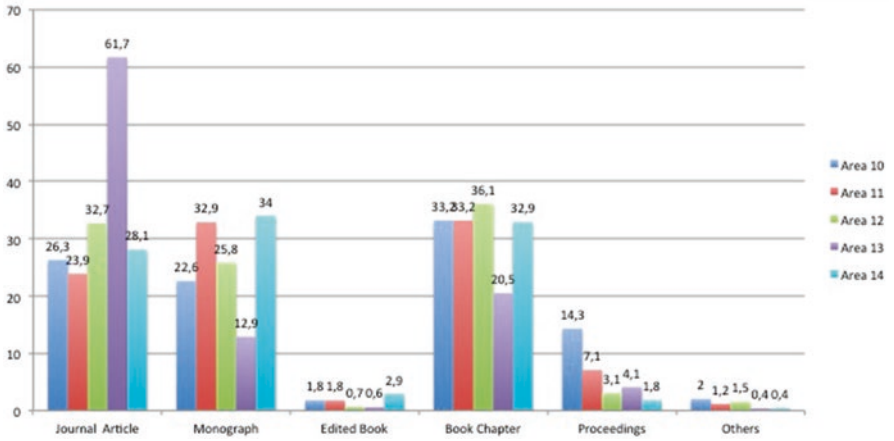


Fig. 3 VQR types of product for CUN areas 10–14 (%) (Source: our elaboration from DB ANVUR 2014 (97%))

In area 12 as in 10, contributions in books account for most of the products (36.1%) and prove most important in terms of evaluation, but with articles in journals only just over 3 points behind (32.7%), followed by monographs or treatises (about 6 points back at 25.8%). Conference papers are of negligible importance (3.1%).

Area 13 presents a very different situation from all the others with articles in journals accounting for well over half of the products (61.6%), followed at a great distance by contributions in books (20.5%), monographs or treatises (12.9%) and finally conference papers (about 4%).

Those submitting products other than articles in area 13 received a much lower evaluation on average.

In area 14 as in 11, monographs or treatises and contributions in books were very similar in importance, accounting respectively for 34% and 32.9% of all products.

Areas 10–14 can be divided into two groups on the basis of the publications submitted for evaluation:

1. The human and social disciplines (areas 10, 11, 12 and 14), where paramount importance attached to the book form with monographs and contributions accounting for the majority of products: 55.6% in 10, 66.1% in 11, 61.9% in 12 and 66.9% in 14.
2. The economic and statistical disciplines (area 13), where articles in journals account for the majority of products, as in the hard sciences.

In fact, close examination of the specific types of publication per SSD reveals a still less homogenous situation. As frequently emerges from the area reports, the administrative logic of the SSD in no way reflects the logic of scholarly

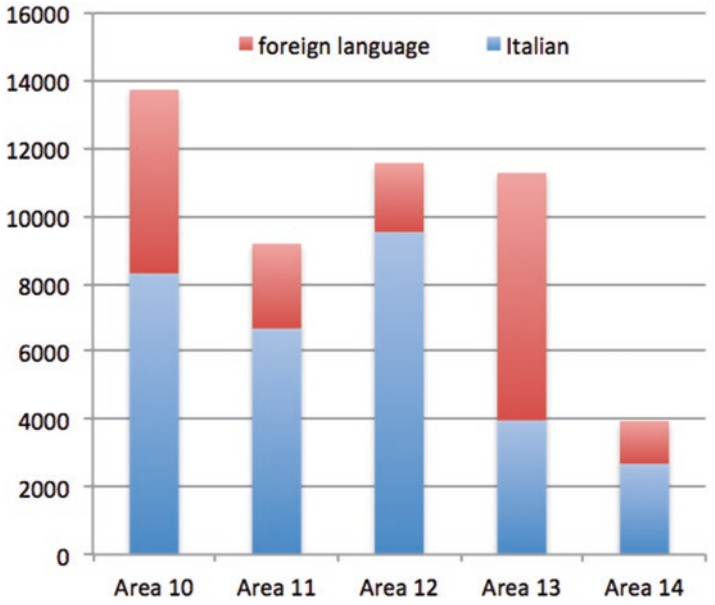


Fig. 4 VQR language of publication by CUN area (areas 10–14) (Source: our elaboration from DB ANVUR 2014 (97%))

communication, which suggests that the SSD should perhaps be grouped differently for the purposes of evaluation.

Consideration of the language used in the different areas (Fig. 4) reveals that Italian is mostly used for publications in 11 (60%), 12 (72%) and 14 (69%), whereas area 13 shows the predominance of another language (63%), which is nearly always English (84.37%).

The present authors believe that VQR 2004–2010 has already given rise to a series of unexpected consequences (Merton 1936). Given that the criterion of internationalisation, for example, is often interpreted as meaning publications in English and that bibliometric indicators have become a goal to be pursued rather than a tool of evaluation, researchers may well attach less importance to the quality of their research than to publishing it in journals with a high impact factor or in English. There is also a risk of concentration on subjects of mainstream research because of their higher citation count to the detriment of healthy thematic and methodological variety.

3.2 The Importance of the Monograph for Evalutators: Parameters and Criteria for the ASN

While VQR is an activity that uses the research products of individuals to evaluate structures, the ASN (for separate SC consisting of one or more SSD) constitutes the requisite qualification in Italy for university teaching posts and thus serves primarily for the evaluation of researchers.

It should be made clear from the outset that the authors have no intention of comparing the two forms of evaluation. Such an act is considered wholly illegitimate, as the VQR and ASN differ not only in their targets – structures and individuals respectively – but also in terms of their temporal span and objects examined (the entire activity of the individual professor and researcher in the ASN and publications in the VQR).

Attention is focused here solely on the importance attached to the monograph by scholars, this time as evalutators rather than evaluatees, which reveals that the criteria adopted, while varying greatly from area to area, confirm the paramount role of the monograph in areas 10–14 (SSH).

The ASN, which expires every 4 years, is awarded on the basis of the analytical evaluation of qualifications, publications and individual contribution to scholarly production as expressed through criteria, parameters and indicators that differ from area to area. Unlike VQR, where the criteria of originality, relevance and so on are the same for all areas, each area establishes its own criteria in the ASN. At the time of writing, the regulations with criteria and parameters for the new ASN have just appeared.[12]

The ANVUR itself recognises that it is impossible to establish criteria in the form of numerical thresholds of access that are valid for all the SC within each of the 14 CUN areas due to the significant and sometimes radical differences in praxis of the SC inside the same area. (ANVUR 2011b).

The document *Criteri, parametri e indicatori per l'Abilitazione scientifica nazionale* (https://www.cun.it/uploads/3974/alt_2011_05_24_n1.pdf) was subjected to automatic text analysis is order to determine the differences existing between the various areas and, in more general terms, between the hard sciences and SSH, as regards the criteria, parameters and indicators to be applied in the selection both of boards of examiners and of candidates.

The corpus is made up of 14 sections, one for each area, for a total of 15,520 tokens and 1570 types, approximately 30% of which appear just once (hapax), with an average of 1108 words per text.

The processing was carried out by means of the open-source program IRaMuTeQ (Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires) (http://www.iramuteq.org/) and the Python programming language

[12] http://www.gazzettaufficiale.it/atto/serie_generale/caricaDettaglioAtto/originario?atto.dataPubblicazioneGazzetta=2016-07-05&atto.codiceRedazionale=16G00130&elenco30giorni=false

developed by Pierre Ratinaud and the Laboratoire LERASS of Toulouse-Le Mirail University, which is capable of performing various types of analysis on large corpora (Ratinaud and Marchand 2012).

The document was subjected to specificity analysis, each positive or negative specificity of a term indicating overuse or underuse (or rarity to the point of total absence) with respect to the "expected" value. This gap is evaluated in probabilistic terms.

Specificity analysis of the terms used in the texts identifying criteria and parameters for the hard sciences (areas 1–9) and SSH (areas 10–14) clearly reveals the difference, the most specific terms for the hard sciences being *ISI*, *citation*, *IF* and *H Index* as against *monograph*, *essay*, *publisher*, *excellence* and *national* for SSH.

The most specific terms for each area are as follows:

Area 1 (Mathematics and computer science): data, citation, value
Area 2 (Physics): value, impact, collaboration
Area 3 (Chemistry): IF, H index
Area 4 (Earth sciences): ISI, H index, journal
Area 5 (Biology): IF, article, catalogue
Area 6 (Medicine): IF, index, bibliometric
Area 7 (Agricultural and veterinary sciences): proceedings, conference
Area 8 (Civil engineering and architecture): innovative, original, project
Area 9 (Industrial and information engineering): international, diffusion, relevance
Area 10 (Antiquities, philology, literary studies, art history): site, essay, publishing
Area 11 (History, philosophy, pedagogy and psychology): monograph, continuity, national
Area 12 (Law): excellence, premise, monograph
Area 13 (Economics and statistics): publisher, methodology
Area 14 (Political and social sciences): essay, relevant, publisher

4 Some Findings Regarding the Evaluation of Monographs in the SSH

As shown by the criteria adopted in the ASN and the results of VQR (2004–2010), the book form (chapters and monographs in particular) still constitutes the primary channel of communication in SSH.

This accounts for the widespread interest regarding the applicability of some quantitative (bibliometric) tools for the evaluation of books in SSH.

On the one hand, the evaluation of the quality of a monograph is impossible with bibliometric methods and necessitates a peer review, a procedure (applied not only in *ex-post* evaluation but also *ex-ante* for the acceptance of contributions submitted for publication) carried out by experts and specialists in the sector. On the other hand, in order to avoid any risk of being totally subjective or indeed arbitrary, this

evaluation should be bolstered through prior definition of the criteria to be applied and the support of data obtainable through bibliometric measurement.

The debate on these questions is quite lively and appears to be moving towards a combination of bibliometric measurement and peer review.

As previously stated, the quantitative tools referred to in this connection are as follows:

1. Coefficients for weighting monographs by type and size
2. Evaluation criteria for the reputation and classification of publishers and series of publications
3. Systems for the use of reviews for evaluation purposes and combining them with other methods of impact evaluation
4. Library catalogue analysis (Torres Salinas and Moed 2009)

This section will draw some conclusions as regards the first three aspects. For the fourth, readers are referred to the study by Maria Teresa Biagetti, Antonella Trombone and Antonella Iacono in this volume.

Coefficients for the Weighting of Monographs by Type and Size

The data presented above confirm that the book is still the primary channel through which scholars of SSH communicate the results of their research and are evaluated. As shown by analysis of the ASN criteria, the publication of a monograph, essay or chapter of a book is still one of the key elements for the qualification of researchers in SSH.

This importance of the monograph is also the reason for the repeated demands for it to be weighted more substantially with respect to other publications, not least in recognition of the kind and amount of commitment entailed by comparison with articles and other contributions.

VQR 2011–2014 has responded in the following terms: "Monographs of a scholarly character and their equivalents can count double if so decided by their authors, in which case they will also count double in the final evaluation of the Institution." (ANVUR 2015, p. 10).

In the previous version (3 September) of the call for submissions, the text went on to describe this as an innovation of VQR 2011–2014 designed to do justice to the effort undertaken by the author(s) of a research monograph.

Evaluation Criteria for the Reputation and Classification of Publishers and Series of Publications

Great importance is also attached for monographs to how and where they are published (Giménez-Toledo and Román-Román 2009; Giménez-Toledo et al. 2013).

As emerged with respect to the ASN, the publishing house is regarded as one of the primary criteria of quality.

A natural point of reference in the literature is the Spanish SPI project (Scholarly Publishers Indicators: books in humanities and social sciences) of the CSIC Grupo de Investigación sobre el Libro Académico (ÍLIA), designed to examine the quality of Spanish and non-Spanish scholarly publishing in terms of the "prestige" of the individual publishers. The data obtained from a questionnaire completed by 11,864

researchers were processed by means of the group's ICEE indicator (Indicador de calidad de editoriales según los expertos) (Giménez-Toledo 2016).

With the recent addition of interactive graphics, the Scholarly Publishers Indicators Expanded is a new resource providing additional information on analysis of the coverage of Spanish and non-Spanish publishers through the inclusion or otherwise of 3948 scholarly publishers in the various international information systems, comprising the Book Citation Index (Thomson Reuters), Scopus book titles (Elsevier), the Norwegian List (CRISTIN), SPI and the Finnish List.

Another interesting development is the GPRC (Guaranteed Peer Reviewed Content) certificate of the Flemish Publishers Association (VUV), developed by the publishers themselves to ensure the quality of their publications.

Publishers applying for GPRC certification must be in a position to prove that the publications in question have been positively evaluated in a peer review process based on internationally accepted standards and agreed upon with the Flemish inter-University Council (VLIR). With respect to the SPI, attention is focused in this case on the quality of the work rather than the prestige of the publisher.

No such experimentation is under way in Italy at present. While criteria for the classification and evaluation of journals are beginning to take shape, there appears to be little likelihood of anything similar being achieved for publishers and series of books.

The various reasons for this include the present difficulties of the publishing market and the lack of funds for research, which often cause scholars to settle for lesser forms of publication, the fact that research monographs do not sell well and decisions as regards publishing are often based on the availability of financial contributions rather than the publisher's prestige, and the refusal of major publishers to base their acceptance of works for publication on the decisions of boards of consultants. A leap forward is required. Greater recourse to open-access publication would certainly help to solve the problem while also fostering the full and free use of research findings.

Systems for the Use of Reviews in Evaluation

In addition to a survey of the relevant international literature and best practices in other countries (Zuccala and van Leeuwen 2011; Zuccala et al. 2014), the use of reviews in evaluation was examined on the basis of data from the above-mentioned SUA-RD dossiers for the period 2011–2013, where researchers were asked to indicate any reviews of their publications.

The exploration has proved largely unfruitful so far, not only because the data are very incomplete (details of where the reviews appeared are missing, for example) and therefore unreliable but also because few reviews are in fact mentioned.

Moreover, there are a number of important points to be decided upon, including precise definition of the review and the difference between it and a simple mention, the weight or prestige of the journal in which the review appears, the scholarly quality of the reviewer and the nature of the review, which can obviously be positive, negative, critical or problematic.

5 Some Open Questions

As shown above, the attempt to identify applicable quantitative tools for the evaluation of monographs, which constitute the primary type of publication in the SSH, has so far proved somewhat fruitless with the means currently available.

In conclusion, attention should be drawn to two questions that are still open and in need of careful consideration, namely the delicate relationship between the evaluation of research and the evaluation of third mission activities in SSH and the risk of unexpected consequences arising from the application of unsuitable tools.

5.1 Evaluation of Research vs. Evaluation of Third Mission Activities in the SSH

The subjects addressed in some areas of SSH are of general interest and not confined to specialists, such as historical research, communication, current events in the cultural and political spheres, the dynamics of contemporary society and issues regarding economic development.

Scholarly rigour is not extraneous but rather essential to the work of informing and guiding public opinion. The possibility of participation in public life and the very concept of active citizenship depend on the radical difference between mass-media chatter and quality information.

This constitutes the third mission of universities, which are responsible not only for scholarly research and teaching but also for disseminating the knowledge developed and produced in the academic world more extensively.

The third mission involves engagement

> in the construction of scientific citizenship, which means greater awareness of citizens as regards scientific research and technological development as well as greater participation in technical and scientific choices, including those of an environmental and 'ethically sensitive' nature. It also means greater economic democracy. If knowledge is now the primary factor of economic growth, constructing scientific citizenship (also) means ensuring that it does not also become a factor of new social exclusion but rather an active factor of social inclusion. In practice, it means the university opening up and presenting itself as a new agora, one of the forums of participative democracy (where citizens gather to obtain information, discuss matters and take decisions) and economic democracy, where it is not only the big corporations that obtain knowledge for innovation but all the citizens that acquire the knowledge necessary for their well-being, for social integration and even for bottom-up entrepreneurship (Greco 2010, p. 9).

The tendency established so far in the field of research evaluation is to separate the quality of scholarly studies completely from their impact on society. While these are obviously different things to be evaluated with different and appropriate methods, scarce consideration for the third mission and its civic function has led in many cases to a view of it as inferior in status and something to be avoided by scholars of any distinction.

The consequence has been serious underestimation of the importance of transferring the results of research, often misinterpreted as no more than simple technological transfer, on a broader scale. Consider, for example, the relationship between current research and school textbooks, which sometimes lag so far behind as to be almost embarrassing and are no more than an assortment of already obsolete formulations and clichés.

Without wishing to overstate the consequences of this attitude, the authors regard it as one of the reasons why the academic world is held in such low regard, often accused of being wholly self-referential and out of step with the requirements of the society outside it. Greater attention to the third mission, which also has an undeniable local dimension, would instead help to embed universities in their local communities.

If the aim is to stimulate and support the involvement of the best-qualified scholars in the dissemination and transmission of knowledge, then it is essential to promote the publishing products and initiatives that seek to meet these needs.

5.2 Unexpected Consequences

Production in the field of SSH has rested for decades on consolidated practice and principles such as a primary focus on the monograph, use of the national language and a tendency to address the national scholarly community first and foremost.

It is hardly surprising that the transformations under way as regards tools (online work helps to expand horizons and foster greater responsiveness to the international scene) and an undeniable tendency towards greater productivity and swifter circulation of the results of research even before they are fully consolidated (with the resulting shift from the monograph to the article in a journal) should have led to radical changes also in the behaviour of scholars.

These are positive in many cases and to be welcomed as changes that will help to boost the quality of the studies carried out in Italian universities.

It is, however, important to distinguish between these and the opportunistic attitudes that could spread among younger scholars – obviously more sensitive to the effects of evaluation, as their careers depend on it – as a result of the adopted evaluation criteria.

One example is slicing the results of research projects into a number of articles instead of a single monograph, thus sacrificing the complex and organic exposition that is in itself a hallmark of scholarly maturity. Another is the mistaken belief that internationalisation simply means publication in English or in foreign journals, which does not necessarily correspond to greater involvement in the international debate.

Those involved in evaluation, including universities and agencies as well as individuals, must be fully aware of their responsibilities and of the effects, desired and undesired, of according priority to one criterion rather than another and thereby inducing scholars to alter their behaviour and adopt unnatural attitudes within their disciplinary sphere.

References

ANVUR. (2011a). Valutazione della qualità della ricerca 2004–2010 (VQR 2004–2010): bando di partecipazione. 7 novembre 2011, http://www.anvur.org/attachments/article/122/bando_vqr_def_07_11.pdf

ANVUR. (2011b). Criteri e parametri di valutazione dei candidati e dei commissari all'abilitazione scientifica nazionale. 22 giugno 2011, http://www.anvur.org/attachments/article/44/documento01_11.pdf

ANVUR. (2015). Valutazione della qualità della ricerca 2011–2014 (VQR 2011–2014): bando di partecipazione. 11 novembre 2015, http://www.anvur.org/attachments/article/825/Bando%20VQR%202011-2014_secon~.pdf

Baccini, A. (2010). *Valutare la ricerca scientifica. Uso e abuso degli indicatori bibliometrici*. Bologna: il Mulino.

Banfi, A., Franzini, E., & Galimberti, P. (2014). *Non sparate sull'umanista. La sfida della valutazione*. Milano: Guerini e Associati.

Bradford, S. C. (1934). Sources of information on specific subjects. *Engineering, 137*(3550), 85–86.

De Bellis, N. (2014). *Introduzione alla bibliometria. Dalla teoria alla pratica*. Roma: AIB.

Faggiolani, C. (2015). *La bibliometria*. Roma: Carocci.

Figà Talamanca, A. (2000). *L'impact factor nella valutazione della ricerca e nello sviluppo dell'editoria scientifica,* IV Seminario Sistema Informativo Nazionale per la Matematica SINM 2000: "un modello di sistema informativo nazionale per aree disciplinari". Lecce, 2 ottobre 2000. http://www.roars.it/online/limpact-factor-nella-valutazione-della-ricerca-e-nello-sviluppo-delleditoria-scientifica/

Garfield, E. (1979). Is citation analysis a legitimate evaluation tool? *Scientometrics, 1*(4), 359–375.

Giménez-Toledo, E. (2016). Assessment of journal & book publishers in the humanities and social sciences in spain. In M. Ochsner, S. E. Hug, & H. Daniel (Eds.), *Research assessment in the humanities. Towards Criteria and Procedures*. Cham: Springer Open.

Giménez-Toledo, E., & Román-Román, A. (2009). Assessment of humanities and social sciences monographs through their publishers: A review and a study towards a model of evaluation. *Research Evaluation, 18*(3), 201–213. https://doi.org/10.3152/095820209X471986.

Giménez-Toledo, E., Tejada-Artigas, C., & Mañana-Rodriguez, J. (2013). Evaluation of scientific books' publishers in social sciences and humanities: Results of a survey. *Research Evaluation, 22*(1), 64–77. https://doi.org/10.1093/reseval/rvs036.

Greco, P. (2010). Una "terza missione" per l'Università. *Scienza e società, 9*(10), 7–10.

Hammarfelt, B. (2016). Beyond coverage: Toward a bibliometrics for the humanities. In M. Ochsner, S. E. Hug, & H. Daniel (Eds.), *Research assessment in the humanities. Towards criteria and procedures*. Cham: Springer Open.

Hicks, D. (2004). The four literatures of social science. In H. F. Moed, W. Glänzel, & U. Schmoch (Eds.), *Handbook of quantitative science and technology research: The use of publication and patent statistics in studies of S&T systems* (pp. 476–496). Dordrecht: Kluwer Academic Publishers.

Hicks, D., et al. (2015). The Leiden Manifesto for research metrics. *Nature, 520*(520), 429–431.

McNay, I. (2003). Assessing the assessment: An analysis of the UK research assessment exercise, 2001, and its outcomes, with special reference to research in education. *Science and Public Policy, 30*(1), 47–54.

Merton, R. K. (1936). The unanticipated consequences of purposive social action. *American Sociological Review, 1*(6), 894–904.

Moed, H. F., & Plume, A. (2011). The multi-dimensional research assessment matrix. *Research trends, 23*, 5–7. http://www.researchtrends.com/issue23-may-2011/the-multi-dimensional-research-assessment-matrix.

Pascuzzi, G. (2014). Soldatini e danni collaterali: i settori scientifico-disciplinari, *ROARS – Return on Academic ReSearch*, 18 Gennaio 2014, online su http://www.roars.it/online/soldatini-e-danni-collaterali-i-settori-scientifico-disciplinari/

Pinto, V. (2012). *Valutare e punire. Una critica della cultura della valutazione*. Naples: Cronopio.

Ratinaud, P., & Marchand, P. (2012). Application de la méthode ALCESTE à de 'gros' corpus et stabilité des 'mondes lexicaux': analyse du 'CableGate' avec IraMuTeQ, in *Actes des 11eme Journées internationales d'Analyse statistique des Données Textuelles. Presented at the 11eme Journées internationales d'Analyse statistique des Données Textuelles*, JADT, Liège, pp. 835–844.

Reale, E. (2008). *La valutazione della ricerca pubblica. Un'analisi della valutazione triennale della ricerca*. Milan: Franco Angeli.

Torres Salinas, D., & Moed, H. F. (2009). Library catalog analysis as a tool in studies of social sciences and humanities: An exploratory study of published book titles in economics. *Journal of Informetrics, 3*(1), 9–26.

Van Noorden, R. (2010, June 16). Metrics: a profusion of measures. *Nature, 465*, 864–866. https://doi.org/10.1038/465864a.

Zuccala, A., & van Leeuwen, T. (2011). Book reviews in humanities research evaluations. *Journal of the American Society for Information Science and Technology, 62*(10), 1979–1991. https://doi.org/10.1002/asi.21588.

Zuccala, A., van Someren, M., & van Bellen, M. (2014). A machine-learning approach to coding book reviews as quality indicators: Toward a theory of megacitation. *Journal of the American Society for Information Science and Technology, 65*(11), 1643–2330. https://doi.org/10.1002/asi.23104.

Guidelines for Peer Review. A Survey of International Practices

Andrea Capaccioni and Giovanna Spina

1 Introduction

In the last few years, research assessment has been progressively spreading in universities and research centres all over the world. This process sparked a lively debate, the implications varying considerably across scientific disciplines. In Natural Sciences, the debate focused on the identification of the most effective instruments and methodologies for analysing scientific production, whereas in Social Sciences and Humanities (SSH), the discussion focused on which tools should be used, namely peer review vs. bibliometric instruments.

The San Francisco Declaration on Research Assessment (DORA 2012), presented in December 2012 by the American Society for Cell Biology (ASCB) and signed by various scientific societies, confirms the international relevance of the debate in this field. The declaration questions the use of bibliometric instruments, such as the Impact Factor, proposing instead the use of a wider range of bibliometric indices together with qualitative assessment instruments. Recently, a group of scholars from different countries returned to those issues, promoting the *Bibliometrics: Leiden Manifesto for research metrics* (Hicks et al. 2015), with the aim of elaborating 10 fundamental principles on the use of bibliometric indices for assessment, in order to avoid the assessment system being "damaged from those same instruments devised and designed to improve it" (Faggiolani 2015, p. 429). This renewed interest in qualitative methods has naturally triggered an updating process of the peer review, whose limitations and critical elements have been known

A. Capaccioni (✉)
Dipartimento di Lettere-Lingue, Letterature e Civiltà antiche e moderne, Università degli Studi di Perugia, Perugia, Italy
e-mail: andrea.capaccioni@unipg.it

G. Spina
Università per Stranieri di Perugia, Perugia, Italy
e-mail: spinaj@hotmail.it

© Springer International Publishing AG 2018

A. Bonaccorsi (ed.), *The Evaluation of Research in Social Sciences and Humanities*, https://doi.org/10.1007/978-3-319-68554-0_3

for a long time, through the introduction of methodologies that integrate the classical peer review processes with citation or bibliometric analysis, or the use of *informed peer review*, which accounts for different assessment elements (Moed 2007; Neufeld and Von Ins 2011; Abramo and D'Angelo 2011; Cicero et al. 2013; Wilsdon et al. 2015; Bertocchi et al. 2015; Baccini and De Nicolao 2016).

Through the analysis of authoritative documents (reports and guidelines), the present chapter aims to show how the European academic and research environments are contributing to rethinking the main methods for assessing research products, particularly within the SSH. The documents refer to the peer review methodology for research assessment, in some cases for the *ex ante* project selection, in others for the *ex post* assessment of research products.

The final section of the chapter is dedicated to some considerations on the most significant ways in which the peer review system has been used in recent assessment exercises carried out in Italy, presenting a few proposals for the future.

2 The European Approach to Peer Review

Europe has been expressing the need to identify the most appropriate ways to conduct an adequate qualitative assessment, and this has encouraged the elaboration of guideline proposals for the management of peer review, coming from both academia and the scientific publishing sector (Turner 2009; Research Information Network (RIN) 2010; Nicholas and Gordon 2011; Committee on publication ethics (COPE) 2013). The present document includes a selection of recent contributions, with a specific – albeit not exclusive – focus on the activities aimed at assessing scientific research products.

The guidelines on peer review released by the European Science Foundation (ESF) in 2011 have the desired characteristics. The *European Peer Review Guide* presents itself clearly as an instrument "to benchmark national peer review processes and to support their harmonisation" (ESF 2011, p. 3). The document identifies seven fundamental principles that need to guide any peer review activity (see Fig. 1).

For the present analysis, Chap. 4 is of major importance (*Peer review methodology*, p. 17–45) as it defines the ways in which effective peer review efforts need to be conducted (Fig. 2).

At this stage, this chapter will focus on some aspects only. Great importance is to be attributed to preparation before the assessment: in this phase, human and financial resources have to be carefully organised, together with the instruments and technical means needed for the refereeing activity, the management of the different phases of peer review, and the preparation of all necessary information (guidelines, reports, etc.) for the reviewers' use, but also for those who are evaluated. A large amount of space is left to the reviewers' decisions, and the eligibility criteria for the examined publications:

1. Excellence	Projects selected for funding must demonstrate high quality in the context of the topics and criteria set out in the calls. The excellence of the proposals should be based on an assessment performed by experts. These experts, panel members and expert peer reviewers should be selected according to clear criteria and operate on procedures that avoid bias and manage conflicts of interest.
2. Impartiality	All proposals submitted must be treated equally. They should be evaluated on their merits, irrespective of their origin or the identity of the applicants.
3. Transparency	Decisions must be based on clearly described rules and procedures that are published *a priori*. All applicants must receive adequate feedback on the outcome of the evaluation of their proposal. All applicants should have the right to reply to the conclusions of the review. Adequate procedures should be in place to deal with the right to reply.
4. Appropriateness for purpose	The evaluation process should be appropriate to the nature of the call, the research area addressed, and in proportion with the investment and complexity of the work.
5. Efficiency and speed	The end-to-end evaluation process must be as rapid as possible, commensurate with maintaining the quality of the evaluation, and respecting the legal framework. The process needs to be efficient and simple.
6. Confidentiality	All proposals and related data, intellectual property and other documents must be treated in confidence by reviewers and organisations involved in the process. There should be arrangements for the disclosure of the identity of the experts.
7. Ethical and integrity considerations	Any proposal which contravenes fundamental ethical or integrity principles may be excluded at any time of the peer review process.

Fig. 1 Fundamental principles of peer review (Figure from ESF (2011, p. 13))

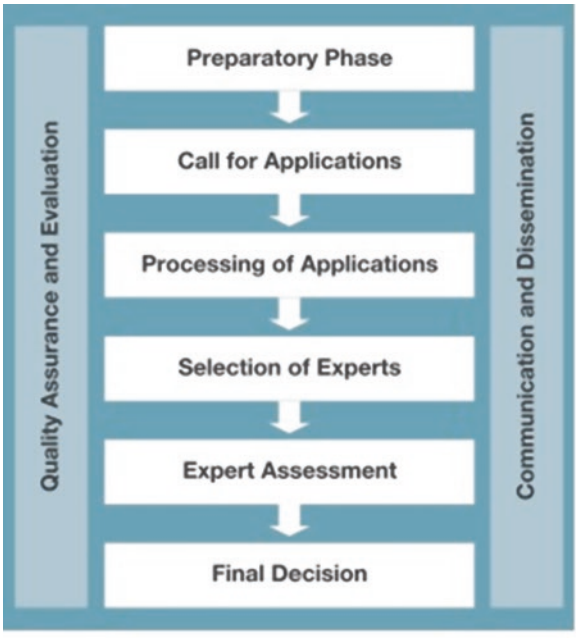


Fig. 2 Description of an effective peer review process (Figure from ESF 2011, p. 17)

> One of the most important and challenging phases of the peer review process is to collect the required number of willing and available experts who would agree to conduct the task of expert assessments both as individual/remote reviewers and/or members of panels and committees as described below (ESF 2011, p. 22).

It is clear that the *European peer review guide* considers the reviewers as key figures in the assessment process, and maintains that the quality of refereeing depends on paying serious attention to their selection and monitoring all the phases of their work. As far as this step is concerned, the ESF guidelines introduce a few brief reflections on a possible contribution from bibliometrics in peer review activities:

> A solid record of publications: bibliometric indices are increasingly used for assessing publication track records. Care should be taken when applying these quantitative measures; these must be used as complementary information and not as sole determining factors in valuing publication track records. An authoritative and elaborate set of recommendations on the usage of bibliometrics in peer review and evaluation is provided in a ministerial report prepared by the French Academy of Sciences (ESF 2011, p. 25).

The use of bibliometric indices should therefore be complementary, not exclusive, in the elaboration of the product's assessment.

The *European peer review guide* not only offers valid solutions that contribute to an effective assessment activity, but it also represents an important instrument for summarising the debate that has taken place in Europe in the first decade of the new century. As for the subsequent years, two documents that appear to be of particular importance have been selected, always following the aforementioned criteria:

- Institut de France, Académie des sciences, *Les nouveaux enjeux de l'édition scientifique* (2014).
- Research Information Network (RIN), *Scholarly communication and peer review: the current landscape and future trends* (2015).

The first aspect that requires attention is that all documents express a strong belief that peer review needs to be considered as a fundamental instrument for assessment activities, that is in no way surpassed by quantitative methodologies. This belief is perfectly in line with some important contributions on the topic, asserting how the scientific world community still considers this assessment method as essential.

> Most researchers – 70 percent – are happy with the current peer-review process; a satisfaction rate higher than those recorded in similar 2007 and 2009 surveys. When asked if peer review helps scientific communication, 83 percent of those we surveyed agreed, with comments such as, "I have had reviews that were very insightful. When researchers get their nose caught in the lab book, we cannot see the forest through the trees. Having a peer look at your science helps expand the overall view" (Mulligan and van Rossumhttp 2014; Mulligan et al. 2013).

In the report issued by the Académie Des Sciences in 2014, the section dedicated to the *Qualité scientifique*, specifically declares "la nécessité d'un examen critique (*peer-review*) des articles par des chercheurs académiques doit être réaffirmée et leur participation à la prise de décision finale doit être encouragée" and that "l'utilisation du facteur d'impact, inappropriée pour l'évaluation des chercheurs, doit être à nouveau dénoncée" (p. 4); and also, "la nécessité du jugement critique

(peer-review) par des chercheurs académiques doit être défendue. Le rôle éditorial des sociétés savantes doit être réactivé" (p. 11). Among other issues, the French report also addresses the relationship between peer review and bibliometrics, referring to a previous document, developed by the same Académie des sciences, and quoted also in the *European Peer Review Guide,* entitled *Du bon usage de la bibliometrie pour l'évaluation individuelle des chercheurs* (2011).

The second chapter of the 2014 document (*Place respective de la bibliométrie et de l'évaluation qualitative par les pairs),* after analysing the possibilities and limits of the systems, both quantitatively and qualitatively, proposes an approach that takes both methods into account.[1] The report states that the assessment of scientific publications needs to be mostly qualitative ("l'évaluation fondamentale doit être qualitative"). However, given the limits of peer review,[2] the report suggests a careful use of quantitative methods:

> En bref, il convient parallèlement d'améliorer, lorsque cela est nécessaire, l'évaluation qualitative en éliminant en particulier les risques de conflit d'intérêt et d'incompétence des évaluateurs, et d'y intégrer la bibliométrie la plus pertinente possible dans les disciplines où cela s'y prête (Académie des sciences 2011, p. 11).

For the purposes of our analysis, attachments 2 and 3 are particularly interesting. Attachment 2 (p. 32–37) outlines the recommendations on the assessment of university professors as included in the above mentioned 2009 Report, which are based upon three key principles: "compétence, transparence et éthique" (Académie des sciences 2011, p. 32). The attachment includes: the code of ethics (duration of the reviewer's term, use of external reviewers, promotion of assessment criteria, etc.); the criteria and the instruments needed in research activities (together with the directions on where to apply the qualitative and quantitative assessment); the criteria for the assessment of teaching activities: the assessment of scientific and administrative assignments; the frequency and the form of assessment activities; the main characteristics of the reviewers; and the criteria for a correct use of the assessments. Attachment 3 is dedicated to the analysis of bibliometric practices according to different disciplines (p. 38–49). The paragraph about bibliometrics within human and social sciences is of particular relevance here. In fact, in this section French academics refer to human and social sciences, clearly stating that:

> Il n'est pas possible et il ne sera pas possible à brève échéance de réaliser des calculs de citations de qualité suffisante pour qu'ils soient utilisés dans un processus d'évaluation, quel qu'il soit. Les divergences énormes rencontrées pour des calculs faits sur les mêmes chercheurs suivant la base de données utilisée (ISI Web of Science ou Google Scholar) en

[1] The report indicates another relevant document, see Institut de France, Académie des sciences, *L'évaluation individuelle des chercheurs et des enseignants-chercheurs en sciences exactes et expérimentales.* 8 juillet 2009, http://www.academie-sciences.fr/activite/rapport/rapport080709.pdf

[2] Amongst the critical issues of peer review, the following can be highlighted: "les manquements fréquents concernant la qualité des évaluateurs, leur éthique personnelle, l'objectivité, la transparence, la qualité de leurs évaluations, et l'analyse trop superficielle des travaux examinés liée en partie au nombre excessif d'expertises demandées aux meilleurs scientifiques", cfr. Académie des sciences, *Du bon usage de la bibliometrie,* cit., p. 9–10.

> témoignent. Ces bases sont à la fois trop peu et trop inclusives et ne peuvent pas refléter efficacement l'activité scientifique en SHS – ce que confirment d'ailleurs les responsables de Thomson Reuters (p. 49).

This strong refusal derives from specific considerations about the nature of a few selected types of research products:

> Plusieurs raisons l'expliquent: des formes de publication nombreuses en dehors des articles, avec, pour les ouvrages, une quasi-impossibilité de tracer des frontières entre ouvrages scientifiques et "grand public"; une quantité totale de publications relativement faible pour chaque chercheur professionnel mais énorme pour chaque discipline ; des publications qui ne se font pas seulement en anglais (ni en français), pour des raisons légitimes; des citations plus nombreuses d'articles " anciens" (avant 2000, voire avant 1980). Il serait donc absurde d'utiliser des indicateurs bibliométriques pour l'évaluation individuelle en SHS – a fortiori pour le recrutement, point nodal des difficultés des SHS en matière d'évaluation (p. 49).

At this point, the document proposes a few suggestions in order to improve individual assessment, also regarding bibliographic and bibliometric aspects:

- standardiser les présentations des CV des chercheurs (discipline par discipline), quel que soit le type d'évaluation auquel ils doivent servir et l'institution concernée, en distinguant en particulier les publications soumises à l'évaluation par des pairs et les autres, y compris en matière d'ouvrages;
- comme cela se fait pour les revues, faire labelliser et le cas échéant subventionner par le CNRS (labellisation par les sections du CoNRS) les collections d'ouvrages scientifiques;
- encourager la création d'un portail web de signalement des comptes rendus en SHS, à l'échelle française ou de préférence européenne, voire mondiale (projet partiellement en cours sous le nom de recensio.net);
- favoriser la présence dans les jurys d'au moins deux scientifiques issues d'autres disciplines (des SHS, voire hors SHS lorsque le profil des candidats le justifie), qui peuvent amener les "pairs" au sens plus strict à expliciter leurs critères de jugement et éviter certains phénomènes de favoritisme. Les recommandations générales proposées dans le présent rapport s'appliquent également aux SHS (p. 49).

The 2015 report entitled *Scholarly communication and peer review*, issued by the Research Information Network, can be considered one of the most informed and up-to-date contributions on the issue of peer review in relation to assessment practices, based also on interviews with some of the actors of assessment activities (professors, researchers, publishers, etc.) In line with the above-mentioned reports, the document aims to show how the academic and the publishing sectors consider peer review as a fundamental instrument within the scientific publishing sectors. "Peer review plays a central and critical role in the systems of publishing and communicating research results, from the perspectives of researchers and also of publishers" (RIN 2015, p. 2). First of all, the report offers an analysis of the main purposes of peer review:

> The first and fundamental purpose is to check the work for 'soundness': has the research been performed to reasonable standards, and are the findings and the conclusions drawn from them in that sense valid? A second purpose is to assess the originality, significance and broader interest of the research: is it simply reporting another example of a well-known finding or phenomenon, or is it new and of wider significance and interest. Third, reviewers are asked to assess the 'fit' with the journal to which a paper has been submitted: is the

> paper likely to be of interest to the readers of this particular journal? A fourth purpose is to help authors to improve the quality of their research and the presentation of their findings: how might the paper be enhanced so that it maximises the credit and rewards that the authors (and the journal) receive as a result of its publication? (p. 6–7)

This is followed by a short evaluation of the role played by journal directors, editors, proofreaders and publishers (p. 8–10) in the supervision process. The passage also includes an overview of the most criticised areas in peer review, often considered as a system "based on human endeavour" (p. 10–14). In this paragraph, the following aspects are taken into consideration: *efficacy*, *burdens on reviewers* (on the ever-increasing workload for reviewers), *expense and delay* (on the costs of peer review), *unfairness and bias* (ethical issues), *publication bias* (aspects related to the selection of publications), *subversions of the peer review system* (examples of peer review gone wrong). The subsequent paragraph examines types of peer review (p. 14–18) and, after a brief description (*single-blind review*, *double-blind review* and *open review*), it offers an in-depth analysis of the relationship between authors and reviewers, introducing a few examples of new methods, i.e.:

> The new journal *eLife* has introduced a similar kind of discussion forum for reviewers, who again submit their reviews in the normal way, but then engage in a collective dialogue with an editor about their reviews (which are revealed to each other) in order to reach an agreed view (p. 16).

This section is followed by some brief reflections on the review of those articles that contain a large amount of data, and by a paragraph dedicated to various recent experiences of *post-publication reviews and comments* (p. 19–22). In particular, it is important to mention some of the platforms "[…] from annotatr (http://annotatr.appspot.com/), through Epistemio (http://www.epistemio.com/) and PaperCritic (http://www.papercritic.com/) to PeerLibrary (https://peerlibrary.org/) where researchers can write and read reviews of published papers" (p. 20), and *post-publication metrics* (p. 22) such as Altmetric (http://www.altmetric.com/), Plum Analytics (www.plumanalytics.com) and Impact Story (https://impactstory.org/).

The document highlights the importance of those instruments that provide information to authors and reviewers, such as guidelines, a clear publishing policy, but also tutorials and other material. It also addresses the issue of remuneration for reviewers, and presents a few examples of *independent peer review platforms* (p. 22–27).

In its conclusions the RIN report insists that nowadays, peer review is to be considered a bedrock of scientific communication, aimed at both academic and publishing sectors. In spite of criticism, review activities have strengthened since the end of World War II, also thanks to the growth of research production. The review activity is able to counter phenomena such as plagiarism and a general decrease in the scientific level of contributions. At the same time, it should aim at guaranteeing an appropriate quality standard, avoiding excessive rigidities, such as those towards innovative or multidisciplinary approaches. In the last few years, some innovative strategies have emerged, which explore mainly web-based approaches and technical methods, able to improve effectiveness, efficacy and fairness of peer review.

According to the report, a significant input in this direction should be coming from a stronger integration between publishing staff, reviewers and authors, and from the direct involvement of scientific communities through the use of social media, with *post-publication review* (compared to the traditional *pre-publication review*), which are still being experimented and are strongly oriented towards single articles (*article-level metrics*).[3] In this way, the role of the reviewer can also acquire a more precise definition, for instance through greater visibility of their activity in academic CVs, or through a remuneration scheme where possible.

2.1 The Italian Context

Drawing on what has been presented above, it is important to provide a few considerations – albeit as an interim measure – which could be implemented to strengthen both review and qualitative assessment efforts.

The previous sections explored how international (mainly European) communities have in recent times begun to reconsider the qualitative method for use in the assessment of research products, and to rethink its main implementation procedures. This final section intends to examine how peer review has been used in the most recent assessment exercises carried out in Italy, and to submit a few proposals.

It has been established that the first step to a correct evaluation of research products consists in preparing an effective and – most importantly – transparent organisation of the review activities. This makes it possible to select the most appropriate criteria in determining the quality of scientific results. In its 2011–2014 VQR report (Valutazione della Qualità della Ricerca – Research quality assessment), ANVUR (the Italian National Agency for the Evaluation of Universities and Research institutes) decided to establish a few general criteria and to assign the elaboration of specific criteria for the respective disciplinary fields to the GEV (Gruppo di esperti della Valutazione – Assessment Experts Group), (see ANVUR 2015a, p. 14–16). With the same purpose, the report also offers a series of guidelines formulated by the Assessment Experts Group (ANVUR 2015b). As far as the non-bibliometric sectors (Social Sciences and Humanities, Areas 10–14) are concerned, the proposed methodology consists in a peer review that is "assigned to external and independent experts selected by GEV (usually two for each research product), whose task is to express an anonymous opinion on the quality of the selected publications" (ANVUR 2015a, p. 14). It is important to point out that the Agency itself proposed that half of the overall number of products in all disciplines (plus one) should be evaluated

[3] On the topic of *pre/post-publication reviews*, see the interesting report issued by the House of Commons: House of Commons Science and Technology Committee, *Peer review in scientific publications: eighth report of session 2010–12*, Vol. I. 18 July 2011, p. 66–67. The whole report constitutes a synthesis of the issues and possibilities of peer review, which are summarised in p. 88–94, http://www.parliament.uk/business/committees/committees-a-z/commons-select/science-and-technology-committee/inquiries/peer-review/

through peer review. Furthermore, reviewers were asked to give an anonymous qualitative judgement, assigning one of the following levels of quality to each product:

(a) innovativeness, intended as the level of originality with which a product introduces a new way of thinking in relation to the scientific object of research, therefore distinguishing itself from the previous approaches on the same subject;
(b) methodological rigour, intended as the level with which a product can clearly present the research objectives and the state of the art in literature, adopts a methodology that is appropriate to the research object, and proves that the aims have been achieved;
(c) proven or potential impact in the respective international scientific community, intended as the level to which the product exerts, or is likely to exert, a theoretical and/or practical influence on that same community, based also on its ability to comply with international standards on research quality (p. 14).

It is important to highlight the intention of providing clear yet brief explanations as to the different ways of intending the criteria, in order to prevent misunderstandings and, most of all, inconsistent ways of implementation. Such issues are common in many assessment exercises. The British Research Excellence Framework (REF, which ended in 2014) also decided to apply the criterion in relation to the proven or potential impact of research within the international scientific community, defining impact as "an effect on, change or benefit to the economy, society, culture, public policy or services, health, the environment or quality of life, beyond academia" (REF 2011, p. 26). This definition has sparked a lively debate in the British scientific community. The discussion centred on the meaning of assessing research impact, and on the most suitable instruments to measure such impact. With respect to the REF definition of impact, it is important to note the difference with the definition adopted by the VQR, which carefully limits the impact "on the same (international scientific) community" (for more on these issues see in this volume the chapter by Blasi, Romagnosi and Bonaccorsi).

Going back to the VQR, first of all it is important to note that, compared to the previous assessment exercise (2004–2010), in the new 2010–2014 exercise, to be completed by the end of 2016, the relevance and "internationalisation" criteria have disappeared (ANVUR 2011), and have been replaced by the impact on the international scientific community. In the current edition, after the qualitative judgements given to each publication, the reviewer is asked to assign one of the following levels:

(a) Excellent (level 1): the publication achieves the highest levels in terms of innovation and methodological rigour, and reached, or it is likely to reach, a strong impact on the related scientific community at a national and/or international level.
(b) Very good (level 0.7): the publication achieves very good levels in terms of innovation and methodological rigour, and reached, or it is likely to reach, a

significant impact on the related scientific community at a national and/or international level.

(c) Good (level 0.4): the publication achieves good levels in terms of innovation and methodological rigour, and reached, or it is likely to reach, an appreciable impact on the related scientific community at a national and/or international level.

(d) Acceptable (level 0.1): the publication achieves sufficient levels in terms of innovation and methodological rigour, and reached, or it is likely to reach, a circumscribed impact on the related scientific community at a national and/or international level.

(e) Poor (level 0): the publication achieves a poor level in terms of innovation and methodological rigour, and reached, or it is likely to reach, a very limited impact on the related scientific community at a national and/or international level.

(f) Not eligible for evaluation (level 0): the publication belongs to areas that are not included in the present exercise, or it includes attachments and/or documents which are not eligible for assessment, or it has been published in the years before or after the 4-year time frame. This category also includes the missing publications against the expected number (ANVUR 2015a).

Even in this case, the effort to clarify the definitions is to be appreciated, although the considerations that describe the scores are generic and therefore they are open to different interpretations. As far as GEV is concerned, at this stage the present document will offer some observations about the decisions made by the experts regarding the Area 11a (ANVUR 2015c). The efforts of GEV have focused on the effective and transparent organisation of the reviewers' work, based on the directives by ANVUR (ministerial decree, calls for tender, guidelines) in relations to assessment criteria. The key points regarding the assessment are (ANVUR 2015c, p. 9–11):

– any product shall be evaluated by two external reviewers, selected independently by two members of GEV;
– the external reviewers can be Italian or foreign, and shall be selected according to "fairness, objectivity, and impartiality";
– the reviewers shall comply with "satisfying standards for scientific quality, with a particular attention to the assessment experience";
– the reviewers' identity shall be kept confidential (their names can only be communicated after the results have been achieved);
– "The reviewers' form is designed so as to allow the reviewer to assign a score for each of the three assessment criteria established by the ministerial decree and the call for tender, that is to say innovation, methodological rigour, and proven or potential impact; the form has a blank space in which the reviewer has a limited amount of words, and is required to insert a brief evaluation summarising the main reasons for the scores given";
– the GEV shall then turn "the information included in the forms of both reviewers into one of the five levels described in the call for tender".
– procedures for settling disputes are to be defined.

In light of this, is it possible to observe the introduction of significant new developments in the management of peer review in this new Italian assessment exercise? This question necessarily requires a complex answer. Without a doubt, the effort of restructuring the whole organisation while considering past experiences has to be appreciated. In our opinion, the issues that still need to be addressed exceed the new developments. Here is a possible list of issues; the fact that ANVUR showed its openness about the use of peer review in bibliometrics is a step towards the acknowledgement of the value of quantitative methods, rather than their fallibility: hence the need to resort to other assessment systems. The same applies to peer review, but with a stronger conviction. *Informed peer review*, that is to say a qualitative assessment that makes use of several systems to improve judgement, is gaining consideration. Such method is also mentioned in the fourth chapter of the ANVUR guidelines about GEV (ANVUR 2015b, p. 4). The present documents maintain the need to invest more on an experimental peer review process able to use bibliometric data in order to support and strengthen the assessment activities.

Measuring the scope and impact of the research is another key issue, although very challenging. However, reviewers should be able to count on a series of reliable and objective instruments that can support them in evaluating the level of diffusion of a monograph. Currently, there are ongoing initiatives that intend to develop a methodology able to ascertain the presence of such products in the catalogues of some important libraries in the world (*Library catalog analysis*) (see the chapters by Biagetti, Iacono and Trombone in this book).

Another instrument that can increase the awareness in the assessment activities of monographs is the review (Zuccala and van Leeuwen 2011; Zuccala et al. 2014). This kind of research output is considered, in this context, as a "recognition" of the monographs' value. Reviews can be taken into account together with the awards and recognitions received. Naturally, the criteria for selecting reviews are to be developed more in depth. For this purpose, it is relevant to mention the work carried out in the last 2 years by a group of scholars on this issue (Giménez Toledo 2016; Giménez Toledo et al. 2013; Giménez Toledo and Tejada Artigas 2015; Capaccioni and Spina 2016).

A separate consideration has to be made about median values, which compare the productivity of a researcher with that of his/her own community (ANVUR 2012). In our opinion, this assessment method should be preserved as it managed to contribute greatly to curtailing the "myth" of quality, sometimes emphasised arbitrarily and "at all costs", even when the quantity was insufficient. Nevertheless, it is important to point out a risk factor: the use of median values and other productivity indicators could involuntarily contribute to the publication of "in progress" essays, articles, works, etc. where the results are presented "in instalments" instead of coming together in a single monograph after a greater or lesser period of silence, as has happened in the past. If this should happen, a "normalisation" process could be introduced through a "correction index", as currently happens with academic age.

We reckon it is fair to highlight the importance of an in-depth analysis devoted to the different types of monographs. Nowadays, the need to develop a more complex breakdown of monographs that should consequently take into account different

assessments according to different purposes is more important than ever, as shown by the efforts made by ANVUR and other assessment agencies. For instance, there should be a distinction between a scientific research monograph, "educational research", and the manual. The same applies to dissemination and the third mission (the disclosure of results with educational and dissemination purposes is one of the research aims). Also in this case, assessments of impact, internationalisation, etc. can be predisposed.

With regard to the internationalisation criterion of research products, it has been said that in the VQR 2004–2010, the internationalisation criterion has been the most problematic, creating major differences in the *referees'* behaviours and judgements, resulting in widespread dissatisfaction (ANVUR 2013). It is important to mention that in the present exercise ANVUR offered some clarifications, referring to "impatto attestato o potenziale nella comunità scientifica internazionale di riferimento" ("proven or potential impact in the respective international scientific community"), (ANVUR 2015a).

In spite of this positive advance, it is important to proceed towards a more detailed "standardisation", able to stem an excessive diversification in the application of such criterion. Undoubtedly, internationalisation cannot be reduced to the fact of presenting a product written in a different language than Italian, or published abroad. Instead, the process of evaluating how a product considers the international debate and bibliography in its specific field should be carried out gradually. The objective is feasible, provided that internationalisation is intended as the ability of disciplinary sectors to have a dialogue with the world of international research. However, before evaluating such ability, it would be necessary to know how to measure it. In this regard, a few suggestions are proposed here: those products that don't even intend to interact with the world of international research are to be placed at the lowest level; above them should be placed the products that demonstrate that at least some international research products have been considered, for instance in the quoted bibliography; above these, those products that have a dialogue with such world, even those coming from Italy that are often unknown; then, there should be the products able to be considered by the world of international research (regardless of the language they are written in); at the highest level, there should be those monographs that have a national and international interest. This gradualness can be evaluated according to the ability of a single product to contribute to such debate, and to the measure in which it was and is able to do it, thanks to its dissemination and possible reception. In this specific area, it would be useful to evaluate as well:

(a) Whether it has been published or translated abroad (monograph or article).
(b) Whether it has been published or not in a ISI-WoS/Scopus journal.
(c) Whether it has been co-authored with one or more foreign authors.
(d) Whether it is the output of research that has been funded with European or international grants.
(e) Whether it is included in international databases or foreign libraries, or it was reviewed in important international journals in its fields, or it received awards or recognition even on an international level, etc.

As for the relevance of the product subject to assessment, we reckon that the referees should determine whether the analysed topic and issues are to be considered significant to the present and/or the future of the discipline, also in an interdisciplinary perspective. To this purpose, it could be useful to identify a few characteristics: does the monograph have a generic scope, or a limited and restricted one? And also: has the product been able to encourage new research, or will it be able to do that in the future? What importance is to be given to the examined issues, and to the proposed interpretations, analysis and solutions? What is the value of such products within the state of the art of the field? What is its ability to influence the scientific community, also in terms of publishing importance? What is the value of the methodology, sources and documentation used (considering also if the results, methods, sources, data, etc. have been, or could be, used by other scholars)?

The criterion of relevance is also connected to that of innovation. In order to explore this aspect thoroughly, the assessor should be able to answer the following questions correctly: to what extent has the product developed new concepts and/or approaches and methodologies, or raised new issues? To what extent has the product used new data, or innovative sources?

To conclude, a brief observation about the ethical code and respect for the fundamental principles of the professional code of conduct. In humanities, this subject applies also to particular aspects, such as the link between scientific research and the researchers' options in terms of schools of thought, cultural traditions, values and cultural and political orientations. Along this line, after examining the solutions experimented in different countries and contexts, the research group reckons it is important to make further efforts in defining the situations where a conflict of interest between the reviewers and those being reviewed might arise, in order to avoid any chance of distorted judgement.

References

Abramo, G., & D'Angelo, C. A. (2011). Evaluating research: From informed peer review to bibliometrics. *Scientometrics, 87*(3), 499–514. https://doi.org/10.1007/s11192-011-0352-7.

ANVUR. (2011). *Valutazione della qualità della ricerca 2004–2010 (VQR 2004–2010)*. http://www.anvur.org/attachments/article/122/bando_vqr_def_07_11.pdf. Accessed 11 July 2016.

ANVUR. (2012). *Sul calcolo delle mediane per l'abilitazione nazionale*. http://www.anvur.org/attachments/article/253/mediane_spiegate_definitivo_14_settembre_2012.pdf. Accessed 11 July 2016.

ANVUR. (2013). *VQR 2004–2010: Rapporto finale di area Gruppo di Esperti della Valutazione dell'Area 11 (GEV 11)*. http://www.anvur.org/rapporto/files/Area11/VQR2004-2010_Area11_RapportoFinale.pdf. Accessed 12 July 2016.

ANVUR. (2015a). *Valutazione della qualità della ricerca 2011–2014 (VQR 2011–2014)*. http://www.anvur.org/attachments/article/825/Bando%20VQR%202011-2014_secon~.pdf. Accessed 11 July 2016.

ANVUR. (2015b). *VQR 2011–2014: Linee Guida dei Gruppi di Esperti della Valutazione (GEV)*. http://www.anvur.org/attachments/article/825/Linee%20guida%20GEV_italiano_~.pdf. Accessed 11 July 2016.

ANVUR. (2015c). *VQR 2011–2014: Criteri per la valutazione dei prodotti di ricerca. Gruppo di Esperti della Valutazione dell'Area 11a (GEV 11a)*. http://www.anvur.it/attachments/article/853/Criteri%20GEV%2011a.pdf. Accessed 11 July 2016.

Baccini, A., & De Nicolao, G. (2016). Do they agree? Bibliometric evaluation versus informed peer review in the Italian research assessment exercise. *Scientometrics*, 1–21. https://doi.org/10.1007/s11192-016-1929-y.

Bertocchi, G., Gambardella, A., Jappelli, T., et al. (2015). Bibliometric evaluation vs. informed peer review: Evidence from Italy. *Research Policy, 44*(2), 451–466. https://doi.org/10.1016/j.respol.2014.08.004.

Capaccioni, A., & Spina, G. (2016). La valutazione della monografia scientifica e il ruolo degli editori: il progetto spagnolo Scholarly Publishers Indicators (SPI). *Bibliothecae.it, 5*(1), 241–255. https://doi.org/10.6092/issn.2283-9364/6128.

Cicero, T., Malgarini, M., Nappi, C.A., et al. (2013). *Bibliometric and peer review methods for research evaluation: A methodological appraisement. Munich personal RePEc archive (MPRA)*. https://mpra.ub.uni-muenchen.de/50470. Accessed 11 July 2016.

COPE. (2013). *Ethical guidelines for peer reviewers*. http://publicationethics.org/files/u7140/Peer%20review%20guidelines.pdf. Accessed 13 July 2016.

DORA. (2012). *San Francisco declaration on research assessment*. http://www.ascb.org/doraold/files/SFDeclarationFINAL.pdf. Accessed 11 July 2016.

ESF. (2011). *European peer review guide: Integrating policies and practices into coherent procedures*. https://www.vr.se/download/18.2ab49299132224ae10680001647/European+Peer+Review+Guide.pdf. Accessed 13 July 2016.

Faggiolani, C. (2015). Contro le unanticipated consequences della valutazione quantitativa della ricerca: il Leiden Manifesto for research metrics. *AIB Studi, 55*(3), 427–438. https://doi.org/10.2426/aibstudi-11366.

Giménez Toledo, E. (2016). Assessment of journal & book publishers in the humanities and social sciences in Spain. In M. Ochsner, S. E. Hug, & H. D. Daniel (Eds.), *Research assessment in the humanities* (pp. 91–102). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-29016-4_8.

Giménez Toledo, E., & Tejada Artigas, C. M. (2015). Proceso de publicación, calidad y prestigio de las editoriales científicas en educación. *Educación XX1, 18*(1), 17–44. https://doi.org/10.5944/educXX1.18.1.12310.

Giménez Toledo, E., Tejada Artigas, C. M., & Mañana-Rodríguez, J. (2013). Evaluation of scientific books' publishers in social sciences and humanities: Results of a survey. *Research Evaluation, 22*(1), 64–77. https://doi.org/10.1093/reseval/rvs036.

Hicks, D., Wouters, P., Waltman, L., et al. (2015). Bibliometrics: The Leiden Manifesto for research metrics. *Nature, 520*(7548), 429–431. https://doi.org/10.1038/520429a.

Institut de France, Académie des sciences. (2009). *L'évaluation individuelle des chercheurs et des enseignants-chercheurs en sciences exactes et expérimentales*. http://www.academie-sciences.fr/archivage_site/activite/rapport/rapport080709.pdf. Accessed 14 July 2016.

Institut de France, Académie des Sciences. (2011). *Du bon usage de la bibliométrie pour l'évaluation individuelle des chercheurs*. http://www.academie-sciences.fr/pdf/rapport/avis170111.pdf. Accessed 12 July 2016.

Institut de France, Académie des sciences. (2014). *Les Nouveaux Enjeux de l'édition scientifique*. Paris. http://www.academie-sciences.fr/pdf/rapport/rads_241014.pdf. Accessed 12 July 2016.

Moed, H. (2007). The future of research evaluation rests with an intelligent combination of advanced metrics and transparent peer review. *Science and Public Policy, 34*(8), 575–583. https://doi.org/10.3152/030234207X255179.

Mulligan, A., & van Rossumhttp, J. (2014). *What you think about the peer-review process: We turn the spotlight on the results of Elsevier studies reflecting the views of more than 3000 active researchers*. http://www.elsevier.com/reviewers-update/story/peer-review/what-you-think-about-thepeer- review-process. Accessed 14 July 2016.

Mulligan, A., Hall, L., & Raphael, E. (2013). Peer review in a changing world: An international study measuring the attitudes of researchers. *Journal of the American Society for Information Science and Technology, 64*(1), 132–161. https://doi.org/10.1002/asi.22798.

Neufeld, J., & von Ins, M. (2011). Informed peer review and uninformed bibliometrics? *Research Evaluation, 20*(1), 31–46. https://doi.org/10.3152/095820211X12941371876382.

Nicholas, K. A., & Gordon, W. S. (2011). A quick guide to writing a solid peer review. *Eos, 92*(28), 233–234.

REF. (2011). *Assessment framework and guidance on submissions*. http://www.ref.ac.uk/media/ref/content/pub/assessmentframeworkandguidanceonsubmissions/GOS%20including%20addendum.pdf. Accessed 11 July 2016.

Research Information Network. (2010). *Peer review: A guide for researchers*. London. Available at. www.rin.ac.uk/peer-review-guide. Accessed 11 July 2016.

Research Information Network. (2015). *Scholarly communication and peer review: The current landscape and future trends. A report commissioned by the Wellcome Trust*. http://apo.org.au/files/Resource/scholarlycommunicationandpeerreviewcurrentlandapeandfuturetrends.pdf. Accessed 14 July 2016.

Turner, S. R. (2009). Best practices in peer review assure quality, value, objectivity. *Journal of the National Grants Management Association, 17*(1), 43–44.

Wilsdon, J., Allen, L., Belfiore, E. et al. (2015). *The metric tide: Report of the independent review of the role of metrics in research assessment and management*. 10.13140/RG.2.1.4929.1363

Zuccala, A., & van Leeuwen, T. (2011). Book reviews in humanities research evaluations. *Journal of the American Society for Information Science and Technology, 62*(10), 1979–1991. https://doi.org/10.1002/asi.21588.

Zuccala, A., van Someren, M., & van Bellen, M. (2014). A machine-learning approach to coding book reviews as quality indicators: Toward a theory of megacitation. *Journal of the American Society for Information Science and Technology, 65*(11), 2248–2260. https://doi.org/10.1002/asi.23104.

Peer Review in Social Sciences and Humanities. Addressing the Interpretation of Quality Criteria

Andrea Bonaccorsi

1 Methodological Foundations of Peer Review in SSH Research Assessment

The first large scale research assessment exercise carried out in Italy was labelled Valutazione della Qualità della Ricerca (VQR), or Research Quality Evaluation. It covered the entire period 2004–2010, an unusually long period, due to the need to merge with the previous research evaluation, which covered the years 2000–2003. This former evaluation was an experimental one and was based on a small sample of products (approximately 18,000) selected by departments and not by individual researchers. The results, published in 2004, were not intended to be used as a basis for performance-based funding, although they were actually used by the Ministry of Universities and Research for this purpose, though with a small weighting in the formula. This situation created great expectations for the new exercise, to be launched in 2011.

The details of the VQR are discussed in various chapters in this volume (in particular, in the chapters by Capaccioni and Spina and by Solimine and Faggiolani). In this chapter we report on the discussions that followed the publication of the results, in particular with reference to the issue of peer review. All panels in SSH adopted peer review as the exclusive assessment methodology, with the exception of the Economics panel (GEV 13).

The arguments for the exclusive use of peer review have been supported by the Agency since the beginning of the research assessment exercise. In 2012 a position paper (Bonaccorsi 2012) was published on the Agency website, formalising the main arguments for the adoption of peer review and supporting them vis-à-vis the

A. Bonaccorsi (✉)
DESTEC, University of Pisa, Pisa, Italy

IRVAPP-FBK, Trento, Italy
e-mail: a.bonaccorsi@gmail.com

© Springer International Publishing AG 2018

A. Bonaccorsi (ed.), *The Evaluation of Research in Social Sciences and Humanities*, https://doi.org/10.1007/978-3-319-68554-0_4

literature. The paper also advocated the creation of an Italian database of indexed and non-indexed journals in SSH, following the lines of the scoping project described in Hicks and Wang (2009), Martin et al. (2010) and Moed et al. (2010). This project did not materialise, however, given strong oppositions from some scientific communities, in particular in Legal studies, which feared the (hidden) introduction of bibliometrics.

There is international recognition of the need to develop research quality criteria that are fit for the purpose in the case of SSH (Committee on the National Plan for the Future of the Humanities 2009; Royal Netherlands Academy of Arts and Sciences 2011; Volkswagen Stiftung 2014), recognising that measuring research performance in these fields is still an elusive goal (Finkenstaedt 1990).

The international debate on the potential of bibliometrics for SSH started more than 25 years ago (Nederhof et al. 1989) and developed since then (Moed et al. 2002; Van Leeuwen 2006; Linmans 2010; Sivertsen 2010; Nederhof 2006, 2011). The Agency believed that there are several compelling arguments for rejecting the use of bibliometrics for the evaluation of individual research products, as it happens in the VQR.

First, it is well known that existing bibliometric databases have limited coverage of materials produced by researchers in SSH, particularly books and non-English journals (Archambault et al. 2006; Ossenblok et al. 2012; Leydesdorff et al. 2011). Scholars in SSH publish in a larger variety of sources (Hemlin 1996; Hicks 2004; Huang and Chang 2008). In particular, they publish more books and book chapters than in other fields, although the share of articles is increasing over time (Engels et al. 2012). The limited coverage of materials makes any indicator calculated within the boundaries of the dataset misleading. This is true also considering new data sources, which have been developed mainly in response to the inadequacy of existing bibliometric sources. The coverage of the Book Citation Index is still limited (Gorraiz et al. 2013), while other database covering books, such as Google books and Scopus, while making rapid progresses, also do not give sufficient coverage (Kousha and Thelwall 2009; Koshua et al. 2011; Sivertsen and Larsen 2012; Sivertsen 2014). The coverage of books and book chapters in existing databases is also problematic for a technical reason, i.e., the duplication of items in chapters and edited books (Leydesdorff and Felt 2012).

Second, there is evidence that the structure of citations in books greatly differ from the structure of citations in journal articles, which form the backbone of bibliometrics (Lindholm-Romantschuk and Warner 1996; Larivière et al. 2006; White et al. 2009). Authors of books not only cite more sources, due to larger available space, but also more diverse by format (including non-serial sources and unpublished sources) (Cullars 1989, 1998; Glänzel and Schoepflin 1999; Knievel and Kellsey 2005), more remote in time and more interdisciplinary (Georgas and Cullars 2005; Hellqvist 2010; Hammarfelt 2011, 2012). Most of the cited sources do not appear in indexed sources, thus do not create a stable reference set. These characteristics make the notion of bibliometric indicators, which must be calculated within the perimeter of a dataset, highly problematic, as noted in one of the most comprehensive treatments of citation analysis (Moed 2005).

Third, there are large differences across fields in the number of publications per period (Piro et al. 2013). The relative discontinuity associated with the publication of books makes the choice of the observation window highly sensitive, reducing the reliability of measurements.

Finally, there is evidence that citation-based indicators do not capture the main dimensions of research quality (Ardanuy 2013; Ardanuy et al. 2009; Bourke and Butler 1996).

For all these reasons the Agency supported the notion that research assessment in SSH disciplines should be carried out only by peer review, with the only exceptions of Economics and Psychology, whose scientific communities largely accepted the use of bibliometrics.

With the support of an Expert group on new indicators and databases, the Agency also studied several alternative methodologies. First, it examined the Spanish experience of book publisher classification (Giménez-Toledo and Román-Román 2009; Giménez-Toledo et al. 2013, 2015). While the main goal of this exercise was considered important, preliminary analyses showed that there is still considerable heterogeneity within series of books on the same topics published by Italian publishers, so that the classification of publishers would not be reliable. As several chapters in this volume argue (among them Solimine and Faggiolani) there is certainly room in the future for addressing the issue of quality of publishers, but there is also a need for further methodological work.

It also studied the use of alternative metrics such as Altmetrics, but concluded that it is not yet mature enough to give a full scale representation of research quality (Tang et al. 2012; Hammarfelt 2014). Similar negative results were found with respect to Library Catalog Analysis (Torres-Salinas and Moed 2009) and to book reviews (Zuccala and van Leeuwen 2011; Zuccala et al. 2014). The chapters by Biagetti, Iacono and Trombone in this volume support these methodological choices.

The only large scale exercise that was considered feasible as a complement to peer review was the classification of journals (Nederhof and Zwaan 1991). This exercise took place in two waves. One, on a small scale, identified a small number of top journals to be used within the VQR exercise. Referees could receive a note about the quality of the journal in which articles to be read and evaluated were published, following the model of informed peer review. The second was associated with the new procedure for recruitment, called National Scientific Habilitation. This was a massive exercise, carried out on more than 60,000 titles (Bonaccorsi et al. 2015; Ferrara and Bonaccorsi 2016).

The results of the VQR were intended to be used by the Ministry of Research as the quantitative basis for the allocation of a share of government funding. This provision created a strong pressure for an assessment methodology that did not create unbalances across disciplines (Ahlgren et al. 2012).

2 Refining Research Quality Criteria for the Peer Review Process

Consequently, the entire VQR exercise in SSH (with the exception of Economics and Psychology) was carried out by using the peer review methodology. However, this methodology does not solve all issues. It is well known in the literature that peer review is not without limitations (Chubin and Hackett 1990; Travis and Collins 1991; Harley and Krzys 2011).

There is a need to ensure procedural fairness (Lamont and Huutoniemi 2011) in the selection of experts and the running of product allocation, scoring and aggregation of scores. There is a need to write the questionnaires in such a way as to reflect epistemic differences and ensure appropriateness (Mallard et al. 2009). Finally, there is a need to avoid any unintended consequence in suggesting implicit value judgements that penalise unorthodox views, original thinking or minority scientific opinions (Guetzkow et al. 2004; Zuccala 2012).

In general, there is a need to anticipate unintended consequences of the evaluation of scientific practices. For example, a concern often voiced in SSH communities is that research evaluation may disqualify books and give prominence to journal articles. According to Hammarfelt and de Rijcke (2015) the impact of evaluation on SSH is visible in the increase in publication in English journals and indexed journals, a trend that, however, started earlier (Kyvik 2003). According to these authors, there is no evidence of a major distortion in terms of lack of incentives to complete research monographs, which are still considered a necessary condition for the demonstration of scholarly capabilities in academic communities (Thompson 2002).

At the same time, there is evidence that research quality criteria adopted in SSH, hence used in the peer review process, are diverse and are not entirely captured by existing indicators. An influential stream of empirical research has shown that communities in SSH define research quality differently (Hug et al. 2013, 2014; Ochsner et al. 2012, 2013, 2014). These criteria reflect epistemic differences (see the chapter by Bonaccorsi in this volume), but also other factors. They create serious methodological and procedural difficulties that must be addressed systematically and with strong scientific support (Ochsner et al. 2016).

The Agency recognised this problem immediately after the VQR 2004–2010 and developed a number of initiatives to refine the process of peer review, both substantively and procedurally. In this chapter I make use of published academic articles that followed the launch of the VQR (2004–2010) and of Agency documentation that was used to take stock of the initial experience and to prepare the second exercise (VQR 2011–2014). In order to address the methodological issues of peer review and to discuss some of the criticisms raised in the fields of Social Sciences and Humanities (SSH), in 2014 and 2015 the Agency convened several group of independent experts (see Appendix 1 for their composition). These experts, usually full professors in their respective disciplines in Italian or foreign universities, were requested to offer their personal and informed views in a climate of intellectual freedom. Discussions were lively and open. No official minutes of the meetings

were taken. At the same time, the groups were requested to contribute pragmatically to the improvement of the peer review procedures to be used in the new VQR 2011–2014, which at that time was in preparation.

The groups were convened in the following areas:

(a) Architecture
(b) Humanities
(c) Legal studies
(d) Business studies
(e) Social and Political science

Each group worked for 6–8 months and produced an internal final report. The publication of the report was not mandatory and each group decided whether to publish it or not. Two of the reports were published on the website of the Agency (Architecture and Social and Political science), three were forwarded internally to the groups of experts in charge of the Research Assessment Exercise (VQR 2011–2014) for their consideration. No official translation into English is available. In this chapter we examine some of the main topics covered in the published documents, making them available through my own translation to the international community of scholars of research evaluation. In this chapter I limit the discussion to the documents published by the groups in Architecture, Legal studies, and Social and Political science.

This chapter will show that many recommendations from these groups were finally received in the new edition of the research assessment exercise (VQR 2011–2014). Listening is an essential component of the work of an evaluation Agency.

3 Peer Review in the Research Assessment Exercise (VQR 2004–2010)

The VQR exercise made extensive use of referees for the peer review. It is interesting to note that the use of peer review was mandated by the law that created the National Agency for the Evaluation of Universities and Research Institutes (ANVUR). More precisely, the considerations that open the decree state that the Agency should make “prevalent use” (our translation) of peer review. This provision comes from the parliamentary debate on the creation of the Agency, which was dominated by the fear that the quantitative approach, associated with bibliometrics, would dominate. Several MPs requested that peer review be defined as the exclusive methodology to be used. In the final text, this wording was attenuated and became “prevalent use”.

The Agency followed this legislative provision quite literally, showing that more than 50% of products evaluated in the VQR had been submitted to peer review. Peer review was the exclusive methodology adopted in the following areas: Architecture (area 8), Antiquities, Philology, Literary studies, Art history (area 10), History,

Philosophy, Pedagogy (area 11), Law studies (area 12), Political and Social sciences (area 14).

It is interesting to report on methodological choices made in those areas that cut across the boundaries between STEM and SSH, in at least three cases. In the case of Architecture, the area was divided in two sub-areas: one mostly related to Civil Engineering in which bibliometric methods were used, and one related to Architectural design, history, conservation and restoration in which peer review was instead adopted in an exclusive manner. This distinction has not been the object of controversy. The second case refers to Psychology. From an administrative point of view, Psychology is classified in area 11 together with History, Philosophy and Pedagogy. This reflects the humanistic origin of psychological studies in the Italian academic system. During the preparation of the VQR, the scientific society of Psychology issued a document in which it was claimed that the international practice of psychological studies places them closer to hard sciences than to humanities. Consequently, it was suggested to adopt evaluation methods used in the international community. This suggestion was fully reflected in the VQR practice, for which the Psychology sub-group used bibliometric indicators.

In the case of Architecture and Psychology this divide was full reflected in a procedure that started shortly after the VQR, i.e., the new National Scientific Habilitation procedure, opened in 2012. In this procedures scientific disciplines were classified ex ante as "bibliometric" and "non-bibliometric" and separate classes of indicators were adopted. Area 8 was then split into two parts, one bibliometric (Civil engineering) and one non-bibliometric (Architecture), while Psychology was entirely classified as a bibliometric discipline.

Finally, in Economics and Statistics (area 13) the panel made the choice to combine bibliometric indicators with peer review.

The most important controversies that originated from the VQR can be summarised as follows:

(a) Object of evaluation
(b) Research quality criteria
    (i) Relevance
    (ii) Originality/innovation
    (iii) Internationalisation
(c) Epistemic pluralism
(d) Impact

These are discussed below.

4 Object of Evaluation

The description of products to be admitted to the evaluation followed the broad list used in the UK, first by the RAE and then by REF. It was a comprehensive list that tried to cover all communication channels used by scholars. The Call for VQR

(7 November 2011) (available at http://www.anvur.it/index.php?option=com_content&view=article&id=28&Itemid=119&lang=it) included the following list of products (our translation).

(a) Journal articles
(b) Books, chapters in books and in conference proceedings, only if with ISBN
(c) Critical editions, translations and scientific comments
(d) Patents awarded to researchers in the 7 years covered by the research assessment
(e) Compositions, drawings, designs, performance, collections and exhibitions; artefacts, prototypes, pieces of art and their projects; databases and software; thematic maps, only if associated with publications that allow their evaluation.

The specific interpretation of these items was left to the guidelines of the expert groups (GEV, Gruppi di esperti della valutazione, or Evaluation expert groups).

There was some debate on the inclusiveness of the list and on the unintended censorship that the research quality criteria created in some scientific communities. As a matter of fact, the list of products was greatly enlarged and modified. It now includes 45 different typologies. Several new typologies have been included, in most cases following the debate raised in the SSH community. In some cases specific suggestions were made by the Research groups convened by the Agency and they were accepted and adopted.

4.1 Object of Evaluation: (a) Architectural Projects

The Research group on evaluation in Architecture addressed this issue at length. It summarized a large debate, that had taken place during and after the VQR, regarding the evaluation of "projects". Projects are technical artefacts, composed of drawings, pictures, photos, maquettes, calculations and narratives, that are materially embedded into a variety of formats. They are the main product of architects in all their fields (buildings, urban, territorial, industrial design, interior design). Architects with an academic appointment usually involve students into the various sections of the project, teaching them during the execution. This kind of field education is considered essential in the discipline.

Although the notion of "design" was explicitly included in the list of admissible products, there were several obstacles. The first was practical: all products should have been submitted to the evaluation in a digital format. No other material embodiment was considered. In the case of projects, this required to pack all documentation into a digital format, eliminating all 3D products. The second obstacle was conceptual: design products (as well as, in other disciplines, software or machines) were admitted to the extent that they were associated with a published source illustrating its content. This was intended to ensure the communication and circulation of the research associated with the project.

Finally, there was a psychological obstacle, repeatedly reported by scholars in Architecture after the VQR: although the project was included under the heading of "design" in the list of admissible products, people felt that the evaluation of the projects would have been poor, compared to published articles and books. Therefore a very small number of projects were submitted.

Two other options had been considered as alternatives to the request of a published article associated with the project, but they were eventually discarded in the preparation of VQR 2004–2010: reviews and awards. The rationale of reviews is sound: Architect A builds an innovative building, Architects B and C comment on the building in academic journals, or in magazines and newspapers. The same can be said for awards as indicators of quality of the projects: if a project wins a competition with others there is an indication of quality. The lack of common definitions of reviews, of any official list of awards and the extreme difficulty in standardising the data suggested discarding these indicators. In addition, it was not clear whether reviews and awards are indicators of research quality, defined in scientific terms with respect to the state of the art in the field, or rather of other relevant aspects, such as the urban or social impact, cost effectiveness, sustainability, or others.

The choice of asking for a published source was well grounded. In an interview in a journal of Architectural studies I defended this choice based on the combination of functional and aesthetic dimensions in the architectural design work (Lo Sasso 2011). I proposed that the functional, or problem solving, activity could be evaluated by the scientific community, while the aesthetic dimension could not.

> We have here the coexistence of opposing principles: while the marking out of the space of the possible worlds demands the use of scientific knowledge (in the broad sense of valid, codified, replicable knowledge), the design generation has much more value if based on a single unique understanding. This characteristic has a crucial implication in terms of evaluation. One can evaluate only what is considered to be knowledge legitimised by a scientific community, or what emerges from forms of communication and themes selected autonomously by the community.…
>
> Strictly speaking, on the other hand, one *cannot* evaluate an aesthetic product. The aesthetic product is the result of a form of expressive rationality that doesn't need the validation of a community, but instead is aimed at *creating* a community of its own – the community of those who value that particular expressive form. That is true in the major works of artists, but it is also central to the world of the architectural project.…
>
> So in a certain sense there is inside the project an element that can be evaluated, and one that cannot be evaluated! That means that one cannot claim that the project is *by itself* the subject of evaluation. The scientific work on the project, before and after it is realised, is the subject of evaluation, as well as the way it is presented to the community in compliance with the rules of scientific literature. In a nutshell, it is the collection of publications that accompany the project and circulate the knowledge in the community (p. 87–88).

I suggested that the project is an activity in which replicable scientific knowledge is combined with non-replicable, unique, non-codified expressive knowledge of an aesthetic type. In order to be evaluated the two dimensions should be disentangled. The distinction between these two forms of knowledge requires the intervention of a scientific community working with the standard forms of scientific communication, like journal articles and books. I added the following remarks:

> From a scientific point of view, surgeons cannot be evaluated for their surgery, jurists for their opinions in court, nor engineers for their calculations, but each of these activities needs to be accompanied by products accepted for publication by their respective scientific communities (p. 88–89).

As a matter of fact, the debate after the VQR made it clear that the scientific communication in the community of architects does not necessarily take the form of publications. Many respected scholars made it clear that when they produce a good project they *do not* publish accompanying articles, but rather leave the appreciation of the project to the material observation *in situ*. People visit the work, take pictures and make drawings, and reflect on the innovative nature of the project. After years, if the project has attracted the interest of the community, other authors will place it in their books as an example. For highly successful architects, in due time there will be a book collecting their main work. Journal articles by the authors of projects are not necessarily associated with quality work.

Therefore the request that a published source be associated with the project was theoretically grounded, but lacked some realism in the consideration of the form of scientific communication in a professional discipline.

The Research group in Architecture worked in considerable detail on this issue (Gruppo di ricerca sulla Valutazione in Architettura 2015). Their document is an interesting piece of (albeit implicit) philosophy of science in disciplines at the crossroads between science and humanities, and with a design orientation.

They start by noting the

> transversal and eccentric position of architecture, as a collection of heterogeneous disciplines that are kept together in most countries, with few differences, by the goal of educating technicians. It is obvious that the intersection of research activities between science, humanities and design produces intrinsically divergent results, sometimes placed at the crossroads.

Within this field, there is a difference between the notion of "design" and the notion of "architectural project" (*progetto architettonico*). The former is a specific type of project referred to a product. The latter is more comprehensive and includes, according to the Research group, "architectural, urban, territorial and design projects". Projects have largely different scales, according to the number of people involved, the stakeholders, the material deployment. At all scales, however, it is possible to evaluate the contribution of any single author. The suggested definition, based on the experience of the University of Bologna, is as follows:

> Product of civil engineering and architecture. Study/project aimed at clarifying essential elements for the design, design of structures or infrastructures in civil engineering, design or drawing of artefacts in architecture, industrial design, or part thereof. In order to be evaluated, the project must have clearly recognisable elements of new and scientific or artistic nature. It must have been awarded a prize, or a mention, in tournaments or competitions, or it must have been published. In the case of impediments to the publication, by nature or contractual clauses, the project will be evaluated, even in a confidential manner, only if associated with a document illustrating its description and all elements allowing for evaluation.

The documents recognise that the object of evaluation cannot be the professional activity, but the activity which takes place "in the public sphere in which projects circulate, are communicated and evaluated". This is a fundamental point with a larger conceptual importance. In academic disciplines in which there is an interface with applications, researchers are often involved in professional activities with clients. These activities may bring prestige to the university and add expertise to researchers, but are not by themselves the object of research evaluation. What can be the object of evaluation is an activity that takes place "in the public sphere". Consequently the Research group suggests that architectural projects are evaluated on the basis of their circulation in "journals, books, catalogues, as well as their participation in competitions, awards and public recognition received".

The document criticises the choice made in the VQR 2004–2010 of suggesting that the medium used for communication was more important than the content. In particular, the provision that design activities could be evaluated only if a publication was associated with it was considered wrong. The group suggested to enlarge the material embodiment of the project, by accepting for evaluation a portfolio, or a PDF document including images, drawings and narratives, even if not published.

Interestingly, the VQR 2011–2014 accepted these suggestions. The new list of objects admitted to the evaluation includes "architectural projects".

4.2 Object of Evaluation: (b) Textbooks

The ministerial decree that launched the VQR and gave the Agency the guidelines for the research assessment explicitly excluded "text and software used prevalently for teaching" (Ministerial Decree art. 5, subsection 2 dated 15 July 2011).

The VQR excluded from evaluation those books that were intended as student textbooks, either at the university or high school level. The rationale was clear: textbooks are not a research product, but an educational product. Scholars summarise their own research and others' research into a book with a view to consolidate the state of the art for transmission to younger generations.

As a general rule, all guidelines issued by GEVs excluded textbooks as objects of evaluation.

This view was challenged by the group of experts in Legal studies. According to them, textbooks are a major channel for authoritative scholars to offer their own research-based and systematic interpretation of debated issues in law. Legal textbooks, according to this view, can be products of research.

The Research group in Architecture recommended a case-by-case approach.

The new nomenclature now includes "specialised university textbook" (or *Manuale universitario specialistico*), a notion that, while admitting the intended use for teaching (textbook), still preserves the requirement that it includes original content (specialised).

4.3 Object of Evaluation: (c) Edited Books

Another controversial issue regards edited books. In the call for participation issued in November 2011 the notion of "books" was left unexplained. What is a book? Does the definition include edited books or only authored books? And authored by how many authors?

The ministerial decree that launched the VQR and provided the Agency with the guidelines for the research assessment explicitly excluded "editorial activities and editing of books" (Ministerial Decree art. 5, subsection 2 dated 15 July 2011).

As a matter of fact most guidelines issued by GEVs excluded edited books (in Italian, "*curatela*"). This orientation was reinforced by the start of the new National Scientific Habilitation procedure in 2012. In the construction of indicators for both candidates and commissioners edited books were explicitly excluded.

The overall rationale was clear. In edited books, particularly if there is no information on the editorial process underlying the approval of the title (i.e., peer review), it is often impossible to evaluate the innovative content provided by the editor. Since the Habilitation procedure involved the large scale calculation of indicators, including edited books of uncertain quality seemed to be unwarranted.

Therefore after 2011 (start of VQR) and 2012 (start of Habilitation) there was a widespread assumption that edited books were not valuable items for evaluation.

The Research group in Architecture discussed this issue at length. They claimed that edited books included one of their most important and peculiar research products, that is, critical work. They suggested to adopt the definition created by the Bologna group ("Edited books are a collection of texts on a specific theme, catalogues of exhibitions, monographs on authors in which the work of text collection, selection and comment goes beyond a short introduction"). In the field of Architecture edited books are often the result of long critical work, which is itself original.

A similar argument was raised by the scientific community in Archaeology. Archaeological excavations are the result of large scale mobilisation of a number of scholars, which contribute to edited books from different perspectives. The editor of these books are the scientific coordinators of the excavation. These edited books have the largest possible value in the community.

Following this criticism, the new VQR 2011–2014 included edited books in the list of admissible products, and clarified that critical editions refer not only to texts, but also to excavations.

5 Drafting the Criteria of Evaluation of Research Quality

The VQR exercise adopted a framework, applied to all disciplines, in which three general quality criteria were defined. The formulation of criteria was not decided by the Agency, but was already defined in the ministerial decree that launched the

exercise (Ministerial Decree art. 8, subsection 2 dated 15 July 2011). The decree articulates the quality criteria as follows.

(a) *Relevance*, to be understood as value added for the advancement of knowledge in the field and for science in general, also in terms of congruence, effectiveness, timeliness and duration of the impact.
(b) *Originality/innovativeness*, to be understood as the contribution to the advancement of knowledge or to new achievements in the relevant field.
(c) *Internationalisation* and/or competitive international potential, to be understood as positioning in the international scenario, in terms of relevance, competitiveness, editorial diffusion and recognition in the scientific community, including explicit collaboration with researchers and research groups of other countries.

The specific implementation of the three criteria was delegated to GEVs in each discipline. The guidelines of GEVs led to formulations that were used in the questionnaires that referees were requested to complete during the peer review process. While there have been slight variations in the wording of the criteria across disciplines, the main thrust expressed in the call for participation was essentially preserved.

During and after the VQR some debate was raised on the three quality criteria, and, above all, on their implementation. Almost all research groups convened by the Agency addressed this issue and formulated suggestions on refinements to be introduced in the definition of criteria. Overall, the suggestions follow three directions:

– Refine the definition of research quality criteria
– Enrich and/or clarify the wording
– Develop qualitative scales of qualitative judgement to be associated with a score

I review the main contributions below. It is interesting that in the new formulation of the VQR 2011–2014 the criteria of relevance and originality/innovativeness have been entirely reworded as follows:

(a) *Originality*, to be understood as the degree to which the work introduces a new way of thinking with respect to the scientific object of the research and is therefore different from the previous approaches to the same object.
(b) *Methodological rigour*, to be understood as the degree to which the work presents in a clear way the goals of the research and the state of the art of the literature, adopts a methodology appropriate to the object of research and demonstrates that the goals have been achieved.
(c) *Current or potential impact on the relevant international scientific community*, to be understood as the degree to which the work had, or is able to have in the future, a theoretical or applied influence on that community, also due to its ability to respect international standards of research quality.

As it will be seen in the next sections, it is clear that the new wording is the result, among other things, of the acceptance of the suggestions of the Research groups.

In particular, the distinction between originality and methodological rigour makes the criteria much more transparent and operational. Also, there is a significant clarification of the meaning of internationalisation.

5.1 Relevance

By relevance was meant a summary judgement of the research work (defined as "product" in the VQR terminology) from the point of view of its importance, value and impact. As already noted, the groups of experts (GEVs) took the general definition of relevance from the ministerial decree and adjusted it, making it operational in the peer review form to be completed by anonymous referees. The Agency decided that the specific formulation should be left to disciplinary expert groups, in order to be close to specific epistemic practices. The Agency only controlled at the central level whether the formulations were consistent with the legislative mandate.

Several research groups advanced suggestions on how to improve the definition and the implementation contained in the VQR 2004–2010 and adopted for peer review. For example, the Architecture group suggested the development of descriptors that might help the referees articulate their judgement. These descriptors focus attention on several aspects of the quality of a research work, without formulating a scale.

Their document offers a comparison between the formulation adopted by the GEV in Architecture in the VQR 2004–2010 and the new suggestion, as shown in Table 1.

According to these suggestions, there are three dimensions of relevance: the choice of the topics, the treatment of the topics, and the circulation of the research work. In order to formulate a judgement, referees should ask questions on all these dimensions. No indication is given about the relative importance of these dimensions.

5.2 Originality/Innovativeness

As already noted, by originality and innovativeness the ministerial decree and the VQR 2004–2010 intended a broad characterisation of advancement with respect to existing knowledge. This is, however, a multidimensional construct. A research contribution might be original for different reasons: due to an original, unexplored topic, or a new method applied to known topics, or both. The tension between topics and methodology was not elaborated in the previous broad definition.

Given this tension, the GEVs tried to develop a questionnaire for referees that introduced methodological issues in separate items.

Table 1 Comparison between the formulation of the relevance research quality criteria in the VQR 2004–2010 and in the suggestion of the Architecture research group

VQR 2004–2010 GEV 8 Architecture	Suggested formulation
Relevance.	Relevance.
Please assign a score to the importance of the research work, in particular to its value and impact with respect to the advancement of the field in terms of scientific knowledge, technical capabilities and/or improvement of architectural and engineering production.	New suggestion. Please assign a score to the relevance of the research work on the basis of the following descriptors.
	Descriptors related to the topics
	The work precisely identifies some topics and is well organised from a logical point of view.
	The topics selected in the work have relevant implications for the discipline.
	The topics selected in the work are present in the national and international debate and may be considered important in the time window covered by the research assessment exercise in the field of Architecture.
	Descriptors related to the treatment
	The topics selected are treated in a pertinent way and are discussed in a rigorous manner from the point of view of argumentation.
	The topics selected are treated with reference to an information base that is solid, extensive and well organised.
	Descriptors related to circulation
	The work is placed in circuits at a national or international level.

Source: our adaptation from Gruppo di ricerca sulla Valutazione in Architettura (2015)

Take for example the GEV in Architecture. Under the research quality criteria Originality/Innovativeness the original GEV utilised a battery of questions (Table 2, left column) aimed at focusing the attention on two distinct dimensions: the topics and the methods. Their formulation seemed to define innovativeness in terms of opposition with existing tenets of the discipline (which I have translated as "received wisdom"). The new formulation suggested by the Research group is more comprehensive: a work may also be considered innovative if it targets topics that have been covered in the past but in a novel way. In addition, the suggestions clearly distinguish the content (topics) and the methods. This distinction has been fully received in the new formulation of the quality criteria for the 2011–2014 exercise.

A similar suggestion can be found in the document prepared by the Social and Political science research group (2015). They start from the new VQR definition, in which Originality is a separate dimension from Methodological rigour, as suggested in the debate. They move beyond definitions and suggest the development of a qualitative scale, associated with scores.

Table 2 Comparison between the formulation of the originality/innovativeness research quality criteria in the VQR 2004–2010 and in the suggestion of the Architecture research group

VQR 2004–2010	
GEV 8 Architecture	Suggested formulation
Originality/Innovativeness	Originality/Innovativeness
Please assign a score to the level of originality and innovation of the research work.	Please assign a score to the level of originality and innovation of the research work on the basis of the following descriptors.
Did the work call into question the received wisdom about architectural and engineering production?	Descriptors related to the topics
	The work addresses topics that are poorly investigated, or proposes new approaches to existing topics through advancements, solutions, or by offering confutations of received wisdom.
Did the work develop new concepts, approaches, methodologies, tools or techniques in its own field in architecture or engineering?	The work places assumptions or positions taken for granted in a different light.
	The work achieves unusual results/confutations/configurations.
	Descriptors related to the methods
	The work experiments with non-conventional approaches and methodologies in an effective manner.
	The work applies analytic methodologies or suggests solutions that are innovative in the field.

Source: our adaptation from Gruppo di ricerca sulla Valutazione in Architettura (2015)

Originality: what is the level of innovativeness of the work?

The approach, the hypotheses and the analysis deliver a contribution to the existing literature and to the advancement of knowledge in the relevant field.

- Highly innovative (score = 5)
- Innovative (score = 4)
- Partially innovative (score = 3)
- Scarcely innovative (score = 2)
- Neither innovative nor original (score = 1)

Methodological rigour: to what extent the goals have been made explicit and formulated with clarity, there is awareness of the state of knowledge on the topics, and an appropriate methodology has been utilized?

- The goals have been made very explicit and have been formulated with great clarity, there is full awareness of the state of knowledge on the topics, and a particularly appropriate methodology has been utilised (score = 5).
- The goals have been made explicit and have been formulated with clarity, there is good awareness of the state of knowledge on the topics and an appropriate methodology has been utilised (score = 4).
- The goals have been made sufficiently explicit and have been formulated with some clarity, there is some awareness of the state of knowledge on the topics and a relatively appropriate methodology has been utilised (score = 3).

- The goals have been scarcely made explicit and have been formulated with poor clarity, there is limited awareness of the state of knowledge on the topics and the methodology utilised is not completely appropriate (score = 2).
- The goals have not been made explicit and formulated with clarity, there is scarce awareness of the state of knowledge on the topics, an inappropriate methodology has been utilised (score = 1).

According to this contribution, originality refers to: (i) approach, (ii) hypotheses, (iii) analysis. This is an interesting example of epistemic discussion. What is important for originality is, first of all, the overall approach, i.e., the framework of analysis and the basic assumptions that researchers use in addressing an issue. The notion of approach is deliberately loose, since it accommodates the main research options of researchers in framing a research problem and identifying a promising path. Second, the research hypotheses. Taken together, these two dimensions give high importance, to use the classical Popperian distinction, to the context of discovery rather than to the context of justification. Highly original research hypotheses should be evaluated positively, irrespective (one might say) of the actual results, provided an appropriate methodology is put in place. Results are contemplated in the third category ("analysis"), which clearly refers to the analysis of research findings. In this formulation nothing yet is said about methodology, or the process that links research hypotheses to results and their analysis.

The issue of methodology is contemplated under separate quality criteria, "methodological rigour". The document of social and political scientists suggests an implementation in which rigour is a function of: (a) explicit research goals, (b) clarity of research goals, (c) state of the art, (d) appropriate methodology.

5.3 Internationalisation

The introduction of "internationalisation" among the evaluation criteria marked a discontinuity in Italian academia in SSH. While international orientation has traditionally been part of the academic life of the most advanced scholars, it was not considered a permanent dimension for all scholars.

The international dimension is firmly established in the legislation and in the ministerial decrees opening the VQR procedure, as well as other ministerial decrees regarding the accreditation of doctoral courses or the National Scientific Habilitation. It can be said that internationalisation is nowadays irreversibly embedded in academic life.

Of the research quality criteria included in the VQR, one of the most controversial was Internationalisation. This is no surprise given the peculiarity of SSH with respect to the use of the national language and communication in books and journals in the national language.

As a matter of fact, the notion of internationalisation has never been challenged in itself. The question is its interpretation. What does internationalisation mean,

exactly? After the VQR some authors and learned societies have voiced the concern that in the VQR exercise there was a privilege assigned to products published in English, irrespective of the content. These communities claimed that the formulation of the VQR opened the way to a wrong interpretation, according to which by internationalisation is meant publication in international journals, usually in English. According to this interpretation, several critics argued, it was common practice to award a higher score to articles in the English language than to articles or books in the Italian language and/or published in Italy. The concern was that referees, faced with a question regarding the degree of internationalisation of the product, would invariably give a low score if the product was written in Italian.

This claim is not based on strong empirical bases. In a study on the scores obtained by articles in History in the VQR, Ferrara and Bonaccorsi (2016) do not find a significant correlation between scores and the prevalent language in journals for products in area 11 (including History, Philosophy, Geography, Anthropology, and Pedagogy). However, despite the evidence the opinion that the inclusion of internationalisation among the criteria was responsible for a significant downgrading of research in SSH is widely held.

This issue has been broadly addressed by research groups. The Architecture research group introduced an initial distinction between the relevance of the research work "for current and future research" in terms of internationalisation. The distinction between current impact and future impact aimed at introducing a reflection on the potential for international impact, even if the original research work was published in the national language and not in English. In other words, the appreciation of the potential impact answered a counterfactual question of the type: "Would the international community have an interest in the research work should it be made available in a suitable language?".

This is a sophisticated evaluative approach that demands the construction of a scenario, or the activation of hypothetical thinking. Is this notion too vague? One might argue that what matters for evaluation is not the future, but the past. Did the research work have any impact on the international community or not? If not, why should we be interested in its potential, which is by definition uncertain? This formulation, which has been adopted by the official documentation of the VQR, is instead very promising for evaluation.

First, it takes into account the specificities of research in Humanities, and, to a certain extent, in Social and Political science. In these fields the choice of the language is not neutral, since the language is itself (in some cases) the object of research, because the objects of research are written texts. This happens, for different reasons, in Literary studies and in Law.

Second, it forces the evaluators to consider that, even if research carried out in SSH is originally written in the national language, it *might* have a wider audience if it were brought to the attentions of scholars abroad. The notion of "potential internationalisation" creates the conceptual distance between current reality and a possible, desired future. Then the question becomes: is there a potential? If not, this should be justified: it is likely that the interest of foreign scholars for a fiscal law in a given country is limited. If, however, the potential is clearly identified, then the

Table 3 Comparison between the formulation of the internationalisation research quality criteria in the VQR 2004–2010 and in the suggestion of the Architecture research group

VQR 2004–2010 GEV 8 Architecture	Suggested formulation
Internationalisation	Internationalisation
Which is the degree of internationalisation of the work?	Please evaluate the relevance of the work for current and future research in the relevant discipline, or in an interdisciplinary perspective (theoretical and methodological implications, advancement of knowledge) in terms of internationalisation, defined as the ability to enter into dialogue with the international research community, on the basis of the following descriptors.
	The work takes into account international production in the specific field.
	The work discusses topics and issues that are relevant at an international level.
	The work has been placed in international circuits (with a certain degree of penetration in the scientific community).

question becomes: did this author undertake any initiative to make the research work available to an international audience? This question may be answered only with reference to the state of the art of the relevant discipline. If it is common practice to publish research results in international journals, or international books, or it is common practice to have one's own books translated into a foreign language, then this practice creates a benchmark.

In this formulation, internationalisation is a dynamic process. The evaluation process should represent the internationalisation dynamics along the time axis and offer quality criteria that are consistent with the stage of maturity that is appropriate for any discipline.

This notion is more complex and sophisticated than just the one based on the language of publication, but has the advantage of being framed in a dynamic setting (Table 3).

As in the other cases, the Social and Political science research group suggested the development of a scale, taking into account the distinction between actual and future (or potential) impact on the international community. Their suggestion articulates the notion of impact in three dimensions: (a) interest, (b) visibility, (c) recognition.

The suggested wording is as follows.

Actual or potential impact on the relevant international scientific community: what influence did the work have, or will have, on this community?

The impact from the point of view of interest, visibility and recognition by the international scientific community is (or will be)

- Very significant (score = 5)
- Significant (score = 4)
- Slightly significant (score = 3)
- Not very significant (score = 2)
- Irrelevant (score = 1)

6 The Issue of Epistemic Pluralism

6.1 The Debate on Pluralism in SSH

One of the crucial epistemic features of SSH is their pluralism. The way in which disciplines carry out the production of knowledge is the object of fascinating literature in the sociology of higher education and the sociology of science (Knorr-Cetina 1999; Whitley 2000; Becher and Trowler 2001). While in hard sciences it is often found that a paradigm is dominant, to the point that almost all scholars share the same research questions, relevance criteria, and experimental techniques (perhaps with one or a few minority views), this is not generally the case in SSH (Lamont 2009). Objectivity may be a normative goal but does not prevent the formation of lively pluralism (Novick 1988). Scientific communities are often organised by schools of thought. The founders of schools may well be alive, but the strength of school differences is felt well beyond the passing of the founders.

Pluralism creates a serious problem for research evaluation. Ex-post research evaluation is not double blind, it is single blind. The referee knows the name of the author(s). There have been suggestions to make the name of the author(s) blind in published articles, but it is obvious that referees might easily retrieve their names with a simple Internet search. The single blindness of ex-post peer review may expose some authors to referees that are not able to professionally evaluate the products.

At one extreme, we might find cases of personal conflict. The evaluation panel may be unaware of the negative personal relationships between colleagues and assign the product to a biased one. More frequently, there is the possibility that referees who adhere to different schools of thought are not able to disconnect their evaluation of the quality from their epistemic and methodological preferences.

The Agency took some measures to handle this issue. First, the behaviour of referees involved in the 2004–2010 exercise was massively examined with the goal of detecting unreliable referees. They were defined as those that gave consistently similar extreme grades (in particular, zero or one) to a large number of products. The number of equal grades that would qualify the referee as unreliable clearly depends on the number of products examined. The probability of assigning all zeros or ones to products decreases with the number of products examined. This probability was calibrated in each discipline against the observed proportion of extreme grades. Referees with a behaviour significantly different from the average were tagged and have not been recruited for the subsequent exercise (2011–2014).

Second, the referee instructions included a more detailed narrative on the importance of declaring the conflict of interest. The definition of conflict of interest was enlarged to cover the case of deep methodological and substantive differences between the referee and the subject evaluated.

Sometimes the suggestion was made to ask the subjects to name those referees that they would *not* accept. This idea was rejected by the Agency. It was believed that this mechanism would have opened the way to reject more severe yet neutral

and professional referees in order to increase the probability of being assigned more lenient judges. In addition, managing veto declarations on thousands of scholars is cumbersome.

This issue has been thoroughly examined by the Social and Political science evaluation research group (2015). Their view is rather sharp.

> Within the same discipline, such as, for example, Sociology and Political science, there are very different disciplinary streams as well as theoretical approaches, research traditions and methodological options that are not only different but often engaged in fierce competition. In this situation a scholar who subscribes to a given approach is evaluated negatively – and in good faith – by a scholar following an alternative approach just because the latter is not in a position to appreciate and appropriately evaluate the quality of the contribution within the chosen scientific benchmarks.
>
> Given this premise, it must be understood that:
>
> (a) No evaluation exercise will ever receive a full legitimation by all evaluated subjects.
> (b) There is no such thing as a perfect evaluation exercise, and this is even more true in Social and Political science.
> (c) At the same time, there are no shortcuts. What is needed, even for a partial success of the evaluation exercise, is that the members of the GEV and referees be aware of this situation and make an ethical commitment with regard to it.

While the premise is rather sceptical of the possibility of fairness in the evaluation, the groups made the effort to identify possible solutions. They can be classified as follows: (a) develop an ethical code to be accepted by referees, (b) use informed peer review, (c) appoint arbitration groups. Before examining these solutions it is useful to compare the approach of scholars in Social and Political science with those of Legal studies.

The issue of pluralism had also been considered by the Legal studies evaluation research group (labelled GRAVAG from the acronym of the Italian name: see Gruppo di ricerca ANVUR sulla Valutazione in area giuridica (2015)). The group convened in October 2014, and after preliminary work it circulated a questionnaire to all learned societies in Legal studies. Almost all societies (20 out of 23) responded with a written contribution. The summary of contributions was discussed in an open seminar in June 2015. Contrary to their colleagues in Social and Political science, the scholars in Legal studies did not even mention the epistemic pluralism in their disciplines, but they had it in mind when they made a number of recommendations under the heading "protection of evaluated subjects". Framing the suggestion in legal language, they discuss the case in which two referees formulate divergent judgements on the same piece of research. The procedure adopted by the VQR 2004–2010 was to appoint a consensus group *within* the GEV. The consensus group could appoint a third referee, which was common practice.

By and large, the responses of the learned societies were critical of this procedure. Among the suggested procedures we find the appointment of the entire evaluation panel (GEV) based on suggestions of learned societies; the random extraction of the members of the consensus group; the exclusion from the consensus group of members of the expert panel who have managed the referee process; the creation of consensus groups with full transparency of their members; the publication of all names of referees.

Taken together, these suggestions point to three alternative mechanisms: (a) delegation from learned societies, (b) selection procedures, (c) transparency. Only the latter suggestion was adopted by the Agency, as we will see. I discuss the suggestions of these research groups, respectively.

6.2 Ethical Code

The suggestion of an ethical code is interesting. From the start the Agency had adopted an ethical code, published on the website (available at http://www.anvur.it/index.php?option=com_content&view=article&id=12&Itemid=337&lang=it).

This code is extended to all experts appointed by the Agency, who accept it by signing the letter of acceptance. However, the suggestion of the research group went beyond. They asked for an initiative aimed at sharing a number of professional and ethical criteria among referees. The suggestion is important. In the future a strong initiative should be taken (perhaps using a MOOC) in order to align all experts with a number of practices.

6.3 Informed Peer Review

In order to mitigate the variability of experts' judgements in the peer review process, the group on Social and Political science suggested to strengthen the "informed" part of the process. Interestingly, in doing so they placed some elements of bibliometrics in the domain of peer review. In practice, they suggested that the score assigned to an article should not be lower (higher) than a given threshold if there was objective information pointing to a high (low) quality. This information was to be taken from citations, whenever available, and classification of journals. With respect to citations, the group suggested a threshold of five citations: if a referee would like to give a low score to an article with more than five citations, he/she should have to justify the decision.

The suggestion is neither practical nor based on a thorough empirical analysis of the relation between citations and peer review scores in SSH. However, it is interesting in conceptual terms. What these scholars are telling us is that bibliometric indicators not only do not induce a bias in the scientific production of SSH, but help to protect it from the distortions of peer review.

6.4 Arbitration Groups

It is well known in the literature that peer review is subject to inter-rater variability. There is no way to avoid this outcome. Consensus groups or arbitration groups are a reasonable solution.

The hot issue here is whether the members of the arbitration group should be the same as the panel. According to the document by the Social and Political science group, there is no objection to including members of the panel (i.e., those who have administered the original peer review by receiving the scores of the referees), but with the inclusion of external members as additional referees. As we will see below, scholars in Legal studies are more sanguine in requesting only external members.

6.5 Delegation by Learned Societies

The summary of opinions of learned societies in Legal studies (GRAVAG 2015) shows that a number of them ask to be involved in the selection of experts. The suggestion that learned societies might have a role in the appointment of members of the expert panel is not new. It was followed in the experience of AERES in France (alongside the appointment of experts upon the suggestion of trade unions). In the Italian experience, the Agency has systematically refused to take it into consideration. The main argument was that the Agency had a legitimation that comes directly from a law, that is, the highest level of legislative authority in the country. The process by which the members of the Board of the Agency are identified (by an independent search committee, from a few thousands self-candidatures), evaluated, screened (forming a short list of 15 names), selected (by the Minister of Research), approved (by the Council of Ministers) and validated (by two independent parliamentary votes) is sufficient guarantee of independence and professionality. The legitimation of the Agency does not come from the delegation of learned societies. In turn, given the legitimation of the members of the Board of the Agency, their choice of experts is legitimate, upon the condition that the selection procedures are transparent.

Shortly after the launch of the National Scientific Habilitation, one of the most prestigious learned societies in Law, the Association of Italian Constitutional Law, asked the Administrative Tribunal to stop the procedures through which the Agency classified scientific journals. The main argument of the Association was that the power to express a scientific judgment could not reside in the Agency, but should be recognised as exclusive responsibility of learned societies, which act on behalf of the scientific community. Interestingly, the Administrative Tribunal rejected this argument and claimed that there is no foundation for the claim that learned societies have exclusive representation of the scientific community. The Agency, by appointing experts and asking them to formulate independent judgements on scientific products, has full legitimation. This apparently technical dispute has significant implications.

In essence, what this Association and in general several (not all) learned societies claim is that the expertise resides in a representative structure. The appointment of experts should be done only upon their suggestions. This amounts to admitting that the expertise of members of the scientific community can only be appreciated by

members themselves, or that the knowledge they use in their expert judgement is completely tacit. No external entity, whatever the legitimation, can understand who the real experts are in a discipline.

I find this argument epistemologically weak and ethically (and politically) untenable. The epistemological argument is regressive since it negates the public and open nature of scientific knowledge. Given the public nature of scientific knowledge, it is possible for external authorities to identify traces of this activity and to select experts based on scientific reputational processes. Using transparent procedures such as the call for candidatures, the publication of CVs and the analysis of publications (with or without bibliometrics, depending on the discipline) is sufficient guarantee for the scientific community of a fair treatment. The claim of learned societies is also ethically untenable: active participation in associative work does not necessarily correlate with high research performance. Political participation in an association should never substitute for scientific reputation. As a matter of fact, no scientific journal would even consider the suggestion that the list of referees for its articles should be proposed by a learned society. In a recent book I have therefore argued that the solution to the issue of epistemic pluralism, and the associated issues of protection of evaluated subjects, should not be managed by delegating the choice of experts to learned societies (Bonaccorsi 2015).

6.6 Expert Selection Procedures

The suggestion that the selection of experts may follow procedures that guarantee epistemic pluralism is a serious one. As noted above, the scholars in Social and Political science are somewhat sceptical about the possibility of establishing procedures and rely mostly on the ethical conduct of panel members and referees. The scholars in Law, coming from a procedural culture, undertook this effort. An interesting suggestion is that panel members who have managed the initial referee process and who receive the referees' scores not be included among those that create the consensus group. This suggestion, however, is not realistic with respect to the small size of expert panels, in which there are a few, or even just one or two, members for each discipline. Another suggestion was to draw the referees randomly from a list. This suggestion was partially adopted by the VQR procedure. The panel members develop the list of referees based on their scientific reputation and availability. Matching between the research product and the referee, however, is not decided by panel members but is done randomly by the computer. If referees accept the task, the procedure stops. If they do not accept within a few days, the computer system moves to another referee. This procedure provides some benefits in terms of fairness and transparency.

Finally, the suggestions regarding selection procedures have been followed by the Agency also in other areas of its operations. One such area is the expert panel in charge of the classification of scientific journals (Bonaccorsi et al. 2015;

Ferrara and Bonaccorsi 2016). This is a delicate task. The main choices made by the Agency to ensure fairness and transparency, were as follows.

- The expert panel has a short term appointment (2 years), after which it is completely changed.
- Each year (or 2 years, depending on the scheduling of research assessment deliveries) an appeal procedure is made available whereby journal editors may offer counterarguments against the decision of the panel (to reject the application for class A journals, or the classification as scientific journal, respectively).
- Each 4 years a complete re-examination of the journal list is carried out in order to correct for false positives, following a detailed and transparent procedure.
- At each renewal of the panel careful consideration is given to issues of pluralism, adopting a rotating procedure (i.e., experts with a given orientation do not succeed each other).
- If a discipline or field exhibits significant and unsettled pluralism, the number of experts is exceptionally augmented in order to represent various options, even if this creates imbalances in the total number (and some extra cost).
- In the case of appeal of a journal against a decision made by the panel, different anonymous referees are appointed to re-evaluate the merit of the journal.

The provision of a short term appointment is aimed at avoiding locked trajectories. It is not unusual for experts at the end of their term to try to suggest candidates for the next term. These suggestions are taken with caution as they may create path-dependency in the orientation.

In summary, expert selection is a key part of the evaluation process. It cannot be done automatically (i.e., by randomly extracting from a list), nor can it be made completely neutral. What can be done is to build a number of procedures that minimise the error, ensuring that there is no path-dependency in the decisions through changes made at regular intervals and involving people having different perspectives.

6.7 Publicity

The suggestion that the names of referees be made public has often been proposed by learned societies in Legal studies. This suggestion has always been rejected by the Agency, and rightly so. The reason is clear: the anonymity of the referee process is a fundamental feature of the working of modern science. It is an institutional invention that goes back to the seventeenth century and is the backbone of the scientific communication system. It is no surprise that it is advanced by the legal scholar community, one in which the double blind peer review process is a relatively new practice.

As a matter of fact, the research assessment carried out in the VQR is not based on double blind peer review, but on single blind review. Members of evaluation panels and referees know the identity of the authors, while the reverse is not true. There is evidence that single blind review is inferior to double blind review with

respect to the ability to avoid distortions due to reputation and incumbency in the ex-ante selection (Seeber and Bacchelli 2017). At the same time, it is not possible to make the identity of authors anonymous in ex-post evaluation – a quick search on the Internet would allow the referee to locate the authors anyway. The question is whether abolishing anonymity on both sides would improve the situation, as is often requested by legal scholars. The answer is, most likely, no. Making the referee accountable to the authors would simply create a distortion in the incentive to provide thorough, professional judgements, fearing future retaliation on their academic careers.

What the Agency accepted, on the contrary, is the publication of the full list of referees at the end of the VQR. Referees were given two opportunities to explicitly request that their names not be published, after which the full list went online in 2014 (the list is still available at http://www.anvur.it/index.php?option=com_content&view=article&id=244&Itemid=306&lang=it). No negative reaction from scientific communities was recorded. Approximately 12,800 referees accepted to be published out of 14,000 in total.

7 Conclusions

This chapter has discussed a process of refinement of the definition of research quality criteria to be used in research assessment, and in particular in the peer review process. The refinement has covered the object of evaluation as well as the drafting and wording of questions addressed to anonymous referees, as well as the procedural rules for expert selection and management. The material has been drawn from published reports of independent research groups in the fields of Architecture, Legal studies and Social and Political science. Other interesting material was produced by groups in Humanities and in Business studies, but was not covered in this chapter.

Ochsner et al. (2016) made the compelling argument that communities in SSH (particularly in Humanities) do not share common criteria for the evaluation of research. On the basis of the evidence reported in this chapter, it may be argued that the definition of agreed criteria is a dynamic conversational process in which convergence is a possible outcome. The conversation has two components: an epistemic one and a social one. On the epistemic side, it is crucial to understand why scholars in SSH communities may hold different beliefs on research quality. Scholars and communities should be left free to voice their concerns and criticisms, but at the same time they should be asked to elaborate epistemic arguments with clarity. If concerns about preserving epistemic pluralism are well grounded, and not based on lack of reflection and/or hidden political agendas, then the evaluation process must provide technical and practical solutions to protect pluralism and to ensure maximum procedural fairness.

The social dimension is no less important. Scientific communities must be left free to engage in open-ended discussions about evaluation, their expectations, their concerns. Evaluation agencies should keep the dialogue permanently open, also accepting criticisms and complaints. There is a delicate balance between the

authority that is inevitably assigned by the legislative framework to evaluation agencies and the legitimation that only comes from the trust and prestige of scientific communities. The authority calls for distance, objectivity, standardisation and comparability; the legitimation calls for closeness, engagement, flexibility and adaptation to differences.

On the basis of the recent experience, I suggest that a process of convergence around research quality criteria that are progressively shared by scientific communities is possible and valuable.

Appendix 1

Architecture evaluation research group
Members: Cristina Bianchetti, Marco Biraghi, Marco Gaiani, Francesco Garofalo (coordinator), Sergio Poretti.

Social and Political Sciences evaluation research group
Members: Franco Garelli, Mauro Magatti, Leonardo Morlino (coordinator), Marino Regini, Irma Taddia, Salvatore Veca.

Humanities evaluation research group
Stefano Poggi (coordinator) Francesco Aceto, Giovanna Brogi, Alberto Camplani, Michele Corsi, Flavia Cuturi, Elena Dell'Agnese, Keir Douglas Elam, Vittorio Formentin, Filippo Maria Pontani, Adriano Roccucci, Mirella Schino, Giovanni Solimine, Roberto Tottoli.

Legal Studies evaluation research group
Vincenzo Militello (coordinator), Mauro Bussani, Cosimo Cascione, Giuseppe Conte, Pasquale De Sena, Carlo Fusaro, Diana-Urania Galetta, Tommaso Greco, Raffaello Lupi, Cinzia Motti, Lorenzo Zoppoli.

Business Studies evaluation research group
Stefano Brusoni (coordinator), Paola Adinolfi, Arnaldo Camuffo, Corrado Cerruti, Vittorio Coda, Alfonso Gambardella, Gianni Lorenzoni, Riccardo Palumbo, Alberto Quagli, Sandro Sandri, Elita Schillaci, Paola Schwizer, Alessandro Ruggieri, Alessandro Zattoni.

References

Ahlgren, P., Colliander, C., & Persson, O. (2012). Field normalized rates, field normalized journal impact and Norwegian weights for allocation of university research funds. *Scientometrics, 92*(3), 767–780.

Archambault, E., Vignola-Gagné, E., Côté, G., Larivière, V., & Gingras, Y. (2006). Benchmarking scientific output in the social sciences and humanities: The limits of existing databases. *Scientometrics, 68*(3), 329–342. 1007.

Ardanuy, J. (2013). Sixty years of citation analysis studies in the humanities (1995–2010). *Journal of the American Society for Information Science and Technology, 64*(8), 1751–1755.

Ardanuy, J., Urbano, C., & Quintana, L. (2009). Citation analysis of Catalan literary studies (1974–2003): Towards a bibliometrics of humanities studies in minority languages. *Journal of the American Society for Information Science and Technology, 81*, 347–366.

Becher, T., & Trowler, P. R. (2001). *Academics tribes and territories: Intellectual enquiry and the cultures of disciplines*. Buckingham: Open University Press.

Bonaccorsi, A. (2012). Potenzialità e limiti dell'analisi bibliometrica nelle discipline umanistiche e sociali. *Nota di lavoro ANVUR*. Available at www.anvur.org/attachments/article/44/valutazione_aree_umanistiche_e_sociali.pdf

Bonaccorsi, A. (2015). *La valutazione possibile. Teoria e pratica nel mondo della ricerca*. Il Mulino: Bologna.

Bonaccorsi, A., Cicero, T., Ferrara, A., & Malgarini, M. (2015). Journal ratings as predictors of articles quality in Arts, Humanities and Social Sciences: An analysis on the Italian Research Evaluation Exercise. *F1000Res, 4*(196). 10.12688/f1000research.6478.1.

Bourke, P., & Butler, L. (1996). Publication types, citation rates and evaluation. *Scientometrics, 37*(3), 473–494.

Chubin, D., & Hackett, E. (1990). *Peerless science: Peer review and U.S. science policy*. Albany: State University of New York Press.

Committee on the National Plan for the Future of the Humanities. (2009). *Sustainable humanities: Report from the committee on the national plan for the future of the humanities*. Amsterdam: Amsterdam University Press. http://www.oapen.org/search?identifier=339995.

Cullars, J. M. (1989). Citation characteristics of French and German literary monographs. *Library Quarterly, 59*, 305–325.

Cullars, J. M. (1998). Citation characteristics of English-language monographs in philosophy. *Library and Information Science Research, 20*(1), 41–68.

Engels, T. C., Ossenblok, T. L., & Spruyt, E. H. (2012). Changing publication patterns in the social sciences and humanities, 2000–2009. *Scientometrics, 93*(2), 373–390.

Ferrara, A., & Bonaccorsi, A. (2016). How robust is journal rating in Humanities and Social Sciences? Evidence from a large-scale, multi-method exercise. *Research Evaluation, 25*(3), 279–291.

Finkenstaedt, T. (1990). Measuring research performance in the humanities. *Scientometrics, 19*(5–6), 409–417.

Georgas, H., & Cullars, J. M. (2005). A citation study of the characteristics of the linguistics literature. *College and Research Libraries, 66*(6), 496–515.

Giménez-Toledo, E., & Román-Román, A. (2009). Assessment of humanities and social sciences monographs through their publishers: A review and a study towards a model of evaluation. *Research Evaluation, 18*(3), 201–213.

Giménez-Toledo, E., Tejada-Artigas, C., & Mañana-Rodriguez, J. (2013). Evaluation of scientific books' publishers in social sciences and humanities: Results of a survey. *Research Evaluation, 22*(1), 64–77.

Giménez-Toledo, E., Mañana-Rodriguez, J., Engels, T. C., Ingwersen, P., Pölönen, J., Sivertsen, G., & Zuccala, A. A. (2015). The evaluation of scholarly books as research output. Current developments in Europe. In A. A. A. Salah, Y. Tonta, A. A. Akdag Salah, C. Sugimoto, & U. Al (Eds.), *Proceedings of the 15th International Society for Scientometrics and Informetrics Conference*, Istanbul, Turkey, 29th June to 4th July, 2015 (pp. 469–476). Istanbul: Bogazici University.

Glänzel, W., & Schoepflin, U. (1999). A bibliometric study of reference literature in the sciences and social sciences. *Information Processing & Management, 35*(1), 31–44.

Gorraiz, J., Purnell, P. J., & Glänzel, W. (2013). Opportunities for and limitations of the book citation index. *Journal of the American Society for Information Science and Technology, 64*(7), 1388–1398.

Gruppo di ricerca ANVUR sulla Valutazione nell'area giuridica (GRAVAG). Verso la nuova VQR: Idee e proposte per la ricerca giuridica. Presentation to the Association of learned societies in legal disciplines. Rome, 26 June, 2015.

Gruppo di ricerca sulla Valutazione in Architettura. (2015). Relazione finale. Available at http://www.anvur.it/index.php?option=com_content&view=article&id=871&Itemid=626&lang=it

Gruppo di ricerca sulla Valutazione nelle Scienze sociali e politiche. (2015). Relazione finale. Available at http://www.anvur.it/index.php?option=com_content&view=article&id=871&Itemid=626&lang=it

Guetzkow, J., Lamont, M., & Mallard, G. (2004). What is originality in the social sciences and the humanities? *American Sociological Review, 69*(2), 190–212.

Hammarfelt, B. (2011). Interdisciplinary and the intellectual base of literature studies: Citation analysis of highly cited monographs. *Scientometrics, 86*(3), 705–725.

Hammarfelt, B. (2012). Harvesting footnotes in a rural field: citation patterns in Swedish literary studies. *Journal of Documentation, 68*(4), 536–558.

Hammarfelt, B. (2014). Using altmetrics for assessing research impact in the humanities. *Scientometrics, 101*(2), 1419–1430.

Hammarfelt, B., & de Rijcke, S. (2015). Accountability in context: Effects of research evaluation systems on publication practices, disciplinary norms, and individual working routines in the faculty of Arts at Uppsala University. *Research Evaluation, 24*(1), 63–77.

Harley, D., & Krzys, A. S. (2011). *Peer review in academic promotion and publishing: Its meaning, locus, and future*. UC Berkeley: Center for Studies in Higher Education. http://escholarship.org/uc/item/1xv148c8.

Hellqvist, B. (2010). Referencing in the humanities and its implications for citation analysis. *Journal of the American Society for Information Science and Technology, 61*(2), 310–318.

Hemlin, S. (1996). Social studies of the humanities. A case study of research conditions and performance in ancient history and classical archaeology and English. *Research Evaluation, 6*(1), 53–61.

Hicks, D. (2004). The four literatures of social science. In H. F. Moed, W. Glänzel, & U. Schmoch (Eds.), *Handbook of quantitative science and technology research: The use of publication and patent statistics in studies of S&T systems* (pp. 476–496). Dordrecht: Kluwer Academic Publishers.

Hicks, D., & Wang, J. (2009). *Towards a bibliometric database for the social sciences and humanities-a European scoping project* (Annex 1 of the report 'Towards a bibliometric database for the social sciences and humanities-a European scoping project'). Sussex: Science and Technology Policy Research Unit.

Huang, M., & Chang, Y. (2008). Characteristics of research output in social sciences and humanities: From a research evaluation perspective. *Journal of the American Society for Information Science and Technology, 59*(11), 1819–1828.

Hug, S. E., Ochsner, M., & Daniel, H. D. (2013). Criteria for assessing research quality in the humanities: A Delphi study among scholars of English literature, German literature and art history. *Research Evaluation, 22*(5), 369–383.

Hug, S. E., Ochsner, M., & Daniel, H. D. (2014). A framework to explore and develop criteria for assessing research quality in the humanities. *International Journal for Education Law and Policy, 10*(1), 55–64.

Knievel, J. E., & Kellsey, C. (2005). Citation analysis for collection development: A comparative study of eight humanities fields. *The Library Quarterly, 75*(2), 142–168.

Knorr-Cetina, K. (1999). *Epistemic cultures: How the sciences make knowledge*. Cambridge, MA: Harvard University Press.

Kousha, K., Thelwall, M., & Rezaie, S. (2011). Assessing the citation impact of books: The role of google books, google scholar and scopus. *Journal of the American Society for Information Science and Technology, 62*(11), 2147–2164.

Kousha, K., & Thelwall, M. (2009). Google book search: Citation analysis for social science and the humanities. *Journal of the American Society for Information Science and Technology, 60*(8), 1537–1549.

Kyvik, S. (2003). Changing trends in publishing behavior among university faculty, 1980–2000. *Scientometrics, 58*(1), 35–48.

Lamont, M. (2009). *How professors think: Inside the curious world of academic judgment*. Cambridge, MA: Harvard University Press.

Lamont, M., & Huutoniemi, K. (2011). Comparing customary rules of fairness: Evidence of evaluative practices in peer review panels. In C. Camic, N. Gross, & M. Lamont (Eds.), *Social science in the making* (pp. 209–232). Chicago: University of Chicago Press.

Larivière, L., Archambault, È., Gringas, Y., & Vignola-Gagné, È. (2006). The place of serials in referencing practices: Comparing the natural sciences and engineering with social sciences and humanities. *Journal of the American Society for Information Science and Technology, 57*(8), 997–1004.

Leydesdorff, L., & Felt, U. (2012). 'Books' and 'book chapters' in the book citation index (BKCI) and science citation index (SCI, SoSCI, A&HCI). In *Proceedings of the ASIS&T Annual Meeting*, Vol. 49.

Leydesdorff, L., Hammarfelt, B., & Salah, A. A. A. (2011). The structure of arts and humanities citation index: A mapping on the basis of aggregated citations among 1157 journals. *Journal of the American Society for Information Science and Technology, 62*(12), 2414–2426.

Lindholm-Romantschuk, Y., & Warner, J. (1996). The role of monographs in scholarly communication: An empirical study of philosophy, sociology and economics. *Journal of Documentation, 52*(4), 389–404.

Linmans, J. A. M. (2010). Why with bibliometrics the humanities does not need to be the weakest link. Indicators for research evaluation based on citations, libarary holdings and productivity measures. *Scientometrics, 83*(2), 337–354.

Lo Sasso, M. (2011). Valutazione della ricerca e progetto: Intervista ad Andrea Bonaccorsi. *Techne, 02*, 86–91.

Mallard, G., Lamont, M., & Guetzkow, J. (2009). Fairness as appropriateness: Negotiating epistemological differences in peer review. *Science, Technology and Human Values, 34*(5), 573–606.

Martin, B., Tang, P., Morgan, M., Glänzel, W., Hornbostel, S., Lauer, G., & Zic-Fuchs, M. (2010). *Towards a bibliometric database for the social sciences and humanities-A European scoping project* (A report produced for DFG, ESRC, AHRC, NWO, ANR and ESF). Sussex: Sience and Technology Policy Research Unit.

Moed, H. F. (2005). *Citation analysis in research evaluation*. Dordrecht: Springer.

Moed, H. F., Luwel, M., & Nederhof, A. J. (2002). Towards research performance in the humanities. *Library Trends, 50*(3), 498–520.

Moed, H. F., Linmans, J. A. M., Nederhof, A. J., Zuccala, A., López Illescas, C., & de Moya Anegón, F. (2010). *Options for a comprehensive database of research outputs in social sciences and humanities* (Annex 2 of the report 'Towards a bibliometric database for the social sciences and humanities-a European scoping project'). Sussex: Science and Technology Policy Research Unit.

Nederhof, A. J. (2006). Bibliometric monitoring of research performance in the social sciences and the humanities: A review. *Scientometrics, 66*(1), 81–100.

Nederhof, A. J. (2011). A bibliometric study of productivity and impact of modern language and literature research. *Research Evaluation, 20*(2), 117–129.

Nederhof, A. J., & Zwaan, R. (1991). Quality judgments of journals as indicators of research performance in the humanities and the social and behavioral sciences. *Journal of the American Society for Information Science, 42*(5), 332–340.

Nederhof, A. J., Zwaan, R. A., De Bruin, R. E., & Dekker, P. (1989). Assessing the usefulness of bibliometric indicators for the humanities and the social sciences: A comparative study. *Scientometrics, 15*(5–6), 423–435.

Novick, P. (1988). *That noble dream: The 'Objectivity Question' and the American historical profession*. Cambridge, MA: Cambridge University Press.
Ochsner, M., Hug, S. E., & Daniel, H.-D. (2012). Indicators for research quality in the humanities: opportunities and limitations. *Bibliometrie—Praxis und Forschung, 1*(4).
Ochsner, M., Hug, S. E., & Daniel, H.-D. (2013). Four types of research in the humanities: Setting the stage for research quality criteria in the humanities. *Research Evaluation, 22*(4), 79–92.
Ochsner, M., Hug, S. E., & Daniel, H.-D. (2014). Setting the stage for the assessment of research quality in the humanities: Consolidating the results of four empirical studies. *Zeitschrift für Erziehungswissenschaft, 17*(6 Supplement), 111–132.
Ochsner, M., Hug, S. E., & Daniel, H.-D. (2016). *Research assessment in the Humanities. Towards criteria and procedures*. Dordrecht: Springer.
Ossenblok, T. L., Engels, T. C., & Sivertsen, G. (2012). The representation of the social sciences and humanities in the Web of Science. A comparison of publication patterns and incentive structures in Flanders and Norway (2005–9). *Research Evaluation, 21*(4), 280–290.
Piro, F. N., Aksnes, D. W., & Rorstad, K. (2013). A macro analysis of productivity differences across fields: Challenges in the measurement of scientific publishing. *Journal of the American Society for Information Science and Technology, 64*(2), 307–320.
Royal Netherlands Academy of Arts and Sciences. (2011). *Quality indicators for research in the humanities*. Amsterdam: Royal Netherlands Academy of Arts and Sciences. Available at https://www.knaw.nl/shared/resources/actueel/publicaties/pdf/qualityindicators-for-research-in-the-humanities.
Seeber, M., & Bacchelli, A. (2017). Does single blind peer review hinder newcomers? *Scientometrics, 113*(1), 567–585.
Sivertsen, G. (2010). A performance indicator based on complete data for the scientific publication output at research institutions. *ISSI Newsletter, 6*(1), 22–28.
Sivertsen, G. (2014). Scholarly publication patterns in the social sciences and humanities and their coverage in Scopus and Web of Science. In E. Noyons (Ed.), *Proceedings of the science and technology indicators conference 2014 Leiden* (pp. 598–604). Leiden: Centre for Science and Technology Studies.
Sivertsen, G., & Larsen, B. (2012). Comprehensive bibliographic coverage of the social sciences and humanities in a citation index: An empirical analysis of the potential. *Scientometrics, 91*(2), 567–575.
Tang, M., Wang, C., Chen, K., & Hsiang, J. (2012). Exploring alternative cyber bibliometrics for evaluation of scholarly performance in the social sciences and humanities in Taiwan. In *Proceedings of the ASIS&T Annual Meeting,* Vol. 49.
Thompson, J. W. (2002). The death of the scholarly monograph in the humanities? Citation patterns in literary scholarship. *Libri, 52*(3), 121–136.
Torres-Salinas, D., & Moed, H. F. (2009). Library catalogue analysis as a tool in studies of social sciences and humanities: An exploratory study of published book titles in economics. *Journal of Informetrics, 3*(1), 9–26.
Travis, G. D. L., & Collins, H. M. (1991). New light on old boys cognitive and institutional particularism in the peer review system. *Science, Technology and Human Values, 16*(3), 322–341.
Van Leeuwen, T. (2006). The application of bibliometric analyses in the evaluation of social science research. Who benefits from it, and why it is still feasible. *Scientometrics, 66*(1), 133–154.
Volkswagen Stiftung. (2014). *What is intellectual quality in the humanities? Some guidelines*. Hannover: Volkswagen Stiftung. Available at http://www.volkswagenstiftung.de/uploads/media/Humanities_Quality_Guidelines.pdf.
White, H. D., Boell, S. K., Yu, H., Davis, M., Wilson, C. S., & Cole, F. T. H. (2009). Libcitations: A measure for comparative assessment of book publications in the humanities and social sciences. *Journal of the American Society for Information Science and Technology, 60*(6), 1083–1096.
Whitley, R. (2000). *The intellectual and social organization of the sciences* (2nd ed.). Oxford: Oxford University Press.

Zuccala, A. (2012). Quality and influence in literary work: Evaluating the 'educated imagination'. *Research Evaluation, 21*(3), 229–241.

Zuccala, A., & van Leeuwen, T. (2011). Book reviews in humanities research evaluations. *Journal of the American Society for Information Science and Technology, 62*(10), 1979–1991.

Zuccala, A., van Someren, M., & van Bellen, M. (2014). A machine-learning approach to coding book reviews as quality indicators: Toward a theory of megacitation. *Journal of the American Society for Information Science and Technology, 65*(11), 1643–2330.

Research Quality Evaluation: The Case of Legal Studies

Ginevra Peruginelli and Sebastiano Faro

1 Introduction

Scientific research assessment exercises are now becoming a widespread practice in all the European countries. The evaluation of scientific research is an extremely important instrument of public policy as a set of actions undertaken by public actors to address a collective problem that is the use of public resources for financing the developments of science.

One of the most critical topics in research evaluation is represented by the choice and definition of evaluation criteria.

The study of Martin and Irvine (1983) highlighted two scientific criteria: (i) quality, as the originality of the given contribution of knowledge; (ii) potential impact that the research output would have on the scientific community.[1]

These criteria are at the basis of the debate on assessment that has flared up in recent years, in particular in non-bibliometric areas like humanities, social and legal sciences. The heterogeneity of research and of scientific outputs, the specific nature of the communication channels – the monograph at the top – the absence of objective data to support the assessment, due to poor coverage in citation databases, are some of the factors that are fuelling the discussion (Moed et al. 1985a, b; Seglen and Aksnes 2000; Nederhof 2006; Martin et al. 2012).

[1] The potential impact is not easily measurable if not in the long term, both due to the limitations related to the difficulty of effectively communicating science and the difficulties experienced by scientists in recognising in time the scientific value presented by a specific author. The real impact is expressed mainly in the citation form, and can be calculated by counting the number of citations of research output in the short term.

G. Peruginelli (✉) • S. Faro
Institute of Legal Information Theory and Techniques of the National Research Council of Italy (ITTIG-CNR), Florence, Italy
e-mail: ginevra.peruginelli@ittig.cnr.it; sebastiano.faro@ittig.cnr.it

© Springer International Publishing AG 2018

A. Bonaccorsi (ed.), *The Evaluation of Research in Social Sciences and Humanities*, https://doi.org/10.1007/978-3-319-68554-0_5

In each country, evaluation exercises usually concern the whole range of scientific areas. The common feature of these exercises is the difficulty of comparing the results obtained from different scientific areas due to their peculiarities. This is particularly evident for the social sciences (in the Anglo-Saxon acceptation, including both law and economics) both in relation with other research areas and within themselves, as has been recognised since the old debate on the two cultures (Snow 1959, 1964). On the one hand there are large differences in the way in which research is organised, results are disseminated, and research outputs are cited. In addition, the audience is quite diverse, as, beyond the academic community, research in these areas is addressed to the vast network of policy-makers and not to general readers, nor entrepreneurs that translate the new knowledge into innovation. At the same time, social sciences in the broad sense are also internally heterogeneous, in particular as regards the degree of professional orientation and of internationalisation. These peculiarities are even more pronounced in legal sciences and particularly in legal scholarship.

The research outputs of legal science represent the work of legal scholarship, which, through monographs, articles, contributions to collections, reviews and manuals, describes and interprets aspects of law exhaustively, logically and systematically (Gutwirth 2009). In particular, the issue of evaluation of the scientific quality of legal publications is now at the centre of the debate in legal academia (among others Flückiger and Tanquerel 2015). Nowadays, in principle, peer review remains the preferred method for assessing the quality of legal scholarship: this is partly due to the failure in this area of a purely metrics-based system (Wooding and Grant 2003). In legal sciences, where research output is usually produced in long written formats, research performance is hard to assess using quantitative indicators: bibliometric methods are not sufficiently capable of measuring research performance in legal scholarship and are not considered reliable for the legal community (Gutwirth 2009).

Until recently, the notion of an explicit and independent assessment of the quality of legal research was considered unusual. Let us consider for example the case of legal journals. They often rely on editorial scientific boards in charge of accepting contributions: these boards do not use independent referees and most of the time they do not make their evaluation criteria explicit. As a result, a harmonised and transparent system of scientific evaluation of publications does not yet exist (van Gestel 2015). It is also not always true in practice that the most prestigious journals and publishers in the legal area have the highest acceptance threshold, and consequently are able to offer the best publications (Jerker and Svantesson 2009 p. 681). Moreover, to date at the supranational level there is neither an established classification of legal journals, nor a harmonised peer review system, nor databases that keep track of citations (van Gestel and Vranken 2011, p. 905).

On the one hand, the pressure to establish procedures and more explicit criteria (indicators) for the assessment of scientific quality in this area has increased over the last few years, but on the other hand there is no general consensus about the way in which the quality of legal research is conceived, thus hindering constructive discussion (Van Gestel et al. 2012; Nourse and Shaffer 2009).

In 1992 Edward L. Rubin, professor of law at the Vanderbilt University Law School (Rubin 1992) argued that there is no theory of evaluation for legal sciences. He stated that what actually leads legal academics to assess a work is based on an undefined concept of quality of judgements. This creates a number of conceptual and practical difficulties that produce confusion and unease in the sector. It is a matter of fact that many of the most heated discussions on legal scholarship concern evaluation and many of these are repetitive and non-productive due to the complete absence of a theory of evaluation. Professor Rubin directly tackles the question of what the foundation for evaluation should be and recommends an epistemological approach for formulating an evaluation theory. Some interesting issues are raised in his paper: the need for using criteria such as clarity, persuasiveness, significance, or the consideration by evaluators of their own uncertainty, especially in the case of topics somehow different from their discipline. A heated debate is still going on over criteria and even about the possibility of objective reliable evaluation. Major critical issues remain and no innovative solutions have been brought forward.

In this context, the aim of this chapter is to describe the emerging debate on the essential criteria for the assessment of legal scholarship, with a brief look at some experiences in Europe. The case study on Italy will be analysed further as well as the Italian debate which focuses, on the one hand, on the growth of the evaluation culture, claiming for the renewal in the assessment of legal research and, on the other hand, on the conservative attitude of legal scholars who are strongly resistant to change.

2 Legal Science as Special Context in the Debate on Research Evaluation

Legal scholarship has an ambiguous identity, somewhere between the humanities and the social sciences, sharing characteristics of both. It shares many of the peculiarities of social sciences as law is a social phenomenon, but where normativity and legal certainty are concerned, legal scholarship is probably closer to humanities. At the same time law is really distinct from both. There is an interesting debate on the role of academic legal research that can be considered as focusing on "law as a practical discipline", "law as humanities" and "law as social sciences" (Siems and Mac Sithigh 2012). All three categories are relevant within the context of legal academia and can be used to map how far institutions, individuals and legal cultures belong to one or more of them.

Every description of law includes a series of interpretations and offers hypotheses about the meaning and scope of legal concepts, rules, principles, that may be confirmed or falsified through scientific research. Roughly speaking, legal scholarship is a scientific discipline in its own right with a methodology that, in its core characteristics, is comparable to the methodology used in other disciplines. However, there is no agreement among legal theorists on the nature of legal doctrine

as a discipline, even independently from differences among national traditions of legal scholarship. Starting from the way legal doctrine has been practised in the course of history, in most countries we can consider it as a mainly hermeneutic discipline also with empirical, argumentative, logical, and normative elements. Description of law is closely linked to its interpretation and the legal scholar is formulating hypotheses about its existence, validity and meaning (van Hoecke 2011).

All of this calls for legal science to be regarded as a peculiar case study in the context of research evaluation.

As already stated, the field of law is more and more remote from bibliometrics. However it is interesting to note that Shapiro (1992) demonstrated that bibliometrics had a long history in the field of law, originated in England and in the United States, even prior to its introduction in the scientific literature. Since the early nineteenth century computing and counting systems referring to legal documentation were in place. Citation indexes began to be produced in the form of tables for judgement quotes (cases). Dating back to 1860 a volume containing a comprehensive citation index was published, followed, a few years later, by texts containing a citation analysis of courts' decisions.

Furthermore, the roots of the citation indexes can be traced back to the practice of publishing, in volumes containing judicial reports, tables of cases that are cited by sentences. A series of reports and indexes were then published, including the cases cited or referenced in a narrative way in the text or in the notes, and an alphabetical lists of cases followed by brief notes on subsequent decisions relating to the authority that issued that decision.

The first volume of citation indexes was published by H.H. Labatt in 1860, arranged alphabetically by name of cases, with brief notes on the following case treatment (Labatt 1860). Then followed another volume compiled by William Wait (Wait 1872). Later, on the heels of the work done, in 1873 Frank Shepard started to print citations to cases of the Illinois Supreme Court, and the work was then extended to a national system of production of printed volumes containing citations by subsequent decisions, laws and other legal sources. Eugene Garfield had the idea of using the Shepard citation principle as an indexing technique for medical literature (Garfield 1979). Then the Science Citation Index was born.[2]

Despite such early interest in bibliometrics, today legal science shuns evaluation practices based on quantitative tools such as bibliometrics, or the use of quantitative indicators that are based on the calculation of citations for the assessment of researcher impact. As Posner remarks:

> Unfortunately quality of scholarship is difficult to measure. It is no good looking at scholarly citations. I am an enthusiast for citation studies, but while they can be used (with a grain of salt) to evaluate scholars within fields, they cannot be used to evaluate field – really a group of fields – as broad as interdisciplinary legal scholarship (Posner 2002).

[2] Shepard's Citations Service is a search service making it possible to reconstruct the history of a court case, going back to the previous levels of judgement until the levels of appeal. In addition, the service makes it possible to see other citations of the judgement by other judges, or journals, or other subsequent resources http://www.lexisnexis.com/en-us/products/shepards.page

The main reason for this approach is perhaps the inability of the bibliometric community to create valid and representative indicators for the academic activities of legal research communities. However, under the influence of exact sciences, legal scholars feel somehow "forced" to follow the evaluation model of the hard sciences (van Gestel and Vranken 2011).

It is clear that anything concerning the assessment of legal research cannot fail to take account the following specific characteristics.

First, *legal science is not monolithic.* While legal sciences are profoundly different from the exact sciences, strong differences exist among the various areas of law. For example, the communication tools used by legal scholars in tax law are very different from those of legal comparatists. Law is not monolithic, neither are the processes behind it. There are different branches of law and different schools of interpretation within these branches.

Second, there is a *strong link between legal scholarship/legal publishing and practice*. Many scientific journals host contributions of practitioners (judges, professionals, public servants) or articles dedicated to operational issues, which perhaps in other scientific disciplines would have a different destination. Many scientific journals publish contributions on practical issues related to the legal profession. It is not at all uncommon for a journal to publish the contribution of a judge, especially if serving on a higher court. Legal scholarship sits in a remarkable and singular position as it is not only the science of law but also, and in particular in commonwealth countries, an influential and authoritative source of law (Gutwirth 2009).

Third, there is a *plurality of "literary genres" and of their specific value.* In addition to monographs and journal articles, search results can be published in other forms like comments on judgements and on legislation, and reviews. Each type of product also has its specific value: legal scholars often disseminate new knowledge through monographs (Di Raimo 2015) and they enrich their CVs through book chapters and journal articles.

A fourth feature to be considered concerns the s*ocietal impact of legal research outputs*. Legal science has a strong impact on society's core principles, such as justice, freedom, human dignity and solidarity. It is certainly not the only science to deal with these values, but it comes close to these principles through the exceptional nature of the law. One of the objectives of a legal scholar of positive law is to affect the "outside" world with his/her research. The purpose of a contribution can be, for example, to convince a court to change direction on a given issue, or to receive a particular reconstruction and interpretation of the rules imposed by the legislator. It can happen that observations/interpretations of legal scholarship can be the source from which significant changes in the legal domain originate. This kind of impact, however, is difficult to measure as in some jurisdictions case law and regulations do not contain doctrinal citations.[3] After all the use of assessment tools that affect the

[3] In Italy, for instance, Section 118, third paragraph of the rules for the implementation of Italian Civil Procedure Code, also applicable in criminal courts, provides that any quote of legal scholarship and authors should be omitted in the judgement.

socio-economic aspects of research is one of the most complex and difficult challenges in which academia and legal research have been involved in the last few years. A relevant issue is linked to the hunting of experts and reviewers committed to assessing impact (Holbrook and Frodeman 2011). This activity requires disciplinary expertise in economics and social sciences and a deep knowledge of the relevant fields of application, not always present in a legal scholar's curriculum.

The fifth point concerns the s*ystem-bound nature of law.* Despite legal science has become increasingly internationalised, the national legal framework is a powerful brake on competition between researchers outside national borders. In fact the relevance of national legal scholarship is mainly undisputed. This is true for the legal branches of positive law, but it is not always true for scholars of international and comparative law, philosophers or sociologists of law, historians of law. Contributions that have a decisive impact on national laws are rarely published abroad. National legal systems are built on different foundations and governed by rules that foreign legal scholars are not always able to understand. An Estonian law journal, for example, might have little incentive to publish a contribution of a Portuguese scholar on a *Supremo Tribunal de Justiça* decision if this judgement is only relevant in Portugal.

Finally, another specific factor is related to *the role of the author.* Unlike colleagues in the exact sciences, a legal researcher usually works alone. His/her focus is on the object of research, on which he/she gives a personal and original interpretation in the framework of an absolute pluralism of views: the contributions of other scientists are often used to outline the context of reference. It should also be taken into account that not always author and reader coincide. Bibliometric studies (Schloegl and Gorraiz 2011) based on the comparison between downloads and citations in the exact sciences show that there is strong correlation between the readers and the citers. On the contrary, this is not the case in the social and legal sciences.

In summary, assuming that academic quality and its assessment is very context-dependent, all the above mentioned features must be taken into account in defining what quality research is about, or when a quality measurement model for legal science is to be provided.

3 Research Assessment Exercises in the Legal Domain: A Brief Look at Experiences in Europe

On a European level the assessment of research has been a widespread and well-established practice for many years. This is the case of the United Kingdom and the Netherlands (the former for almost 30 years, the latter for at least 15 years): the position of British and Dutch universities in the main world rankings, and the dissemination of publications in major international scientific journals are remarkable factors in these countries.

The effects arising from evaluation programs, thus the connection between evaluation and the criteria for grant awarding and funding are of course different from country to country (AUBR 2010). A brief analysis of some national experiences (the Netherlands, United Kingdom, France and Germany) is presented below.

3.1 The Netherlands

According to the *Inhoud Standaard evaluatie Protocol* 2015–2021 (Standard Evaluation Protocol 2015) in the Netherlands the debate on the evaluation of research in the humanities and in the social and legal sciences is particularly intense. Within the Dutch scientific community there are some concerns about the possible misapplication of indicators for the evaluation of research and also about a certain abuse of bibliometrics.

These concerns are at the heart of the Leiden Manifesto for research metrics developed in 2015 by several experts in the field (Leiden Manifesto 2015). The ten principles of the Manifesto, in the form of recommendations, refer to key issues such as the supporting role of quantitative in respect of qualitative evaluation, the consideration of the differences between disciplines in the publications and citation methods, the diffidence towards meticulous enumeration and comparisons, for example, in the calculation of citations (as opposed to the use of multiple indicators), regular monitoring of indicators' validity in light of the many changes that have taken place in research and in the objectives of evaluation.

With respect to research in the legal sciences, a recent survey conducted in 2015 on researchers and professors of Dutch law faculties about evaluation of the quality of research shows how Dutch jurists give preference to qualitative rather than quantitative methods (van Boom and van Gestel 2015). While scholars say they prefer blind peer review (anonymous authors and reviewers) rather than reviews carried out by editorial boards, they are doubtful about peer review by independent external experts due to the difficulty of finding neutral experts in the field of legal research. According to the survey, a lack of trust is often expressed with respect to the Netherlands Organisation for Scientific Research (NWO),[4] the main body of research funding committed to select the best research proposals. At the same time, however, respondents seem satisfied with the research policy of their faculties and with the possibility they are given to choose research topics, language in which to publish, research methodology and participation in broader research projects.

Furthermore, the Centre for Science and Technology Studies (CWTS)[5] at Leiden University has set up a working group on the development of indicators for the evaluation of research in the social, human and legal sciences (WISSH),[6] which is working in two directions: the first focuses on how scholars from these disciplines

[4] https://www.nwo.nl/en

[5] https://www.cwts.nl

[6] See https://www.cwts.nl/research/working-groups/social-sciences-humanities-law.

conduct their research and communicate their results both within academia and outside; the second direction deals with the implementation of possible indicators to assess the performance of scholars in these specific domains.

3.2 United Kingdom

The UK has long been engaged in research evaluation exercises that have evolved in the methodology and in the object of evaluation. The programme called REF (Research Excellence Framework)[7] is the current system for assessing the quality of research in publicly funded UK higher education institutions (HEIs). It replaces the Research Assessment Exercise (RAE), last conducted in 2008.

The results of the 2014 REF[8] show the high quality and enhanced international standing of research conducted in UK and demonstrate that the quality of submitted research outputs has improved significantly since the 2008 RAE, consistent with independent evidence about the performance of the UK research base. Furthermore, the results also demonstrate the wide range of outstanding social, economic and cultural impact that research has contributed to.

It is interesting to note that legal research and scholarship within universities were only established in the 1960s while before the teaching of law was mainly in the hands of practitioners. Since that period teaching and research at universities have progressed enormously and today cover the entire range from doctrinal, contextual and theoretical approaches.

The most reliable method for assessing quality legal research has been and remains peer review. This is mainly due to the fact that it may be used to evaluate any product type (monographs, book chapters, collections of essays, journal articles – the majority of submissions – comments on judgements, research reports, databases).

The previous evaluation exercise (RAE 2008) saw 67 institutions operating in the field of law, which submitted the research work of 1700 academics (Douglas 2015). University law departments were ranked according to their outputs, research environment, value and esteem obtained for their work. As regards criteria for the assessment, it is worth mentioning that the panel has not considered place or type of publications as appropriate indicators for the assessment of research outputs.

In REF 2014 great attention was paid to reduce the variability between the scores given by reviewers in different disciplines and to produce as much consistency of standards as possible. Great consideration was given to measure the performance from an international perspective and for this purpose international academic members were included in the panel. The instructions for reviewers made clear that all types of research and outputs were subject to evaluation on an equal basis and for

[7] See http://www.ref.ac.uk/

[8] REF 2014: The results, December 2014. http://www.ref.ac.uk/media/ref/content/pub/REF%2001%202014%20-%20full%20document.pdf

law this meant examining and assessing mainly books as well as their parts, journal articles, digital material, web content, case notes, conference papers, research reports, working papers, critical reviews, textbooks where original research was included (although this issue raised several problems due to the difficulty to assess originality for such material). Originality, significance, rigour were the criteria for assessing the outputs and this was done by attributing a score to each output determined on the basis of a scale from zero to four.

The REF 2014 Panel overview[9] reported that the areas of law covered by the research activity under evaluation were: international law, criminal law, criminology, theory of law, public law and human rights, commercial law, European law. Major contributions in comparative law concerned common law rather than civil law. As for the methodology of legal research, a remarkable variety of approaches were highlighted: doctrinal, theoretical, historical, socio-legal, critic and also empirical. Furthermore, the Sub panel for law (Sub panel 20) noticed an increasing sophistication of legal arguments with strong social implications, and evidence of this comes from the many PhD published monographs dealing with innovative and challenging issues.

The new element of the last evaluation exercise was the impact that the research produced. For this purpose case studies were to be provided showing the influence of the piece of research in the outside world, its change and tangible benefits for the economy, society, culture, services beyond academia. As stated in REF, impact relevant to law include improvements in legal frameworks for regulating property rights, more effective dispute resolutions, implementation of environmental regulations and other matters in public policy, influence of research in the interpretation of law and in the formulation of new laws and regulation of citizens' life at a national, European and international level.[10]

The evaluation of the economic and social impact of research has encountered several problems as such consequences are difficult to envisage in the field of law (Glendinning 2010; Piazza and Matteucci 2015). This criterion has raised a heated debate in the national press and within the scientific community regarding the substantive meaning of this kind of assessment and the identification of the most appropriate tools to measure the impact.

An additional element of assessment is the research environment provided by the submitting institutions. This information, given in narrative form, explains the strategy and objectives of research, people involved and development of careers, infrastructure and contribution to the discipline. The criteria for such evaluation are the vitality of the institution and its programmes, i.e., its research strategy, and, quite important, the ability to support the research programmes in the future, in other words the sustainability of the research environment.

In July 2015 an important report was published, commissioned by the Higher Education Funding Council for England (HEFCE), the government funding body of

[9] REF 2014 Panel overview reports: Main Panel C and sub-panels 16–26, p. 70–74. www.ref.ac.uk/media/ref./content/expanel/member/Main%20Panel%20C%20overview%20report.pdf

[10] REF, Assessment Framework: guidance and criteria, 2011. www.ref.ac.uk/about/guidance/

the English university system. This Report (Snowden et al. 2015) states that a wider use of quantitative indicators could be useful in the transition to a more open research system. However, excessive use of this type of indicators without adequate peer review may have negative consequences, lending itself to misunderstandings and disputes. According to the expert group that prepared the report, the assessment by metrics-based analysis should support, not supplant, expert judgement. With the report, an annotated review[11] of the professional literature and a simulation of evaluation[12] obtained by using quantitative indicators have also been published. This analytical work has generated several comments in the scientific community and even in the press, as proof of how this subject is also interesting for the public as a whole. For example *The Guardian*[13] has stated that "the metric tide is going up", even in a situation where peer review, with its imperfections, is to be maintained as the primary basis for the evaluation of research. The newspaper highlights how a mature system of research needs a balanced and variable judgement as well as quantitative and qualitative indicators.

3.3 France

In France the *Observatoire des sciences et des techniques* (OST)[14] and the *Haut Conseil de l'évaluation de la recherche et de l'enseignement supérieur* (HCERES)[15] that replaced AERES (*Agence d'évaluation de la recherche et de l'enseignement supérieur*) are in charge of the evaluation of the higher education institutions and of the *Réseau des écoles de service public*. HCERES is responsible for:

- Evaluating higher education institutions and groupings, research bodies, scientific cooperation foundations and the French National Research Agency, or, where applicable, overseeing the quality of evaluations carried out by other bodies.
- Evaluating research units on the request of the overseeing institution, in the absence of validation of evaluation procedures or in the absence of a decision by the overseeing institution to use another evaluation body or, where applicable, to have validated research unit evaluation procedures carried out by other bodies.

[11] Literature Review. Supplementary Report I to the Independent Review of the Role of Metrics in Research Assessment and Management. Available at http://www.hefce.ac.uk/PUBS/RERE PORTS/YEAR/2015/METRICTIDE/TITLE,104463,EN.HTML

[12] Correlation analysis of REF2014 scores and metrics. Supplementary Report II to the Independent Review of the Role of Metrics in Research Assessment and Management. Available at http://www.hefce.ac.uk/media/HEFCE,2014/Content/Pubs/Independentresearch/2015/TheMetricTide/2015_metrictideS1.pdf

[13] Available at http://www.hefce.ac.uk/media/HEFCE,2014/Content/Pubs/Independentresearch/2015/TheMetricTide/2015_metrictideS2.pdf

[14] http://www.obs-ost.fr/

[15] L'Haut Conseil de l'évaluation de la recherche et de l'enseignement supérieur was created with Law no. 660 of 22 July 2013. See http://www.hceres.fr/

The mission of HCERES includes the definition of quality assessment criteria and the publication of evaluation procedures. It does not provide a specific regulation or policy framework for legal research and no rating or ranking of journals exists[16].

It is worth mentioning that a number of scientific delegates for legal disciplines that have contributed to the evaluation of research units after the creation of HCERES have resigned in the form of an open letter. Their letters, addressed to its president, express some doubts about the work of HCERES, particularly concerning the lack of transparency, independence and impartiality. The peer review should be organised in a way that the experts of the visiting committees and the scientific delegate representing the HCERES belong to the same discipline of the evaluated research unit. Furthermore, reliability and transparency of assessment dictate that the report prepared under the authority of the evaluation committee's chair should be subjected to a contradictory process so as to ensure the acceptance of the conclusions and therefore their usefulness for the university and especially for the research team.

A particular mention should be given to the publication of HCERES 2016–2020 Strategic Plan[17] which reaffirms the principles of independence, transparency and impartiality that underpin ethical evaluation. With this document, Michel Cosnard, President of HCERES, hopes to strengthen the country's confidence in the evaluation of its higher education and research system

3.4 Germany

Over the past few years Germany has put strong effort into better managing and funding academic institutions with a view to improve the research system. The intense debate that has been generated around it also involved legal scholarship.

The interesting Report of November 2012 "*Perspektiven der Rechtswissenschaft in Deutschland. Situation, Analysen, Empfehlungen"*[18] elaborated by the German Council of Science and Humanities (*Wissenschaftsrat*)[19] is driven by the idea that it is necessary to strengthen the role of legal science by intensifying exchanges within and outside the discipline, opening up to an ever growing internationalisation in the

[16] AERES, in 2008 and later in 2010, published a list of legal journal titles that has not been updated by HCERES

[17] http://www.hceres.com/PUBLICATIONS/Institutional-documentation-Legal-texts/Strategic-plan-2016-2020

[18] The English version entitled "*Perspectives of Legal Scholarship in Germany. Current Situation, Analyses, Recommendations*" is available at http://www.wissenschaftsrat.de/download/archiv/2558-12_engl.pdf. The original version is available at http://www.wissenschaftsrat.de/download/archiv/2558-12.pdf.

[19] http://www.wissenschaftsrat.de/home.html.

research environment. The report (German Council of Science and Humanities 2012, p. 33) states that

> The internationalisation and Europeanisation of law – particularly the gradual opening of the closed legal system of the nation-state (the framework which gave rise to both the German Civil Code and the constitutionally guaranteed fundamental rights) – demands a methodology that critically reflects and integrates international perspectives

and again

> Since legal scholarship is concerned with norms, scholarly progress does not simply refer to the production of original knowledge, but aims also at securing and integrating existing knowledge into new contexts. The most innovative research often emerges when a number of different legal fields meet.

Therefore, according to the German Council of Science and Humanities' recommendation, experts in the discipline are expected to work towards the development of a framework for an unambiguous evaluation process in the specific domain of legal research.

The Council's recommendations address the need for an agreement among the experts involved in the assessment of scholars and their outputs to establish quality standards and evaluation procedures specifically designed to assess quality legal research. It is mainly the professional associations that are called upon to operate in such direction also in consultation with the German Jurist Faculty Association, as the official representative of the law faculties and departments. The main objective is to have available improved evaluation procedures to assess research outputs' quality and German legal scholarship on the whole. Also recommended is the strengthening of exchange and cooperation among experts in the discipline in a way that more efficient and informed criteria and procedures be followed.

4 Legal Research Evaluation in Italy

4.1 Introduction

In Italy the evaluation of university-based research has witnessed two main periods:

1. Research produced in the period 2001–2003 was evaluated by the National Committee for the Evaluation of Research (CIVR, *Comitato di Indirizzo per la Valutazione della Ricerca*)[20], which published its results in 2006.
2. In 2006, a new Agency for the Evaluation of University System and Research (ANVUR, *Agenzia Nazionale per la Valutazione del Sistema Universitario e della Ricerca*)[21] was established, taking over CIVR's role. However ANVUR only became operational in 2011.

[20] http://civr.miur.it/

[21] http://www.anvur.org

ANVUR is responsible for quality assurance of teaching (also at a doctorate level), research assessment, definition of criteria for national selection procedures and career progression for academic staff. It is also committed to drawing up guidance for the evaluation units in charge of allocating national funding by request of the Ministry of Education, Universities and Research.

A first exercise of research quality evaluation was carried out for the period 2004–2010. It produced 14 final reports on all areas of science. This documentation, together with the 2013 ANVUR final report[22] is the foundation of today's debate and of the last 2011–2014 evaluation exercise.

The 2011–2014 evaluation exercise was organised in 16 evaluation areas. For each area, the ANVUR Governing Board selected a group (Group of Experts for Evaluation, hereinafter GEV) composed of qualified Italian and foreign experts, selected according to their scientific expertise and previous experience in evaluation procedures.

With respect to legal science, GEV (GEV 12) followed the above criteria and adopted only peer review without being supported by citation databases. According to the VQR 2011–2014 call,[23] GEV 12 accepted the following eligible categories of research output[24]:

1. Scientific monographs and related output such as scientific treaties, critical editions, critical manuals (not purely educational or explanatory works), book translations (if considered as an hermeneutic work with a critical approach).
2. Journal contributions, including for example review essay, forum contribution on invitation by the journals' editorial boards, comments on court judgements.
3. Book contributions (including preface/postface essays), edited collection of essays, critical items in dictionary or encyclopaedia.
4. Other research output (only if accompanied by evidence in relation to the publication/production date) such as databases and software, if characterised by originality and legal significance.

GEV 12 decided not to include some categories of research outputs as they were not considered "sufficiently significant" for research assessment in the legal area. These were, for example, concordances, scientific comments, research bibliographies, letters, catalogues with introduction essay, part of catalogues, repertoires, corpora.

[22] http://www.anvur.org/rapporto/

[23] Bando VQR 2011–2014, p. 5 and 7. http://www.anvur.org/attachments/article/825/Bando%20VQR%202011-2014_secon~.pdf

[24] Evaluation of Research Quality 2011–2014 (VQR 2011–2014): Criteria for Evaluation of Research Output by Group of Evaluation Expert for Area 12 – Legal Sciences. Available at (http://www.anvur.it/attachments/article/855/Criteri%20GEV%2012_Eng.pdf), p. 8 and subsequent available in English

4.2 Peer Selection, Assessment Criteria and Indicators

Each category of research output is primarily assessed by two external referees, each identified independently by two different GEV members who are in charge of the Scientific Sector (SSD) to which the research output belongs. Alternatively, the evaluation is carried out internally by GEV members, making sure there is no conflict of interest and following the same procedure adopted for external referees.

The selection of external reviewers, either Italians or foreigners, follows the principle of fairness, objectivity and impartiality and the anonymity in the preparation of reviewers list. Reviewers are chosen from amongst the most authoritative scholars and specialists in the field. The evaluation results and their association with the expert reviewers is not publicised.

Evaluation by external reviewers or GEV internal members is based on a special form. This has been constructed according to three evaluation criteria, namely (i) originality, (ii) methodological rigour, (iii) proven or potential impact. There is also an open field for other comments. A mandatory short summary statement of the reasons justifying the scores is required.

For the three criteria mentioned, which altogether contribute to quality assessment, GEV 12 has indicated a set of indicators for each of them in order to better specify the characteristics of legal research outputs and to support reviewers in their work.

One part of the English version of the mentioned GEV 12 working document entitled "Evaluation of Research Quality 2011–2014 (VQR 2011–2014): Criteria for Evaluation of Research Output by Expert Evaluation Group for Area 12 – Legal Sciences"[25] is quoted below:

> Sub a) originality:
>
> – [the output] presents new acquisitions, for example in relation to sources of law, doctrine or jurisprudence; or to legal research topics, also interdisciplinary, previously unknown or less researched;
> – or it develops critical arguments relating to theoretical frameworks or pre-existing historical-legal contexts, useful to the advancement and/or deepening of knowledge in the field of reference and/or the discipline in general;
> – or it adequately stands within the legal discipline and relevant literature on the subject matter, proposing new interpretative lines in relation also to materials already analysed;
> – or it introduces developments or innovative methods of analysis, which lend themselves to being applied to other topics and/or issues, contributing to the evolution of concepts and existing theories;
> – it lacks originality if the contribution is primarily descriptive and is merely a review of existing data and opinions already present in other publications.
>
> Sub b) methodological rigor:

[25] *Ivi*, p. 12 and subsequent.

– it is characterised by the use of an appropriate methodology for the type and the object of the work in line with the state of the art, which can also be used by other scholars who wish to deal with the same or analogous/similar theme;
– it applies this methodology in a rigorous manner, for example in relation:
  - to the appropriate sources or documentation used;
  - or consistency in the development of the arguments;
  - or the ability to interrelate the subject analysed, its aims and its possible theoretical consequences and/or its application.

Sub c) attested or potential impact on the international scientific community of reference:

– it is the result of an explicit partnership with researchers and research groups in other countries (e.g., as part of international projects);
– or it has had, or may have, a circulation within the scientific community in other countries for linguistic reasons (publication or translation in another language) or editorial (presence in collections of international essays or in international journals) or has generated interest (e.g., because of reviews or references in other countries' literature);
– or due to its structural characteristics – e.g., issues and/or questions dealt with, theoretical implications and/or methodological approaches, possibly interdisciplinary, and the capacity of dialogue with the international research community and scholars of other legal systems – it should be considered a significant contribution to current and future research within the relevant scientific domain and has consequently been a primary reference, or at least important, or valuable to those who were to deal with the issue, even foreigners.

According to GEV 12 this third criterion is of fundamental importance in light of the profound changes that have hit the law and legal research even in legal sectors traditionally characterised by strong national roots. The attention beyond the national context is a relevant added value to evaluate the quality of a contribution along with the other two criteria of originality and methodological rigor.

Of course the risk of subjectivity is always quite high, as is the case of non-convergence of scores. In this case GEV 12 has provided for the creation of consensus groups. The consensus group, with an explanatory report, proposes the final score of a submission to GEV. In case of strongly divergent assessments a third expert can be called in support of the evaluation.

5 The Italian Debate in Legal Sciences

In recent years much debate has arisen in Italy over legal research evaluation. Some initial contributions pointed to the risk of bureaucratisation, disproportion between promotion and control, and the failure to distinguish more clearly between evaluation and measurement (Cassese 2013). Furthermore the effectiveness and

significance of research evaluation, including whether or not to adopt the measures proposed by ANVUR to assess scientific production, has been largely discussed.

The criteria used by ANVUR for the 2004–2010 evaluation of legal outputs have been criticised for not having involved the entire scholarly community in the establishment of indicators and "robust" methods in support of assessment. However, it is evident that the launch of the research evaluation process put in place by ANVUR has led to widespread awareness of the importance of evaluation that was mostly absent before. In recent years, mainly in connection with the establishment of ANVUR, a wide body of literature has been produced in Italy on the assessment of research quality, with reference to the whole range of scientific areas as well as to legal science (Di Cristina 2013, 2014, 2015; Della Cananea 2014; Militello 2014; Marcello 2014; Gasparini Casari 2011).

The Italian debate focuses on some aspects that are briefly analysed below, based not only on the numerous and interesting documents produced by associations and scientific societies connected to legal science, but also on the intensive work that GEV 12 has done in the ANVUR research evaluation exercises.

Reference is made here to the final report by GEV 12 within VQR 2004–2010[26] and to the already mentioned working document produced by GEV 12 on criteria for evaluating research outputs for VQR 2011–2014,[27] as well as to the opinion on the use of bibliometric indicators for the assessment of legal research expressed in 2014 by the Conference of the Scientific Association of the Legal Area (CASAG).[28] The main topics under debate are briefly presented below.

Legal Research Outputs Under Evaluation In Italy there are many literary genres for legal studies and research. In addition to scientific monographs, either educational or academic, journal articles, chapters of books, as well as comments on case law constitute a very important literary genre. Case notes, in fact, for their extent and substance, are often similar to journal articles. They also maintain an undisputed importance in terms of quality, but even encyclopaedias and treatise entries are quite a widespread genre, especially by mature scholars. Of course, these genres have different weights and quality levels and the discussion on quality levels is still open.

Monographs and journal articles were the "literary genres" on which, before VQRs, a thorough discussion within national scientific societies had been settled. Journal articles and monographs under evaluation in the VQR 2004–2010 each amounted to 30% of the total of outputs. Therefore, none of these genres have quantitative dimensions like the third group of outputs, which includes both book chapters and contributions to other collective works, which account approximately for 40% of the total.

[26] http://www.anvur.org/rapporto/files/Area12/VQR2004-2010_Area12_RapportoFinale.pdf

[27] See note 25.

[28] http://www.associazionedeicostituzionalisti.it/download/Jvy0AePS84owLDwij3bSn07t0n6ZcqEYKVqDd4Ljuxk/casagparere-indicatori-biblioetrici-a.pdf

Also in VQR 2011–2014[29], 95% of the outputs presented and evaluated in the legal field belonged to the three main literary genres. However, compared to the previous VQR there is the primacy of journal articles, and no longer of the contributions in edited volumes. The first category reached 38.04% of the total (they were 32.76% in the previous VQR), the latter dropped to 30.57% (from the previous 36%), so close to the literary genre of monographs that currently represent 26.17% of the total. The growth of journal articles may be due to the spread of anonymous peer review procedures in the journal articles, which are perceived by the authors as a guarantee of a better quality of content. Of course, disaggregating the data in the various branches of law there are some sectors where the primacy is still held by contributions to edited volumes compared to journal articles (in particular, comparative private law, constitutional law, international law, European law, history of law, criminal law procedure). In the legal field, monographs remain the kind of publication that receives the highest percentage of judgements of excellence (9.99%), followed by journal articles (8.27%).

The establishment of a classification system of literary genres in use in scientific legal research based on a scale of complexity from monographs to comments on cases is a widely discussed topic in Italian academia. In fact, the choice of the type of scientific outputs and their weight directly impacts on the quantity and quality of the scientific production of legal scholars.

As previously mentioned the call for VQR 2011–2014 indicated which types of legal publications are eligible for evaluation and established clearly identified subsets for each type. The document does not offer a definition of any of these types of outputs. In particular, manuals and textbooks of purely educational and popular type were excluded from the scope of the VQR 2011–2014 for the legal area. It is important to note that a lively debate is currently underway on manuals, as it is not unusual to find critical manuals, not merely educational. While research monographs are to be distinguished from books whose purpose is merely to summarise and disseminate acquired knowledge, also in the form of manuals, a distinction should be made on a case-by-case basis, verifying whether outputs are manuals that are purely educational, without any aspect of originality.

Internationalisation This is a favourite theme of Italian academia, which reaffirms that internationalisation is not to be intended as simply writing in English or being cited in international databases. Support in this direction is provided by the motion approved on 20 April 2012 by the *Accademia dei Lincei*, (Accademia Nazionale dei Lincei 2012) which states that in the human sciences it is not possible to entrust research evaluation to bibliometric indicators recorded almost exclusively on publication in English language journals without taking into account: (a) the specific nature and the object of study of the disciplines (including Italian positive law); (b) the characteristics of disciplines where primarily monographs are produced, which

[29] VQR 2011–2014. Rapporto finale di area Gruppo di Esperti della Valutazione dell'Area giuridica GEV 12, p. 32 and ff. http://www.anvur.org/rapporto-2016/files/Area12/VQR2011-2014_Area12_RapportoFinale.pdf

are not recorded by bibliometric indicators; (c) the difficulty of accessing foreign journals not interested in specific national topics (e.g., Italian economy and society).

Legal sciences are by nature linked to national languages and many areas of research are strictly local so that the vehicular language is the national language, that of the legal system analysed. Empirical data demonstrating this aspect is provided by the works published in English: their percentage is 5.6% of the overall products presented by the Italian legal scholars for the VQR 2004–2010 assessment.[30] In addition, only 0.5% of these contributions are produced through an explicit partnership in the form of co-authoring with a university teacher or a researcher working abroad.

In Italy there is a small percentage of publications in languages other than Italian in public law areas as compared to private law areas, except for tax law and those sectors related to a European, comparative and international focus. In fact, constitutional law and public, administrative, criminal and ecclesiastical law are all under the average. Law journals that use the national language as the only language of communication are partly disadvantaged. The direct consequence is that these journals are excluded from international channels, and are often not included in the main citation databases at an international level. This is a crucial issue, as legal scholars do not necessarily consider English as the vehicular language, but value the country's own language, that is, the legal language as the only one being able to represent and reflect the features and value of a jurisdiction.

Internationalisation can, and perhaps should, be represented by the impact within the international scientific community, in co-authoring and in the composition of the editorial and scientific committees of editorial series and, therefore, in the level of prestige that the series gets from the community of scholars in the discipline outside Italy, even if the paper is written in Italian. It follows that the international, European and comparative dimension is certainly an added value, but it cannot and must not be a quality indicator.[31]

In fact, the GEV 12 for the 2011–2014 VQR states that internationalisation has its own value and even in areas not constitutively open to an international dimension, consideration of a scope beyond a national context is a relevant factor in assessing the quality of research products, certainly together with indicators like originality and methodological rigor.

The Value of Citations Citations have two main functions. First, they identify the resources that an author examined and used in developing and writing his/her work (documentary function of citation), and their second purpose is to help the reader locate (and access) the same resources an author has used by providing the necessary

[30] Rapporto finale di area Gruppo di Esperti della Valutazione dell'Area giuridica (GEV 12) per la Valutazione della Qualità della Ricerca 2011–2014, p. 94–95.

[31] Also Germany and in particular also the German Scientific Council on the legal science reaffirms the importance of the international dimension by stressing the need for internationalisation of legal research, which is listed as a requirement to be implemented in order to allow the maintenance of the qualified role of legal science in global context.

bibliographic information. Citations provide both authorial verification of the original source material at the time they are used and the needed information for readers to later find the cited source.

For legal scholars the function of a citation list is to demonstrate the amount of given readings, rather than the quality of the authors and documents cited. A legal citation therefore, is mostly an expression of diligence in research and serves those who make the citation more than those who are quoted, or receiving the citation (Conte 2015; Santilli 2012). This value, therefore, alters the meaning of "citation unity", where bibliometric indexes are often used profitably for the hard sciences. Citations do not necessarily measure the importance of a publication or document or the degree to which an author's thought is original. They may, however, serve to give further information regarding a certain problem, and direct the reader to further literature. It is common practice amongst legal scholars to cite ideas that are considered wrong or to contradict their own opinion as an author. One also has to consider, for instance, the fact that citations are often made by researchers who are part of a legal "school of interpretation", and are done so in order to oppose an opinion or just to prove what is known.

Peer Review It is widely accepted that peer review is the most valid form of research evaluation in legal science. As highlighted in VQR 2004–2010, at the moment the so-called bibliometric evaluation methods are mostly technically unsuitable and unusable for legal research. In fact, in this evaluation exercise bibliometric evaluation has been used only to check the degree of correlation between the two methods (informed peer review and bibliometric evaluation) and to verify its robustness.

Furthermore, in the VQR 2011–2014 it is confirmed that the evaluation method for legal research is peer review.

This is also visible in the final VQR 2004–2010 report and is reflected in the opinion on the use of bibliometric indicators for the assessment of legal research expressed in 2014 by CASAG.

Both documents reassert their opposition to an evaluation based, either in whole or in part, on bibliometric indicators, stating in the final report that

> Nothing indicates that bibliometric indicators – with their advantages and drawbacks – can be considered a sort of excellence to which, sooner or later, one must evolve. Moreover, nothing in the ratings made by ANVUR shows that these indicators provide better results than peer review.

According to CASAG:

> In fact, there is a high risk that once [bibliometric indicators] are introduced, they will end up becoming the exclusive evaluation method, replacing the more challenging judgement of peers. Even apart from this risk, the added value of incorporating such indicators has to be totally proved and it is out of the question that, in case of discrepancy (that is, if the peer review gives a result in contrast to that resulting from the use of bibliometric indicators) prevalence should be assigned to the former.

In particular, CASAG has expressed its view on the document entitled 'Preliminary specifications for an Italian bibliometric database in the areas of humanities and social sciences' prepared by the Working group "Database and new

indicators" appointed by ANVUR and presented at a meeting of scientific societies held in Rome on 20 January 2014.[32] According to CASAG, the creation of a citation database would involve manipulation of data, such as the specious use of citations to the advantage of one's own group.

Ranking of Publishers, of Journals and of Editorial Series The Italian debate and experience in the field of journal evaluation clearly demonstrates that assessment of the container can never replace the assessment of the content. National scientific societies and institutions, directors and members of editorial boards, individual scholars have all expressed concerns, doubts and criticism about the validity and effectiveness of journal classification and ranking. The judgement on the "quality" of the publishing environment is not directly and fully transferable to a research product. This presumably should also apply to monographs: here critical issues and concerns have been expressed about a possible ranking of publishers and editorial series. With respect to the ranking of legal journals, the VQR 2004–2010 proceeded to this complex and controversial activity: complex, because there is no precedent in the Italian university system, and controversial, because different components of the scientific community (national associations and scientific societies, directors and members of the scientific committee of various journals, individual scholars) expressed doubts and criticisms on a number of aspects. In addition to the classification of journals themselves, some critical aspects concerned the equating of online journals to traditional ones, rating management and its use for purposes other than those of informed peer review.

The realisation of ratings for scientific journals is based on the perspective of informed peer review, in support of those (especially external experts) who should assess the scientific outputs (Della Cananea 2015). It is in this sense, and within these limits, that a classification of legal journals has been implemented and is helpful. The rating has therefore no presumptive value but only an informative one. The experience carried out has confirmed the adequacy of those precautions. According to the GEV 12, in about 8000 valuations made on journal articles, there has not been a single case in which a reviewer has felt obliged to justify its decision to award an excellent rating to an article in a journal not included among those placed in the first level. The classification has therefore provided to evaluators more informative elements without affecting in any way their judgements and evidently without preventing judgements based only on the quality of content.

The landscape of journals is extremely diverse too. It is quite difficult to rank journals in which there are technical insights, despite attention for innovative and jurisprudential development. It is not even easy to rate journals published by one (often more than one) department or university press: they are very heterogeneous and often discontinuous.

It is also necessary to avoid an unintentional but no less negative effect, i.e., to discourage the submission of articles to journals. An effect of this type may depend

[32] http://www.anvur.org/attachments/article/570/Documento%20base_specifiche%20banca%20dati%20riviste%20umanistiche%20e%20sociali.pdf

on equating online journals to traditional ones, which have limits on the space available for the accepted contributions and longer publication times. For this reason, GEV expressed the view that there should be a periodic review of ratings and that it should be done frequently, to make the review mechanism both certain and solid, subject to any adjustments, in order to encourage the progressive improvement of the journals.

Today ANVUR publishes updated lists of scientific journals and Class A scientific journals for the purpose of *Abilitazione Scientifica Nazionale* (ASN) which was introduced as a prerequisite for applying for tenured associate or full professor positions at state-recognised universities. These lists are the result of the first phase of the overall review opened in late 2015, carried out by ANVUR with the support of the working group of journals and scientific books (Working group – WG *Riviste e libri scientifici*), composed of well-known scholars. The lists cover all non-bibliometric areas. At this stage, the WG *Riviste e libri scientifici* has focused primarily on the selection of international qualified journals not previously included in the lists.

With regard to monographs, each publisher may have, and in fact often does have, very different series based on traditions, objectives and parameters. For the moment there are no Italian rankings of publishers and editorial series. To set up such ratings, the following indicators should be taken into account:

– Process indicators such as the prestige of the steering committee, the use of pre-established and transparent procedures for the selection of the contributions, the existence of a committee of qualified referees.
– Output indicators, such as book reviews published in international and national journals and other relevant information required to provide useful information on the quality and impact of research outputs.
– Diffusion indicators, such as the presence in libraries of monographs published in the series.

Impact on Society The issue of measuring the impact of legal research is extremely complex and problematic due in large part to the difficulty of measuring impact only using the tools adopted within the academic community.

The use of research evaluation tools relating to the socio-economic aspects of research, however, is one of the most complex and difficult challenges that academia and the research environment is experimenting in recent years. This type of impact involves awareness of both the individual scientific areas of research and the social and economic context within which the research activity places itself. An example is the impact that a comment to a judgement may have on changes in the jurisprudence.

Citation Databases No national citation databases in the field of legal research are available in Italy for articles published in Italian journals. Law reviews that have received accreditation from international databases such as Web of Science and Scopus are very few. The VQR 2004–2010 states that only 0.03% of the Italian legal scholars' works are included in such databases, selected for the calculation of

bibliometric indices. The test calculation in legal research has revealed that bibliometric evaluation is systematically stricter than peer review.

An additional issue is whether, in the long term, a database of the Italian legal journals should be created and be made available to scholars and institutions in charge of funding legal research. In the US there are proprietary databases (among all Lexis Nexis and Westlaw) providing access to digitised texts of American law journals. Within these databases indices are provided such as the famous Shepard's Citations Service allowing the measurement of citations received by an article (Shapiro and Pearse 2012). As legal scholarship citations in case law are not forbidden, this service allows the measurement of citations received by an article in the judgements.

Moreover, for some years now, tools for measuring quantitative indices have been developed that are based on various parameters (impact factor, citations from journal articles, citations from cases, currency-factor). An example is "Law Journals: Submissions and Ranking 2008–2015", available on the website of the Washington and Lee University School of Law, which is based on data of JLR – Journals and Law Reviews Westlaw.[33]

The Italian debate focuses on possible operative approaches to solve the problem of lack of a bibliographic/citation database for the legal domain. At the moment, a wide network of institutional Open Access repositories are being put into place, but there are no disciplinary archives comparable to the former Social Science Research Network (SSRN). As with most prestigious universities in the world,[34] institutional policies might encourage the population of new institutional repositories by implementing different business models. However, in Italy there are resources like the DoGi database, developed by the Institute of Legal Information Theory and Techniques of the National Research Council (ITTIG-CNR)[35] and ESSPER,[36] which could be redesigned to become a multi-disciplinary Open Access repository, which could also be useful for supporting the evaluation process.

The relevance of these listed issues in the national legal academic debate makes it necessary to move forward extensive consultations among various legal scientific communities and involve all stakeholders in order to open up a wide-ranging, collective reflection on such a delicate and relevant scenario for the enhancement of scientific legal production.

[33] http://lawlib.wlu.edu/LJ/

[34] Harvard Law School open access motion. Cyber.law.harvard.edu/node/4289.

[35] The DoGi database, born in 1969, is, in the Italian national context, one of the most relevant sources for online research of legal literature. The free online database provides abstracts of articles published in more than 250 Italian legal periodicals. www.ittig.cnr.it/dogi/en/Index.php

[36] ESSPER is a bibliographic database freely available providing access to the main Italian journals in economics, law and the social sciences. www.biblio.liuc.it/scripts/essper/default.asp

6 Conclusions: Key Issues of the Ongoing Debate

The Italian and international experiences concerning the quality of legal scholarship is more important today than ever, because law appears to be a "discipline in transition", which can no longer simply rely on its historical authority (van Gestel 2015; van Gestel and Micklitz 2014; van Caenegem 1987). The most central issue is the fact that there is no uniform approach to the evaluation of research outputs in legal science. Although peer review is the gold standard amidst all evaluation methods, the popularity of bibliometrics is progressing: citation counts as well as the impact factor score of journals are now emerging factors in the peer review process. Yet, both peer review and the use of metrics have serious shortcomings. Increased consideration should be given to methodological justification, and theory building in scholarly legal publications (van Gestel and Micklitz 2014), increasing the transparency of editorial boards' quality criteria for the assessment of legal research outputs, and to serious and fruitful discussions on the European benefits and disadvantages of each research-evaluation system. Legal science started late compared to other non-bibliometric disciplines. To a certain extent, this can be considered an advantage: it means that the establishment of a new process can start from scratch, with no battles to change an existing system. Constructive discussions pertaining to evaluation, along with new procedures, should be accelerated, at least in Europe, even if it is an evolving process. Again, what is most important is that these discussions are improved through the transparency of adopted criteria.

One part of the literature (van Gestel and Vranken 2011; van Gestel 2015; Gutwirth 2009; Epstein and King 2002; Siems 2008), has highlighted some key questions at the core of the debate on legal research assessment at a European level. These are reported below in the form of questions based on the current international debate.

- Following the research assessment methodologies of various European countries, similar content-based criteria such as originality, significance, depth, societal impact are adopted. Is there a general consensus on the value and interpretation of such criteria?
- Is it possible to assess legal research on the basis of international bibliometric evaluation techniques more or less widely accepted in other scientific disciplines?
- How reliable is peer review?
- What are the advantages and disadvantages of ranking law journals?
- Is the relation between legal scholars and legal practice important for research assessment?
- Is the harmonisation of legal research assessment methodology at a European level desirable in the years to come?

These specific questions are likely to become the framework for future debate, with a potential to involve other countries that only recently have addressed the question of legal research assessment. Legal scholars as well as policy makers of each country are the main actors of this discussion which has just begun.

Adopting the criteria, evaluation processes and methods that are or have been used in other sciences is not a good solution. However, it would be appropriate to create transnational standards for legal research quality assessment, taking into account the actual internationalisation of research in this area and the increasing mobility of students and development of international law schools. The establishment of harmonised standards or of generally accepted quality indicators is a challenge to be met, despite the differences between national assessment methods, various publishing cultures and many academic traditions.

The Italian context is an interesting example to be investigated. The evaluation exercise is very structured and complex and the debate within the legal community is very alive as well as the active role of scholars in the discussion. Currently, the main topics that are leading the debate are: the definition and value of socio-economic impact, internationalisation, and the qualifications of peers involved in peer review. It is more desirable to move forward with a selection of peers who possess reviewing skills as well as the type of expertise required in relation to outputs under assessment, rather than focusing on an abstract distinction between disciplines and sub-disciplines.

Of course, the reflections and considerations presented in this chapter are situated within a cultural context of constant change, given that there is a continual need to reposition certain topics, as well as a need for more investigations concerning the field of legal assessment. There is an opportunity at least for a European-wide common approach, and we conclude by suggesting four main points for further discussion: (1) quality indicators should not be imposed upon legal scholars from a top down perspective; (2) academic freedom is priceless, hence scholars themselves should be the ones who decide what quality entail; (3) the quality of legal scholarship is not a goal that needs to be reached but a journey that needs to continue; and (4) transparency and accountability is to be valued in the legal evaluation process, thus forming a strong evaluation culture.

Last but not least, we recommend that the scientific community play a central role in the evaluation discussion, and that scientists be recruited to establish new and appropriate methods of measurement, so as not to encourage the intolerance of some scientists towards the evaluation process.

References

Accademia Nazionale dei Lincei. (2012). Criteri di valutazione della ricerca scientifica nelle scienze morali con riferimento all'attività dell'ANVUR. www.lincei.it/files/dichiarazioni/Mozione_Anvur_20-4-2012.pdf.

AUBR. (2010). *Assessing Europe's university-based research*. European Commission – Expert Group on the Assessment of University-Based Research – DG Research. http://ec.europa.eu/research/era/docs/en/areas-of-actions-universities-assessing-europe-university-based-research-2010-en.pdf

Cassese, S. (2013). L'ANVUR ha ucciso la valutazione. Viva la valutazione! *Il Mulino*, no.1.

Conte, G. (2015). Le reciproche interferenze tra evoluzione della ricerca giuridica e valutazione della medesima. In G. Conte (Ed.), *Evoluzione e valutazione della ricerca giuridica* (pp. 9–16). Naples: ESI.

della Cananea, G. (2014). L'ANVUR e le banche dati dell'area giuridica. *Diritti lavori mercati, 11*(1), 211–220.

della Cananea, G. (2015). La valutazione della ricerca giuridica dopo il Duemila: verso una nuova koinè culturale? In G. Conte (Ed.), *Evoluzione e valutazione della ricerca giuridica* (pp. 17–39). Naples: ESI.

Di Cristina, F. (2013). Risultati e prospettive della valutazione della ricerca nell'area giuridica. *Munus, 2*, 395–437.

Di Cristina, F. (2014). L'attività di valutazione e la qualità della ricerca giuridica nel Rapporto finale del GEV12. *Rivista trimestrale di diritto pubblico, 64*(1), 267–273.

Di Cristina, F. (2015). Un workshop internazionale su La valutazione della ricerca nelle Humanities and social sciences. *Rivista trimestrale di diritto pubblico, 1*, 269–270.

Di Raimo, R. (2015). la valutazione delle monografie nelle scienze sociali, secondo il sistema attuale. In G. Conte (Ed.), *Evoluzione e valutazione della ricerca giuridica* (pp. 241–252). Naples: ESI.

Douglas, G. (2015). The search for excellence: Assessing legal research in the United Kingdom. In G. Conte (Ed.), *Evoluzione e valutazione della ricerca giuridica* (pp. 147–166). Naples: ESI.

Epstein, L., & King, G. (2002). The rules of inference. *Chicago Law Review, 69*, 1–209.

Flückiger, A., & Tanquerel, T. (2015). *L'évaluation de la recherche en droit/Assessing research in law Enjeux et méthodes/Stakes and methods*. Bruxelles: Bruylant.

Garfield, E. (1979). *Citation indexing – Its theory and application in science, technology, and humanities*. New York: Wiley.

Gasparini Casari, V. (2011). La valutazione delle riviste giuridiche. *Il Diritto dell'economia, 24*(3–4), 673–693.

German Council of Science and Humanities. (2012). *Perspectives of legal scholarship in Germany. Current situation, analyses, recommendations*. Berlin.

Glendinning, C. (2010). *Piloting the assessment of research impact in social policy and social work*. Report of a seminar at the Leeds University on 26 Mar 2010. http://www.social-policy.org.uk/wordpress/wp-content/uploads/2014/05/REF-impact-pilot.pdf

Gutwirth, S.. (2009). The evaluation of legal science. The Vl.I.R.-model for integral quality assessment of research in law: What next? Brussels. *It takes two to do science. The puzzling interactions between science and society*. Available at: http://works.bepress.com/serge_gutwirth/16/

Holbrook, J. B., & Frodeman, R. (2011). Peer review and the ex ante assessment of societal impacts. *Research Evaluation, 20*(3), 239–246.

Jerker, D., & Svantesson, B. (2009). International ranking of law journals: Can it be done and at what costs? *Legal Studies, 29*.

Labatt, H. J. (1860). *A Table of Cases in California: As Affirmed, Overruled, Modified, Commented Upon, or Altered by Statutory Enactment*. San Francisco: H.H. Bancroft and.

Leiden Manifesto. (2015). http://www.nature.com/polopoly_fs/1.17351!/menu/main/topColumns/topLeftColumn/pdf/520429a.pdf

Marcello, D. (2014). Evoluzione e valutazione della ricerca giuridica. Cronaca di un convegno. *Rassegna di diritto civile, 35*(4), 1376–1382.

Martin, B., & Irvine, J. (1983). Assessing basic research. Some partial indicators of scientific progress in radio astronomy. *Research Policy, 12*(2), 61–90.

Martin, B. R., Nightingale, P., & Yegros-Yegros, A. (2012). Science and technology studies: Exploring the knowledge base. *Research Policy, 41*, 1182–1204.

Militello, V. (2014). L'identità della scienza giuridica penale nell'ordinamento multilivello. *Rivista Italiana di Diritto e Procedura Penale, 57*(1), 106–132.

Moed, H. F., Burger, W. J. M., Frankfort, J. G., & Van Raan, A. F. J. (1985a). The application of bibliometric indicators: important field- and time-dependent factors to be considered. *Scientometrics, 8*(3), 177–203.

Moed, H. F., Burger, W. J. M., Frankfort, J. G., & Van Raan, A. F. J. (1985b). The use of bibliometric data for the measurement of university research performance. *Research Policy, 14*, 131–149.

Nederhof, A. J. (2006). Bibliometric monitoring of research performance in the social sciences and the humanities: A review. *Scientometrics, 66*(1), 81–100.

Nourse, V., & Shaffer, G. (2009). Varieties of new legal realism: Can a new world order prompt a new legal theory. *Cornell Law Review, 95*, 61–138.

Piazza, S., & Matteucci, M. C. (2015). La valutazione d'impatto socio-economico e culturale della ricerca nelle scienze umane e sociali: uno studio di caso in una Università del Regno Unito. *Journal on Research Policy & Evaluation, 3*(1).

Posner, R. A. (2002). Legal scholarship today. *Harvard Law Review, 115*, 1325.

Rubin, E. L. (1992). On beyond truth: A Theory for evaluating legal scholarship, 80 *California Law Review*, 80(4), 889–963 (Reprinted in *Readings in Race and Law: A Guide to Critical Race Theory*, Alex Johnson, ed., West, 2002).

Santilli, G. (2012). Bibliometria e avvocati: la valutazione della ricerca in campo giuridico, 2012–02-06 http://www.roars.it/online/bibliometria-e-avvocati-la-valutazione-della-ricerca-in-campo-giuridico/

Schloegl, C., & Gorraiz, J. (2011). Global usage versus global citation metrics: The case of pharmacology journals. *Journal of the American Society for Information Science and Technology, 62*(1), 161–170.

Seglen, P. O., & Aksnes, D. W. (2000). Scientific productivity and group size: A bibliometric analysis of Norwegian microbiological research. *Scientometrics, 49*(1), 125–143.

Shapiro, F. R. (1992). Origins of bibliometrics, citation indexing, and citation analysis: The neglected legal literature. *Journal of the American Society for Information Science, 43*, 337–339.

Shapiro, F. R., & Pearse, M. (2012). The most-cited law review articles of all time. *Michigan Law Review, 110*(8), 1483–1520.

Siems, M. M. (2008). Legal originality. *Oxford Journal of Legal Studies, 28*, 174.

Siems, M. M., & Mac Sithigh, D. (2012). Mapping legal research (1 November 2012). *Cambridge Law Journal*, *71*(3), 651–676. Available at SSRN: http://ssrn.com/abstract=2181374

Snow C. P. (1959). *The rede lecture*. Cambridge University Press. http://s-f-walker.org.uk/pubsebooks/2cultures/Rede-lecture-2-cultures.pdf.

Snow, C. P. (1964). *Le due culture*. Feltrinelli: Milan.

Snowden et al. (2015). *The Metric tide: An agenda for responsible indicators in research.* https://www.theguardian.com/science/political-science/2015/jul/09/the-metric-tide-responsible-indicators-research

Standard Evaluation Protocol 2015–2021. (2015). Association of Universities in the Netherlands (VSNU), Netherlands Organisation for Scientific Research (NWO), Royal Netherlands Academy of Arts and Sciences (KNAW), éd. 2014. *Standard evaluation protocol 2015–2021: Protocol for research assessments in the Netherlands*. Available at https://www.knaw.nl/nl/actueel/publicaties/standard-evaluation-protocol-2015-2021.

Van Boom, W. H., & Van Gestel, R. (2015). *Rechtwetenschappelijk Onderzoek – Uitkomsten Van Een Landelijke Enquête* (Legal Research – Results of a National Survey), May, 2015. Nederlands Juristenblad 2015/20. http://ssrn.com/abstract=2609277

van Caenegem, R. (1987). *Judges, legislators and professors, chapters in European legal history*. Cambridge: Cambridge University Press.

Van Gestel, R. (2015). Sense and non-sense of a European ranking of law schools and law journals. *Legal Studies, 35*, 165–185. https://doi.org/10.1111/lest.12050.

van Gestel, R., & Micklitz, H.-W. (2014). Why methods matter in European legal scholarship. *European Law Journal, 20*, 292–316. https://doi.org/10.1111/eulj.12049.

van Gestel, R., & Vranken, J. (2011). Assessing legal research: Sense and nonsense of peer review versus bibliometrics and the need for a European approach. *German Law Journal, 12*(3), 901–929.

Van Gestel, R., Micklitz, H.-W., & Poiares M. M. (2012). *Methodology in the new legal world* (May 1, 2012). EUI Working Papers LAW No. 2012/13. Available at SSRN: http://ssrn.com/abstract=2069872 or https://doi.org/10.2139/ssrn.2069872

van Hoecke, M. (2011). Legal doctrine. Which methods for what kind of disciplines. In M. van Hoecke (Ed.), *Methodologies of legal research: Which kind of method for what Kind of discipline?* (p. 31). Oxford: Hart Publishing.

Wait, W. (1872). *A table of cases affirmed, reversed or cited in any of the volumes of the reports of the State of New York*. William Gould & Son: Albany.

Wooding S. & Grant J. (2003). *Assessing research: The researchers' view.* Joint Funding Bodies' Review of Research Assessment, U.K. Available at http://www.rareview.ac.uk/reports/assess/AssessResearchReport.pdf.

Part II
The Role of Books and Monographs in SSH Research and Their Evaluation

More, Less or Better: The Problem of Evaluating Books in SSH Research

Geoffrey Williams, Antonella Basso, Ioana Galleron, and Tiziana Lippiello

The truism says that humanities scholars publish books. Like all truisms it is partially true, but not entirely. Extended to the entirety of the SSH, it is definitely not the case as the wide variety of disciplines also means a wide variety of dissemination patterns. What is true is that in many disciplines books are a recognised and valued output, but it is far from easy to take them into account in a clear way.

In order to improve the evaluation of books in the Italian system, the RobinBa project took a close look at the place of books and book chapters in research in the social sciences and humanities in Italy. Rather than just rely on any one approach to books, this project combined both quantitative and qualitative methodologies using databases as well as questionnaires and focus groups to find out not only where humanities scholars are publishing, but why they are doing so, and to what extent these habits and reasons are taken into account in the evaluation exercises. In addition, it also looked closely at how books are dealt with in other European countries.

This chapter is based on the project RobinBa (The Role of Books In Non-Bibliometric Areas), supported by Anvur with Call for proposal no. 1/2014.

G. Williams (✉)
Laboratoire Litt&Arts, Université Grenoble-Alpes, Grenoble, France
e-mail: williams@evalhum.eu

A. Basso
Dipartimento di Economia, Università Ca' Foscari, Venice, Italy
e-mail: basso@unive.it

I. Galleron
UFR Littérature, linguistique, didactique, Université de Paris III, Paris, France
e-mail: ioana.galleron@univ-paris3.fr

T. Lippiello
Dipartimento di Studi sull'Asia e sull'Africa Mediterranea, Università Ca' Foscari Venezia, Venice, Italy
e-mail: lippiello@unive.it

© Springer International Publishing AG 2018

A. Bonaccorsi (ed.), *The Evaluation of Research in Social Sciences and Humanities*, https://doi.org/10.1007/978-3-319-68554-0_6

After a presentation of three main approaches to book evaluation, this chapter will focus on understanding the reasons of SSH scholars for publishing books. In order to do so, it will first analyse patterns for scientific publication in two Italian universities, then move to the analysis of answers obtained from SSH scholars through focus groups and a survey about their publication practices and motivations. It will conclude with some recommendations about how to improve book evaluation.

1 Books: The Nature of the Beast

The first problem is defining what exactly is a book. According to Wikipedia, a book is:

> a set of written, printed, illustrated, or blank sheets, made of ink, paper, parchment, or other materials, fastened together to hinge at one side. A single sheet within a book is a leaf, and each side of a leaf is a page. A set of text-filled or illustrated pages produced in electronic format is known as an electronic book, or e-book.

This is fine as it stands, but it does not say what differentiates a book from a periodical, or a long report, for instance, neither of which qualifies for an ISBN. This definition also treats e-books as simply a fully structured, but unprinted, paper edition, when e-books should, and certainly will, in the near future, be much more than this.

Another simple definition is that a book has an ISBN, and is thus recorded as such in a library database system. Unfortunately, this simple definition is remarkably lacking in clarity. Rather than saying what a book is the ISBN Manual ("ISBN Users' Manual: Interational Edition, Sixth Edition" 2012) is more concerned with what it is not. Thus, amongst those that are not, we find:

- Continuing resources treated in their entirety as bibliographic entities (individual issues may qualify for ISBNs);
- Abstract entities such as textual works and other abstract creations of intellectual or artistic content;
- Ephemeral printed materials such as advertising matter and the like;
- Printed music;
- Art prints and art folders without title page and text;
- Personal documents (such as an electronic curriculum vitae or personal profile);
- […]

This then differentiates books, report and periodicals, but it does not tackle the question of size, leaving some to adopt a fairly brutal quantitative definition. For example, in Lithuanian research evaluation a book is minimally 40,000 characters multiplied by a field based coefficient, which is eight for the SSH, thus corresponding to a book with eight chapters. This is an essentially bureaucratic decision on what is and what is not a book. Furthermore, such definitions only apply to countries adhering to the ISBN, while not all do. As African specialists pointed out during a focus group, they may have research in countries that do not have ISBN, and research is after all international.

Thus, simple book definitions raise several questions, some of which are circular: how many pages for the work to be declared a book rather than an article? Do you give equal value to all books? If you apply a typology, what is the basis for the selection? How do you differentiate academic from non-academic, if at all? And then, there is the question of e-books. How do you handle the e-documents when in theory an electronic document does not have pages, these being purely a convention needed for printed editions?

Even if broad categories for defining books are accepted, other difficulties remain. In research, references are made to the monograph, generally perceived as the work of a single author. However, in some systems multiple authors are accepted, and anyway why should a co-authored work be less considered? Then, there is the question of collected works. Is there a difference between a chapter and an article? Is full recognition given to the editor of a collected work? Is it really possible to differentiate proceedings from collected chapters? Is it even necessary, or useful to differentiate them?

So-called textbooks are another problem, as these may be instruction manuals aimed at secondary schools, which can be simply a compilation of notes and secondary sources, or genuine reference books as the Cambridge Textbooks in Linguistics series, to cite only those, which are major works of scholarship. Similarly, a book review can be a detailed analysis performed only by a true field specialist, or just an exercise of praise performed by a complacent colleague.

None of these questions have easy answers or any answer at all. The reality is that much depends on the disciplines involved and on the evaluation policies of different institutions countries. A bibliometric based system, even when it calls partially on peer review, obviously applies a very different definition to a qualitative system as the REF. When evaluators are aware of the social impact of research, as in the Dutch SEP, then they are going to value a much wider range of outputs, and take into account field and discipline specificities. In a one-size-fits-all process as those found in many ex-communist countries, a rigid system is applied, often with disastrous results.

In 2008, Humanities in the European Research Area (HERA) published a survey of evaluation systems in Europe (Dolan 2007). This obviously covered, amongst other issues, the place of books and book chapters. The predominance of books in humanities research is stressed, as well as that of its complications, notably the coverage by commercial databases and the question of citation indexes. Interestingly, the report noted that humanities books seek:

> *Wider audience:* To a much greater degree than is the case in science disciplines, humanities authors will also target the general public through the medium of books and book-chapters. There is also a strong tradition of writing for the non-scholarly press. (Dolan 2007, 27)

This clearly points towards what the Dutch call ‘hybrid publications’, written as much for a scholarly audience as the general public. This is very much the case for historians, but also in art history and architecture, where what might be seen as a coffee table book because of its wealth of illustrations is also highly erudite.

The HERA report is obviously now out of date and until the results of the pan-European survey being carried out in the ENRESSH Cost action (CA15137) (Galleron 2016) are available, the picture will remain incomplete. Nevertheless,

much work has already been published, and within RobinBa, we were able to carry out a small survey of evaluation systems and how they deal with books.

Consequently, before looking at the attitudes and perceptions of Italian scholars, and describing the methodologies we applied, we shall briefly overview the situation in a number of other European countries.

2 Book Evaluation Across Europe

Different evaluation systems apply very different definitions of what actually constitutes a book and what book types are acceptable in an evaluation exercise. It thus seems useful to start by seeing some of the possibilities. This is not an exhaustive study, but does seek to see how different systems perform. For the purpose of this analysis, we divide evaluation systems into three main groups: qualitative systems as in the UK and the Netherlands; quantitative systems as in the Czech Republic and Poland; and the database systems as in Norway and Flanders. France has a qualitative system for the evaluation of research units rather than the research itself (Williams and Galleron 2016) and the French speaking part of Belgium also has a looser system (Zaccai et al. 2016). This means that in France books are not given any particular status in the global system, and whilst they are taken into consideration in individual assessments, there are no clear criteria for either their inclusion or assessment. This is also the case in many other countries. Consequently we shall concentrate on systems that have explicit criteria and collect quantitative data about research outputs, even if their methodology is qualitative.

2.1 Qualitative Evaluation Systems

In theory, a qualitative system does not need rigid definitions of outputs as the material to be submitted is not limited and all outputs are to be judged on their own merits. Such systems can be expensive to put into place, but when large amount of research funding is at play, the return on investment can make the cost worthwhile. Only two European countries have such systems: the UK and the Netherlands.

Large-scale research evaluation procedures have been in place in the UK for many years now, having started with the Research Assessment Exercise (RAE) and continued by the current Research Excellence Framework (REF), upon which UK Universities rely for a major part of their research funding. REF covers disciplines by allocating them to one of four panels. Social Sciences are to be found in C and the Humanities in D, with each panel being sub-divided into sub-panels corresponding to broad disciplines. Panels and subpanels can take into consideration what is specifically valued in those disciplines, and the SSH panels use exclusively qualitative analyses. It is the researcher who decides what he or she wants to put forward and as evaluation is qualitative a tight definition of a monograph is not needed.

The importance of books and the difficulties involved in their evaluation has been treated in important reports (Vincent et Wickham 2013; Wilsden 2015) and this will obviously have influenced the debate whilst precluding the need for anything but a broad definition. The REF guidelines recognise the importance of books in panels C & D declaring that:

> Research evaluation in book-oriented fields is more challenging than for article-based subject areas because counts of citations from articles, which dominate traditional citation indexes, seem insufficient to assess the impact of books. The Book Citation Index within Web of Science is a recent response to this issue since journal citations on their own might miss about half of the citations to books. However, some academic books are primarily written for teaching (e.g. textbooks) or cultural purposes (e.g. novels and poetry) and citation counts of any kind may be wholly inappropriate for these. (REF 2012)

It goes on to state that books were most frequently submitted in panels C and D. The final breakdown of submitted data shows the importance of books, but also demonstrates that in all disciplines journal articles dominate, as can be seen in Fig. 1.

The dominance of journal articles is most apparent in the business and management disciplines, while in the core humanities disciplines the ratio of books to articles shows a more marked preference for authored books. However, it must be borne in mind that this data does not reflect the entire researcher output, but what researchers or their institutions deem their most significant contribution to their discipline.

The situation is different in the Netherlands, where a new evaluation protocol was introduced in 2014 to cover the period 2015–2021 (Association of Universities in the Netherlands (VSNU) et al. 2014). Unlike the UK, the Netherlands do not have a full performance-based funding system, and so the implications are somewhat dif-

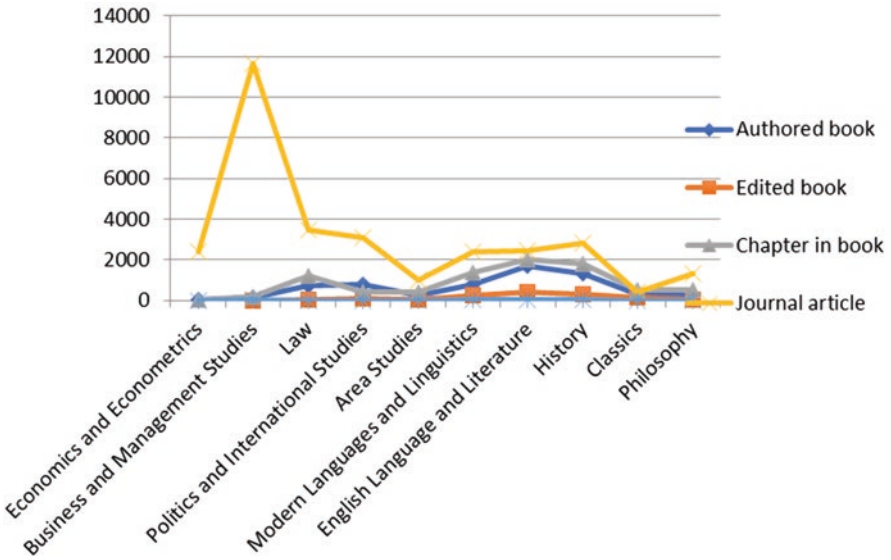


Fig. 1 Breakdown of publication types for selected disciplines in REF 2014

ferent. However, as (Wouters 2014) has pointed out, the repercussions of these national evaluations can be important within the institutions and can cause considerable loss of prestige to institutions, research groups and lead researchers.

Like in the REF, the evaluation is not bibliometric and social impact is seen as of great importance. There is no standard set of indicators but a variety of possible indicators is proposed with, for example, scholarly books being retained as a measure of research quality whilst outreach activities are also given value. As such, monographs require only a broad definition and there is no typology of outputs. As emphasised in a 2011 report on *Quality Indicators for Research in the Humanities,* books and book chapters maintain their importance, whatever the language they are published in:

> Fair consideration needs to be given to monographs and international publications in languages other than English, meaning, for example, that primacy cannot be accorded to bibliometric indicators that are as yet still based on databases consisting primarily of English-language journal publications. (KNAW, Royal Dutch Academy 2011)

Thus, these qualitative approaches differentiate between broad disciplinary area, with the Social Sciences and Humanities benefiting from an exclusively peer reviewed evaluation. In theory, all research outputs can be considered, although in the REF some filtering is carried out by the submitting universities, with some discarding proceedings entirely, even if this can be highly detrimental to some disciplines such as lexicography which maintains a more engineering approach to output and therefore value proceedings.

2.2 Quantitative Evaluation Systems

In many ex-communist countries, a fairly bureaucratic one-size-fits-all quantitative system is applied, with some rather arcane points applied to different outputs. The definition of a monograph applied in Lithuania has already been mentioned; here we shall concentrate on Poland and the Czech Republic.

In Poland, the purely quantitative system, based on a complex point system, clearly privileges journal articles over books. As one informant pointed out, "in Poland, there is a very strong (implicit) assumption that in contemporary science, journals are the main channel of scholarly communication for each discipline. In effect, even the most prestigious monograph (25 points) has to be assessed at a lower level than the most prestigious article (50 points)".

In evaluating books, a strict typology of outputs is applied:

Monograph, excluding handbooks and textbooks, including lexicons, critical editions, encyclopaedias, and conference proceedings. A monograph has to fulfil the following conditions: (1) it has to be reviewed and it has to present an original research problem, (2) it has to contain a bibliography, and (3) its length should be at least sixty author sheets (240,000 characters or approximately 36,000 words). There is no limit to the number of authors.

Edited Volume it is a monograph with editor(s). There is no limit to the number of editors. (informant)

Textbooks are excluded, but it is admitted that Polish researchers often attempt to pass these off by excluding the words "textbook" and "manual" from the title. Proceedings are only accepted if referenced in WoS or Scopus.

The Czech Republic had a similarly complex system, but this has been reviewed and a new system is being put into place in which there is a move from a quantitative to a mixed system with expert panels.[1]

In the current protocol, a strict typology is applied, with papers in proceedings only being accepted if the volume is referenced in the WoS or Scopus databases. Given the complexity of the procedures, this excludes even the biggest of congresses. As in Poland, a complex point system is applied, which tends to give less value to edited volumes whilst emphasising the importance of single or multi-author monographs. Thus, according to the protocol:

> A specialist book presents the original results of a research that has been conducted by the author of the book or a team of which the author was a member. A book is a non-periodical specialist publication of at least 50 printed pages of text, without photographs, pictures, maps and similar annexes, published in a printed or electronic form and reviewed by at least one generally regarded expert from the appropriate field (however not from the author's workplace) in the form of a reader's review. This regards a precisely defined issue of a certain field, contains the formulation of an identifiable and scientifically recognised methodology (explicitly formulated methodological foundations also in monographs leading to applications and/or formulations of a new methodology based on the current theoretical research in the given field). The formal attributes of a specialist book are references to literature in text, bibliography, summary in at least one world language, possibly notes and source bibliography. ("Methodology of Evaluation of Research Organizations and Evaluation of Finished Programmes (valid for years 2013–2015)" 2013).

2.3 Data-Base Systems

An increasing number of countries are using data based protocols that allow them to follow in detail research output. These have been well documented in the literature and so only a very brief overview will be made here.

Norway has a performance-based funding system that makes use of the national database, CRISTIN – Current Research Information System in Norway (see Sivertsen 2015). This was the precursor of much further development, with the approach now having been adopted by other countries, the most recent being Portugal. In this system, only peer-reviewed products are accepted, including journals and series with ISSN; books and book chapters with ISBN are defined as scholarly output.

> The indicator is based on a division of publication channels (journals, series, book publishers) in two levels: level 1 and level 2. Level 2 contains the most selective international

[1] See the following link for a discussion in English: http://vyzkum.cz/FrontClanek.aspx?idsekce=695512

> journals, series and book publishers and may not contain more than 20 per cent of the publications worldwide in each field of research. Articles in journals and series are given 1 point on level 1 and 3 points on level 2. Articles in books (with ISBN only) are given 0.7 points on level 1 and 1 point on level 2. Monographs are given 5 points in level 1 and 8 points on level 2. The points are fractionalized in the level of institutions according to the institution's share of contributing authors. (Sivertsen 2015, 473)

Outputs are formally defined, but restricted by their bibliographic status and the imposed notion of academic production. The system clearly privileges internationalisation, meaning publications in English. As concerns monographs, the inevitable question is how publisher prestige is judged. A recent comparative analysis of SSH evaluation of books in five countries (Denmark, Flanders, Finland, Norway and Spain) tackled this issue (Giménez-Toledo et al. 2016), and will be looked below with reference to Spain.

Belgium has a dual system with Flanders adopting a performance-based system for allocating central funds. A production database is maintained by Vlaams Academisch Bestand voor de Sociale en Humane Wetenschappen', (VABB- SHW), the Flemish Academic Bibliographic Database for the Social Sciences and Humanities. Research funding through this system represents a considerable amount of income for Flemish universities (details in Engels et al. 2012). VABB-SHW breaks down book outputs into three groups with monographs having the biggest weighting (4), whilst edited books and books chapters receiving 1, as do journal articles. Proceedings papers are recognised, but only receive half a point. Thus books are clearly seen as the highest valued output, but even if they are aggregated by publisher and series, no prestige weighting is applied.

Spain has evaluations carried out by two agencies, ANECA and CNEAI. Books, and book chapters, are evaluated through a variety of criteria amongst which a weighting based on the prestige of the publisher using the database established by the Research Group on Scholarly Books (ÍLIA) at CSIC. In this database, publishing houses are grouped in different categories, according to their perceived editorial specialisation; therefore, some have been ranked by scholars only in literature, or philosophy, etc., while others appear in several rankings and integrate also a "general" category (known by a majority of scholars having completed the survey).

The criteria adopted for the publisher prestige lists have been specified (Giménez-Toledo et al. 2015a, b) and are based on extensive questionnaires involving SSH scholars as well as analyses of the actual production process leading to four main criteria for presence on the list:

1. The prestige perceived by the academic community towards individual book publishers.
2. Book publishers' thematic specialization.
3. Original manuscript selection process.
4. Presence of book publishers in other information systems.

The development and use of the database has been the subject of a number of publications and a full description of the database (Giménez-Toledo et al. 2015a, b) and access to the results is available on line.[2]

[2]The SPI database is accessible at: http://ilia.cchs.csic.es/SPI/index.html

3 Books in Italian SSH Research and Evaluation

Having summed up the situation in a number of European countries, it is clear that no definition or protocol of evaluation can be applied everywhere. There is a wide variety of evaluation systems (Ochsner et al. 2016) and therefore a wide variety of outlooks. It is clear that books are important to many humanities researchers in particular, but there is also a move towards article publishing that is either simply recognised as being the case, as in the UK REF, or being imposed by a higher weighting as in Poland. What is often missing is an appreciation of what the SSH scholars themselves value and how they see the evaluation system. This has been attempted in France (Williams and Galleron 2016) and in Switzerland (Ochsner et al. 2013), and changing patterns have been monitored elsewhere (Engels et al. 2012). The RobinBa project has attempted to get an overview of scholarly perceptions in the humanities in Italy, whilst directly engaging the scholars themselves.

3.1 The RobinBa Project

The Italian evaluation system has been detailed elsewhere in this volume so we shall concentrate on the RobinBa project and its outcomes.

RobinBa, The **R**ole **o**f **B**ooks **i**n **N**on-**B**ibliometric **A**reas, was ANVUR funded research conducted in 2015–2016. It was led by a small team from the University Ca 'Foscari along with two researchers from the French LiCoRN[3] research group and a team at the University of Macerata. The work detailed here is essentially that of Ca′ Foscari and LiCoRN.

The project was an investigation into book publication practices and how Italian humanities researchers see the place of books in their research, aimed at providing recommendations about how to improve book evaluation. It was essentially a pilot study concentrating on a small number of disciplines and the output of the two Italian universities involved. It was a mixed approach study, making use of data submitted for evaluation along with questionnaires and focus groups, the output of which was analysed both statistically and using Computer Assisted Qualitative Data Analysis Software, in this case Atlas-ti.[4]

3.2 Quantitative Analysis

Quantitative analysis was based on the material submitted for the VQR from Ca′ Foscari (CFU) and Macerata (UMC). The two files from CFU and UMC were merged, and cleaned of inconsistencies, such as publications declared multiple

[3] www.licorn-research.fr

[4] www.atlasti.com

times (once for every author pertaining to the same university, when they were co-authored by scholars from the same institution). During the lifetime of the project, two more sources of information were used:

– a file with complete information about all Italian SSH output between 2010 and 2012 (All_Italy2010–12);
– a file with outputs submitted to evaluation in REF2.

Here, we use only the Italian data. Files used contain information about the publications of 319 scholars in area 10 (there are 4978 scholars in "All_Italy2010–12" file), and about the publications of 149 scholars in area 11 (4395 for "All_Italy2010–12"). These two areas are those on which the project decided to focus as most representative of the social sciences and the humanities, while the law disciplines covered by area 12 are, in certain academic systems, considered as constituting a separate field, distinct from the SSH. The latter were also treated in depth in another ANVUR funded project. In the meantime, the two areas are representative of the "non-bibliometric areas", even if bibliometric measures may also be debatable in other areas of research.

It is clear that the number of scholars was not stable throughout the studied period; some scholars retired, moved to other universities, or have been recruited during the period, but we lack information about career changes; on the other hand, it is to be noted that "All_Italy2010–12" data covers a period of 3 years, while CUF and MCU data cover a period of 10 years. Therefore, it cannot really be said that the studied sample represents 6,4% out of the total number of scholars in area 10, or 3,3% in area 11. Still, the figures handled are significant and allow the formulation of conclusions about publication habits and trends. This affirmation is further supported by the observation that conclusions drafted in the first phase of the project on the basis of data from Ca′ Foscari (available since October 2015) have not been contradicted by the addition of data from Macerata in February 2016.

It is also to be noted that some disciplines from areas 10 and 11 are not represented in both Ca′ Foscari and Macerata. When looking at the number of scholars involved in these disciplines for the entire of Italy, it can be observed that they represent only 1.5% in area 10; therefore, these "missing" scholars can be deemed as not biasing in a significant way the observations. The percentage is much higher in area 11 (14,51%), but it concerns two specific fields, which are Pedagogy and Psychology. In both cases, observation of publication habits in these disciplines, on the basis of data for the entire of Italy, indicates a good degree similarity with kin disciplines from the same area, included in ROBINBA study, even if one can note some significant differences with regards to papers in journals and chapters in books (see Figs. 2 and 3).

As can be seen in these two tables, the typology used in the project is much more limited than that applied by the VQR. In fact, considering the scope of our research, focusing on certain outputs and merging the others in broad categories appeared to be legitimate. For the final aggregation of data see Annex 1.

When analysing book publication in comparison with other dissemination practices, the data revealed a quite surprising situation. Books are indeed important in areas 10 and 11, especially when compared with the structure of the output in scientific areas

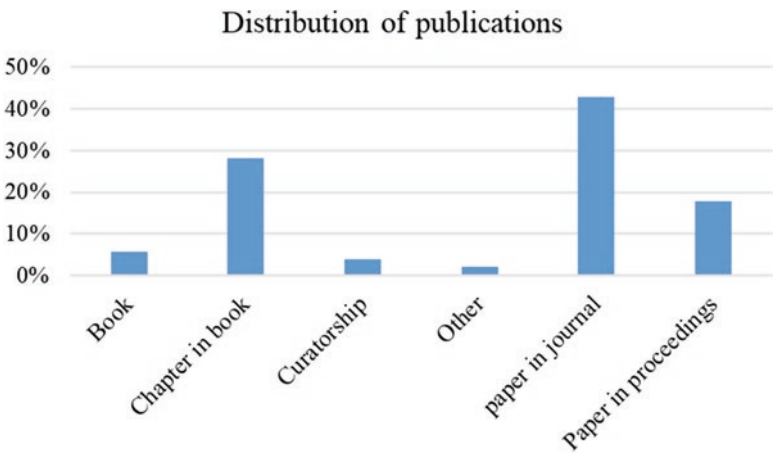


Fig. 2 Publications in "missing" disciplines M-EDF01, 02 and M-PSI02, 03, 06, 07 (based on "All_Italy2010–12")

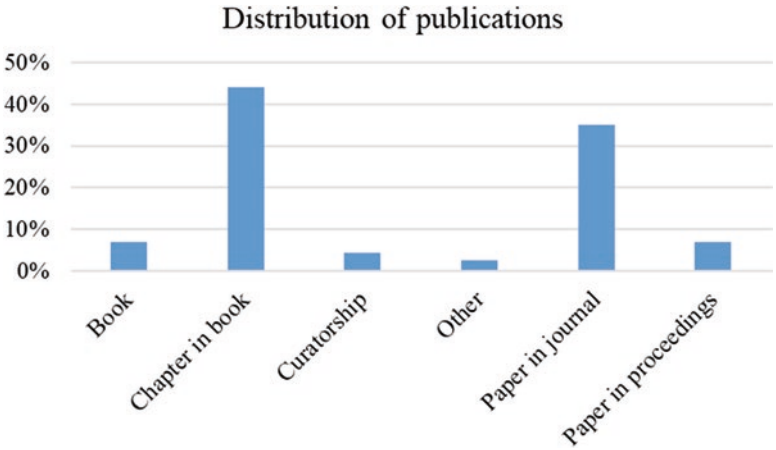


Fig. 3 Publications in disciplines M-PED01, 02, 03, 04 and M-PSI01, 04, 05, 08 (represented in CFU and MCU)

1–9. This is all the more the case if "chapters in books", pointing towards collective books, are added (see Annex 2). Of course, both universities being SSH oriented, they cannot be considered representative of publication habits in areas from 1 to 9, but the difference in publication habits is well documented in literature (Mutz et al. 2013) and confirmed by the repartition of publications that can be observed on the basis of "All_Italy2010–12" file (Fig. 4).

However, contrary to what was expected, books, whether individual (monographs) or collective, do not represent the main publication channel in areas 10 and 11 in these two universities. This is better shown by Figs. 5 and 6.

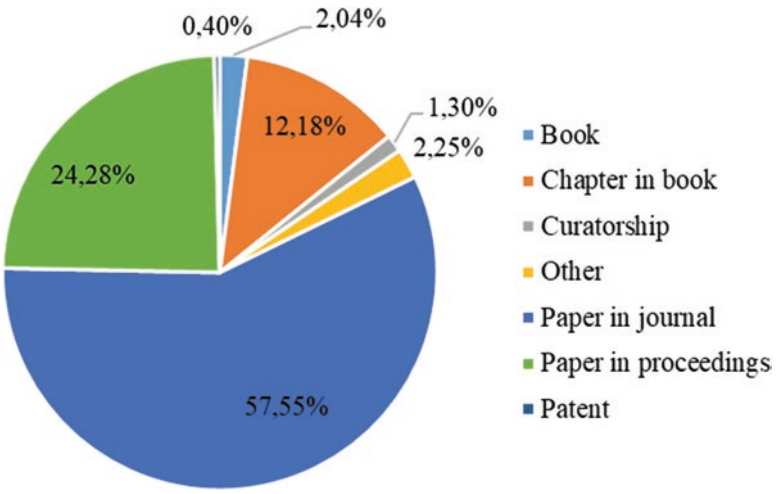


Fig. 4 Output for areas 1 to 9 and 12 to 14 (publications in MCU and CFU, 2010–2012)

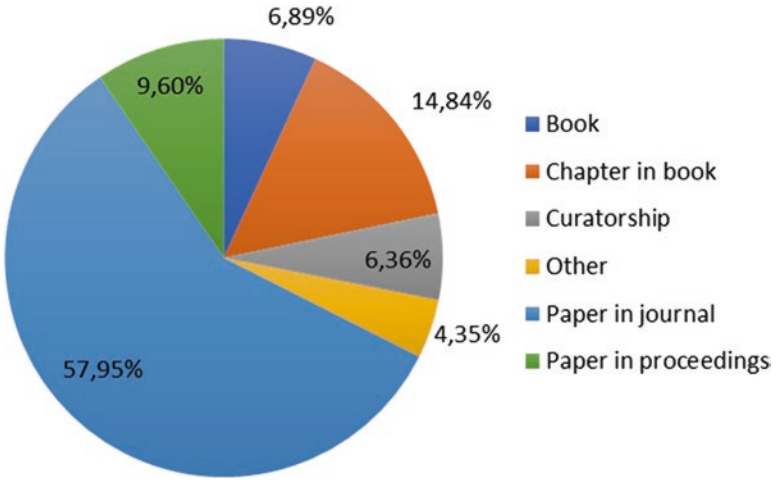


Fig. 5 Distribution of publications (2004–2014) in area 10 in CFU and MCU

Papers in journals account for almost 60% in area 10, and for almost 50% in area 11. In both areas, when considered as monographs or similar, books represent quite a small and stable percentage of some 7%. An important contrast can be noticed between chapters in books, only 14,84% in area 10, while they cover 31,55% in area 11 (more than the double). The role of books remains important in non-bibliometric areas, and especially the role of collective books, but these figures indicate that publishing in journals is gaining traction.

Figures show that publication habits vary considerably from one discipline to another, within the same area, and even between kin sub-disciplines. Book share

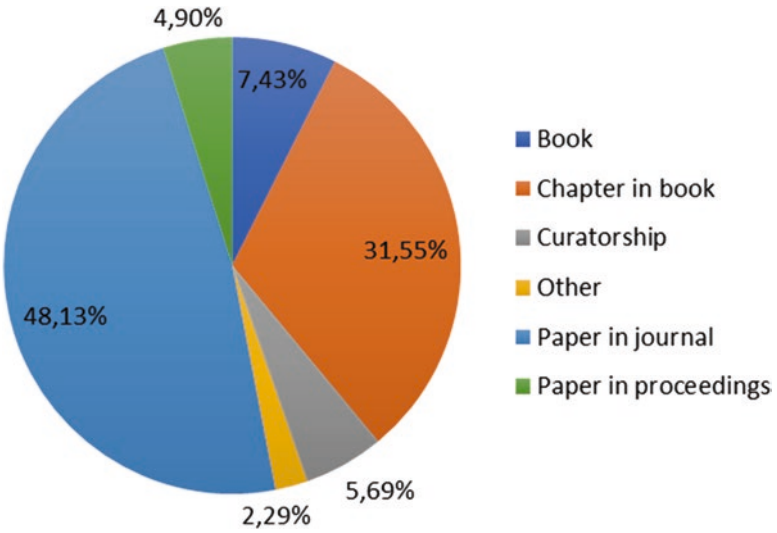


Fig. 6 Distribution of publications (2004–2014) in area 11 in CFU and MCU

goes from 0% to 15,38% depending on the sub-discipline, while book chapters can account for up to 85,14% in some fields. At the same time, papers in journals can represent as little as 5,41% of the total output in some disciplines, while the same category covers 80% of the total output in another discipline. This clearly brings into question any lumping of disciplines that fails to account for such variation, and points towards the need of a more ample and qualitative approach. This investigation was largely beyond the scope of ROBINBA, even if focus groups organised in Macerata and Ca′ Foscari allowed to probe the ground.

In both areas, papers in proceedings observe a slow, but undeniable decline, to the benefit of papers in journals. The turning point seems to be situated somewhere around 2012, and can be interpreted as translating, to a certain extent, the impact of the VQR and its emphasis on publications in peer-reviewed periodicals. However, it is to be noted that the good reputation of papers in journals has not had a significant impact on chapters in books; if the figures decrease, in area 10, between 2011 and 2013, they observe a sharp increase in the following year, and then reach almost the same level as in 2004 (some 36%); this level is only 3 points underneath the peak of chapters publication observed in 2011. The evolution is much more hectic in area 11, where chapters in books tend to pass alternatively through good and bad years, but again values at the beginning and at the end of the period are comparable (40,6% of the total output in 2004; 40,8% of the total output in 2014), while the peak of publications is superior only by 5 points to what can be seen as an average value (40% of the output) (Figs. 7 and 8).

This relative stability of chapters in books in the context of a decrease of papers in proceedings can be interpreted as a hedging strategy of scholars, who may simply

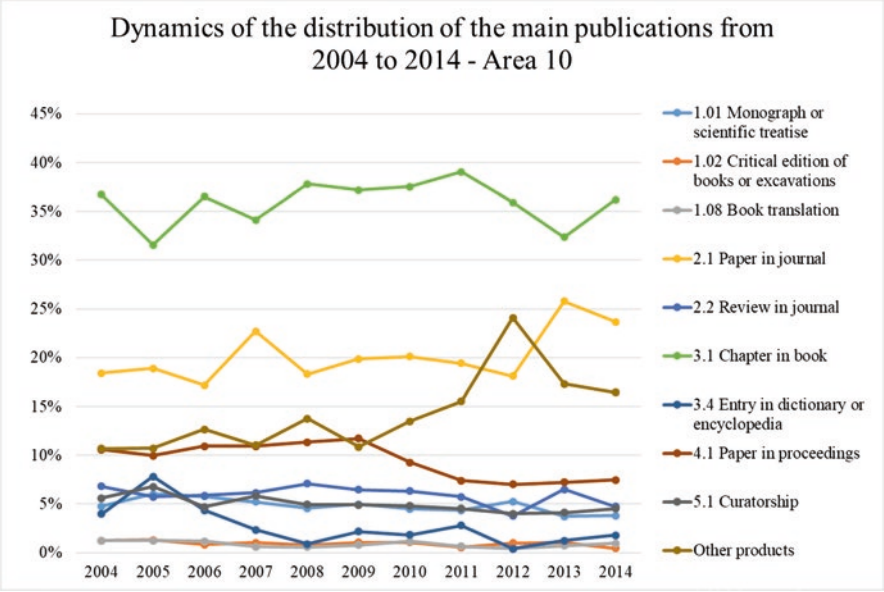


Fig. 7 Dynamics of the distribution of the main publications from 2004 to 2014 – Area 10 in CFU and MCU

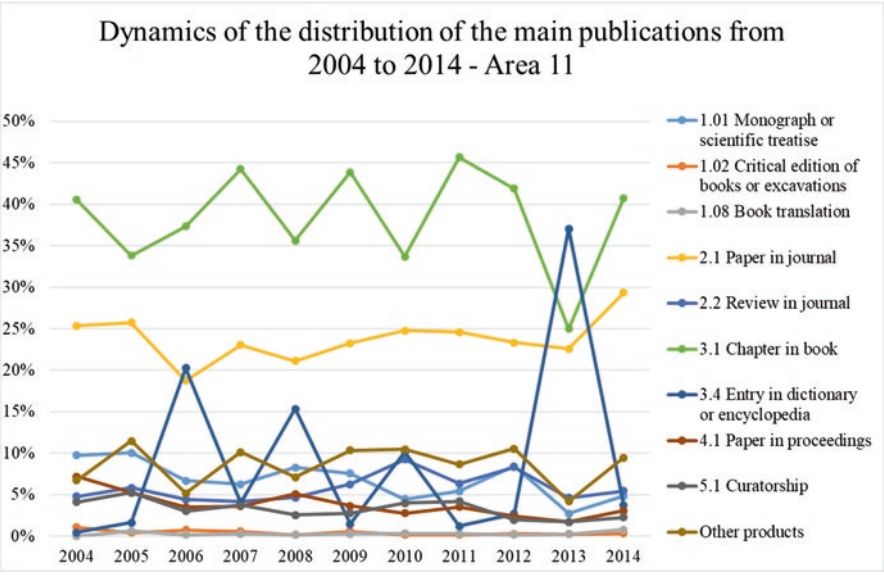


Fig. 8 Dynamics of the distribution of the main publications from 2004 to 2014 – Area 11 in CFU and MCU

have debaptised proceedings and transformed them into the more prestigious (or, at least, the more accepted) collective books.

On the whole, the data invite us to be cautious when focusing on books as the main or the most important publication type in the SSH. While it is clear that an evaluation of the SSH research based only on papers in journals would be meaningless, there is no reason to focus on books only, and even less to impose monographs as the crown output – particularly when judging the activity of young scholars. A nuanced approach is needed, rather than systematically overrating one category of outputs over another.

Can such a qualitative approach be alleviated, to a certain extent, by a semi-automated exploitation of the bibliographical data? In order to explore this possibility, ROBINBA tried to observe the peer-review practices in publishing houses, on the one hand, and to test, on the other hand, the usefulness of league tables of publishers. Both approaches are not the object of this chapter, and therefore we will not present them in detail. What appears however from both is that reputation is not necessarily a reliable criteria: in many cases, selection procedures are far from transparent, and interviews with scholars (see below) tend to suggest that even when a rigorous peer-review procedure is declared by the publisher, this may not exactly be the case in practice. One of the reasons of the increase in the number of papers published in journals may be, therefore, the fact that they usually rely on more robust procedures when it comes to assessing the quality of a piece of research; indeed, our questionnaire revealed that the rigorousness of the peer-review procedure is one of the criteria of quality to which scholars are particularly attentive. There is also a mismatch between the publishing houses where scholars publish their books and those they state as being prestigious: one can observe a clear national bias, even in disciplines abroad oriented (such as Hispano-American and Spanish language and literature studies). For all these reasons, one cannot necessarily equate quality with a book published in a given publishing house – with the exception of a few Anglo-Saxon based, and truly international, scholarly publishing houses. In order to create a reliable prestige indicator, there are different steps to be taken, maybe in the form of the Flemish initiative, where labels have been awarded to the publishing houses respecting clear peer-review specifications and practicing a reasonable selection.

3.3 Survey Data and Qualitative Analyses

In order to deepen our understanding of book publication practices, ROBINBA also adopted a qualitative approach. Observation of figures indicated, as stated above, that SSH scholars make use of a whole variety of outputs. What figures do not say is why scholars choose a type of publication over another, and how this relates to their perceptions and representations of research quality. In order to gather such information, a questionnaire and several focus groups were organised. In the following section, we will describe both methodologies, before focusing on some of their most interesting findings.

The questionnaire was adapted from one initially built for the French IMPRESSH study (Williams and Galleron 2016). It was then translated into English and Italian. The aim of the questionnaire was to understand and compare different practices of dissemination and to collect scholarly opinions about the evaluation processes; in particular, the focus of the questionnaire was to identify how books are weighted in different fields in the humanities. A first questionnaire was in English and was sent to members of humanities departments in Italy (mainly in Area 10 and Area 11), and, for comparative reasons, in Germany. Only the Italian data are discussed here. A second version, in Italian, was sent exclusively to Italian scholars, either directly or via the ROARs (Return on Academic Research) Facebook page. A total of 578 usable answers were received. See Annex 3 for the English version.

As it can be seen in the following figure, the largest percentage of respondents is in the 40–60 years bracket. This also translates into seniority in academia, as the percentage of senior researchers (at least 7 years of research experience after the PhD) is far superior to that of junior researchers (from 2 to 7 years of research experience after the PhD): 88,5% vs. 11,5%. As such, answers describe publication practices of those who have already acquired maturity in academia (Fig. 9).

In terms of gender, there is a balance between women and men, with a slight difference in favour of men (+8,8%). With regards to the research fields, there is a predominance of answers from scholars of area 10 (67, 6%), followed by area 11 (28,1%) (see Fig. 10).

Responders to questionnaires were free to voice their opinions, which led to a wide variety of responses. These were then listed and analysed statistically.

To these, we added the analysis of opinions expressed by experts during a kick-off meeting at Ca′ Foscari University of Venice and others voiced during three focus groups. These were recorded and data analysed using Atlas Ti. The qualitative information from the questionnaire was also dealt with in this way. It is to be noted that focus groups also produced some interesting reflections on the nature of books and the value of peer review, which were tacitly used in the previous sections of this paper.

What appears clearly from both questionnaires and the focus groups is the mismatch between aspirations and even practices of scholars, in terms of dissemination of their results, and the top-down imposed criteria, towards which they play for career reasons.

When scholars have been asked to state “all forms of dissemination they can think of”, the answers clearly showed an interest for reaching the broadest audience possible, whether formed by peers or by the lay public. Interviewed scholars obviously think in terms of channels for dissemination and their potential for reaching out. Therefore, publication forms are of lower concern, even if some voiced preferences because of the dissemination potential of certain types; indeed, when specific outputs are mentioned, they are obviously elicited because they target a large readership, this being the case of editions, dictionary and encyclopaedia entries as well as of catalogues. The main channels mentioned are illustrated below (Fig. 11).

There are some sharp differences in answers between areas,[5] but the commonalities are here of greater interest than the disagreements. To start with, one observes a

[5] For instance, area 10 has a stronger preference for repositories and posters, while area 11 makes more use of exhibition material. The latter is obviously discipline specific, the other two are less clear.

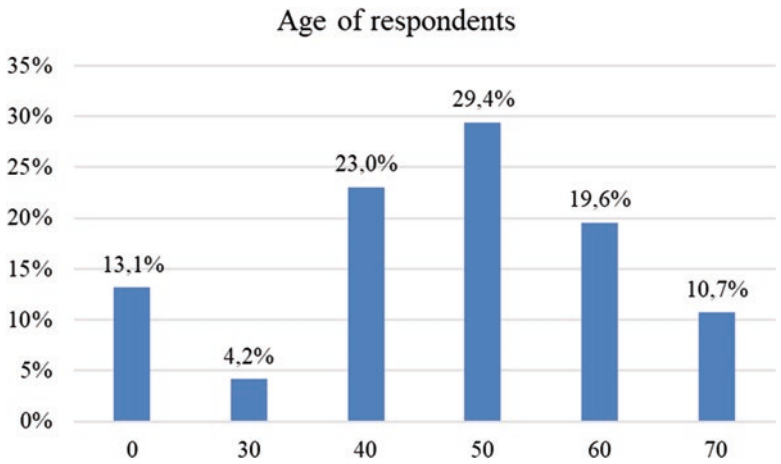
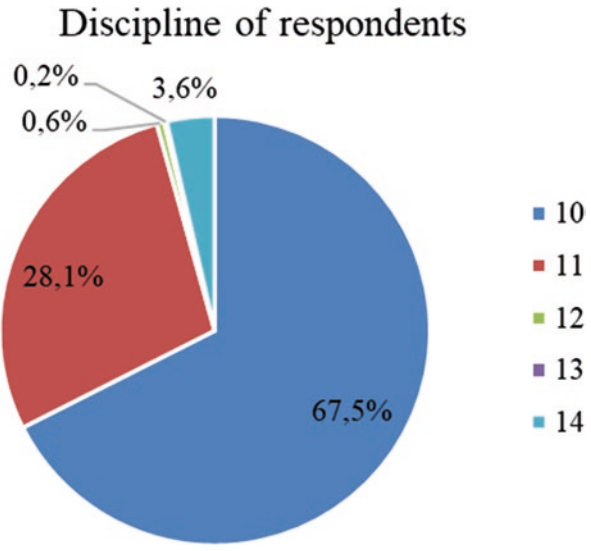


Fig. 9 Age of respondents to the survey
Legenda
0: missing
30: 25–34 years old
40: 35–44 years old
50: 45–54 years old
60: 55–64 years old
70: 65–74 years old

Fig. 10 Breakdown of respondents by discipline



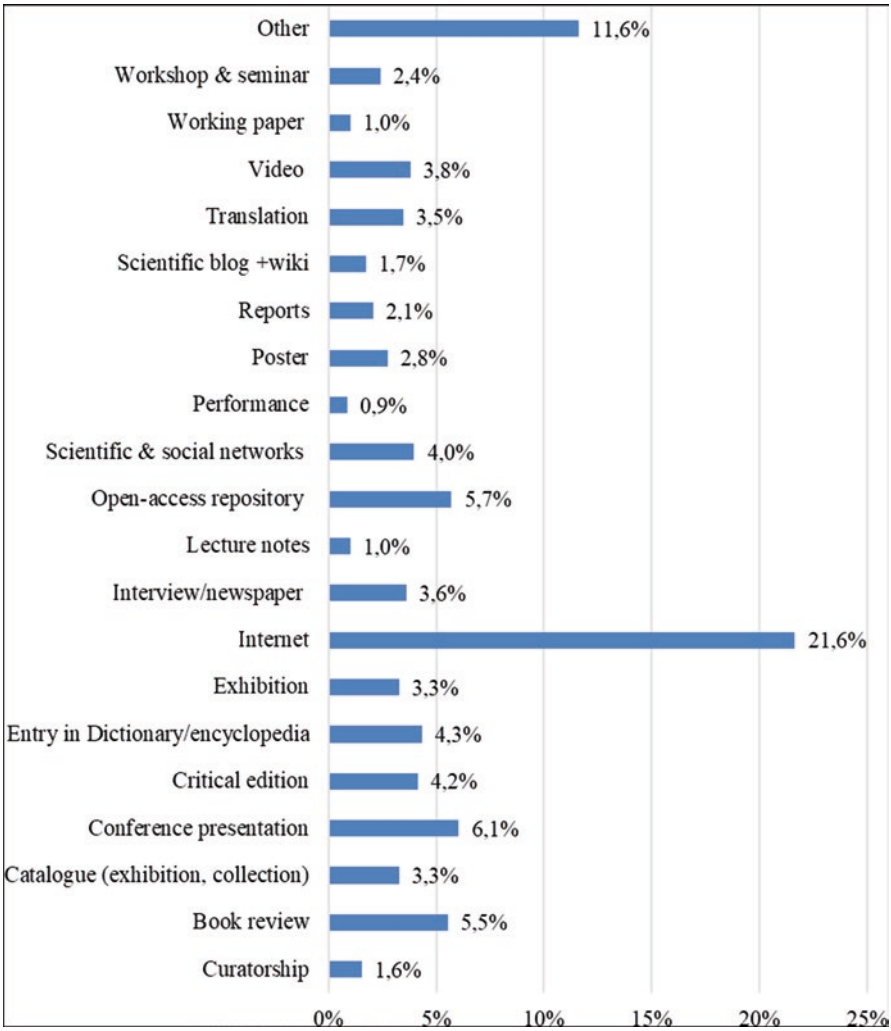


Fig. 11 Answers to question 1 of the questionnaire ("Please tell us all forms of dissemination you can think of, in addition to those reported below"). Percentage of total number of multiple responses. Categories below 0.9% of responses omitted

large number of answers in "Other", covering a list of many possibilities not listed here, thus demonstrating the variety of potential output channels. Maybe the most significant result is the clear and strong interest manifested in new technologies, such as Internet or online repositories as academia.edu. This underlines a desire to disseminate results of research using all possible channels rather than simply provide outputs that conform to evaluation criteria. It is unclear as to whether age represents a criterion here, even though it is mostly those over 40 who see Internet as significant (Fig. 12).

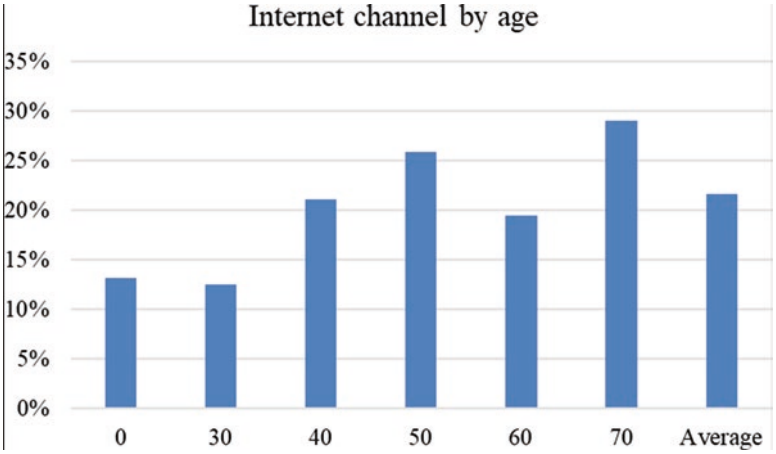


Fig. 12 Breakdown by age of respondents that mention "Internet" as a desired communication channel

Focus groups confirm the interest for reaching as large as possible an audience. One of the interviewed states:

> *I am using the classic channels (journals, book chapters, monographs) but also looking to alternative and more democratic ways of spreading scientific knowledge. As a specialist in African Studies I have realized that most of the published materials we highly value in western universities is not accessible to students, scholars and colleagues of non-developed countries.*

This remark is interesting as it is too easy to forget that scholarly impact is not only concerned with the country in which the work is written. If we aim at true internationalisation, and not simply at homogenised dissemination in English, then these multiple audiences should be considered.

However, many say that Internet and online publications have not fully entered the norm:

> *[Paper publications] are the traditional channels which are considered in the evaluations; on line publications are still not adequately valued.*

One respondent gave a very detailed picture of the importance of different practices declaring that:

> *Till some years ago, when we were urged to publish articles in peer-reviewed journals, our most prestigious ways of dissemination were book chapters (due to the fact that a book focuses on a specific research subject, and that book chapters are usually broader and more ambitious than journal articles), conference proceedings (which sometimes contain very original and up-to-date essays) and monographs. Since 1990s the field of digital scholarly editing has gained more and more importance in philology. Publication of journal articles is now increasing, under the pressure of recent editorial policy.*

Printed items (monographs, chapters in edited books and articles in specialized journals) are still the best and most durable form of dissemination of research results; but online publications and newspaper articles ensure a better timing and (often) a higher visibility.

Free developments in the field, or the use of a form of dissemination over another because it is the most fit to the topic, appear therefore to be hindered by external factors, such as institutional constraint, whose scientific validity is to be questioned.

Answers to the question 3 ("*Is there any reason why you are using this/these channel/s instead of others*?"), as well as the corresponding remarks during the focus groups, comfort this interpretation. The overall reason for both junior and senior researchers is conformism: they tend to resort to standard means of dissemination in their field, which, as it happens, is also what is evaluated. However, what is considered standard is not a fixed notion, and can change with the age, so that junior researchers tend to quote articles as being the standard dissemination type.

Alternative to this rationale, the choice of dissemination forms and channels is clearly pragmatic, especially amongst junior researchers. They prefer proceedings and articles, because:

The Call for Papers circulate quickly among scholars and evaluation is always based on blind peer-review, which ensures the quality of each publication.

Thus speed of publication, as well as a quality check, is seen as of importance as well as the visibility of the source, with Internet and repositories being seen as key channels:

I think the web is the best way to disseminate results (i.e. academia.edu)

Journal articles, especially those available online, may be the only tool to effectively disseminate findings in a lean, timely way.

In the same way, while senior researchers follow their field practice, some of them call for more innovative approaches. A good example is the following:

Not particularly. I try to follow standard practice in my field, but I am also quite open to new/innovative, and non-traditional, forms of dissemination, e.g. webinars, blog posts, etc. as ideal venues for the non-technical dissemination of research findings, project outcomes, etc. This is because part of my applied research is of interest to professionals (e.g. practising translators) and to a wider audience outside academia (e.g. English language teachers and non-specialist students).

Other answers underlined that proceedings allow for a rapid exchange, particularly with the audience at the event. Journal articles also benefit from speed of publications whereas book chapters and books are slower in disseminating findings, even if they have the advantage to allow for a more considered approach. Sometimes, this "slowness" is also compensated by other advantages:

Book chapters often come about after specialized events – it is easier to get innovative ideas published here than in traditional peer-reviewed journals which tend to be more conservative.

The ideal situation remains that of a choice guided solely by the nature of the topic:

Some major topics deserve books, other (less massive but relevant as well) need articles. If well connected with other chapters, book chapters too may be the best way to treat an object. Editing and supervising the production of a miscellaneous work may be a relevant scientific task, if the work is not conceived as a mere sum of self-referential contributes.

Focus groups also suggested that monographs are fit for representing the outcome of a broader and more complete research; they are perceived and meant as scientific and challenging outputs, because they require a long time (research for writing a monograph may require years) and when innovative, as they sometimes are, they can make a real difference. Unfortunately, scholars feel they can seldom consider only this aspect, since they are under lots of institutional pressure. Indeed, the urge is strong to publish books more than anything else, as revealed by the answers to the question 13: "*In your department, how is a book weighed compared to an article or a book chapter?*" (Fig. 13)

This shows clearly that departments are still asking for books and are not in line with either current trends or the wishes of researchers. If the informants consider that the relevance of books, and the pressure to publish those, has somewhat decreased since the internationalisation criteria has been added by ANVUR, we are still here in front of a clear case of the tail wagging the dog. There is therefore no surprise that when asked what are their usual dissemination forms, scholars confess

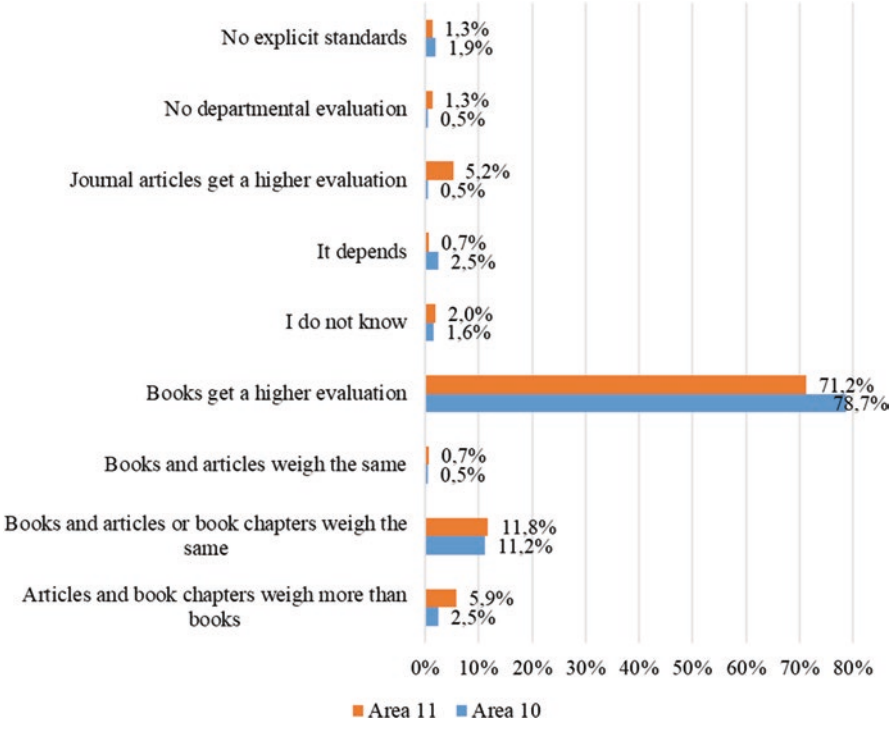


Fig. 13 Answers to question 13 of the questionnaire ("In your department, how is a book weighed compared to an article or a book chapter?"). Percentage of respondents (single answer)

very conservative attitudes, in definite contrast with the new technologies so widely mentioned in response to question number one of the questionnaire (Fig. 11).

Two other interesting observations are suggested by the focus groups. On the one hand, whilst books are considered very important in the SSH, and even fundamental in the career of a researcher, the interviews often do not always reflect the prominence of books over articles or other outputs. What matters is the quality, rather than the container. On the other hand, even if Italy has elaborated detailed typologies of books, it appears that the definition of monographs as "a rigorously peer reviewed output", a discriminant factor in certain research evaluation systems, does not appear clearly from interviews or in the institutional definitions.

4 Concluding Remarks

Humanities scholars do still publish books, but that is only part of the picture, as dissemination patterns are changing and there is a clear move towards more journal publishing. This move is most noticeable amongst younger researchers and may be for both evaluative and normative reasons. However, the speed of publication and peer review also seems to be prime factors, since getting a book into print is a lengthy procedure. The questionnaire and focus group data tend to show that books and book chapters are more appropriate for later stage researchers wishing to publish a state of the art, but this does not make them better than a good journal article in a journal with impact factor.

It may well be better to distinguish classes of publication with articles and proceedings being more suitable for work of immediate significance, an aspect that is particularly important for young researchers. On the other hand, international journals are difficult to access and so publications in high impact journals are more a criteria for judging senior researchers. Books and book chapters also suit more the publication needs of senior researchers for more elaborated productions.

There remains the question of quality of books. It is clearly not possible to lay down rules as to what is quality, and in this respect bibliometrics will not help. Possibly the best way to improve quality is to recognise that book publications should be the exception and not the norm. They suit certain types of research and at a certain time in the research cycle. Placing them as a quasi-obligation for recruitment or career advancement can only lead to what in the USA has been termed *Mickey Mouse publishing*, and the survival of poor quality pay-to-publish houses. Classifying publishers is an activity that is fraught with difficulty, but it is obviously necessary to carry out a survey along the lines of that done is Spain, preferably in collaboration with other research units working internationally in the field.

Annex 1: Broad Categories of Outputs: Aggregation of VQR Categories in ROBINBA Categories

VQR categories	ROBINBA typology
1.01 Monograph or scientific treatise	Monograph or scientific treatise
1.02 Critical edition of books or excavations	Critical edition of books or excavations
1.03 Scheda bibliografica	
1.04 Indice	
1.05 Bibliografia	
1.06 Pubblicazione di fonti inedite	
1.07 Commento scientifico	
1.08 Book translation	Book translation
1.09 Tesi di Dottorato	
2.1 Paper in journal	Paper in journal
2.2 Review in journal	Review in journal
2.4 Scheda bibliografica	
2.5 Abstract in Rivista	
2.6 Traduzione in Rivista	
3.1 Chapter in book	Chapter in book
3.2 Preface/Postface	
3.3 Brief introduction	
3.4 Entry in dictionary or encyclopedia	Entry in dictionary or encyclopedia
3.5 Traduzione in Volume	
3.6 Recensione in Volume	
3.7 Schede di Catalogo, repertorio o corpus	
4.1 Paper in proceedings	Paper in proceedings
4.2 Abstract in proceedings	
4.3 Poster in Atti di convegno	
5.1 Curatorship	Curatorship
7.01 Working paper	
7.02 Rapporto di ricerca	
7.03 Catalogo di Mostra	
7.07 Performance	
7.08 Mostra	
7.09 Esposizione	
7.13 Banca dati	
7.14 Software	
7.16 Other	Other products

All “neglected” forms of publications (scheda bibliografica, indice, etc.) were grouped under the “Other products” category (which includes, naturally, the “other” category from VQR).

Annex 2

Table 1 Distribution of the main typologies of publications by scientific areas in Ca′ Foscari and Macerata Universities

Scientific areas	Book	Chapter in book	Curatorship	Other	Paper in journal	Paper in proceedings	Patent	Total
1. Scienze matematiche e informatiche	1,26%	0,23%	2,74%	4,80%	41,26%	49,71%	0,00%	100,00%
2. Scienze fisiche	0,00%	0,00%	0,00%	4,50%	82,00%	12,00%	1,50%	100,00%
3. Scienze chimiche	0,83%	0,00%	0,44%	0,79%	63,79%	32,79%	1,36%	100,00%
4. Scienze della terra	1,26%	0,00%	0,00%	5,02%	53,97%	39,75%	0,00%	100,00%
5. Scienze biologiche	0,64%	1,03%	0,51%	1,92%	58,97%	36,67%	0,26%	100,00%
6. Scienze mediche	1,70%	16,48%	0,00%	0,00%	62,50%	19,32%	0,00%	100,00%
7. Scienze agrarie e veterinarie	2,81%	13,11%	2,58%	9,13%	52,46%	19,91%	0,00%	100,00%
8. Ingegneria civile e architettura	3,57%	5,36%	2,68%	2,68%	74,11%	11,61%	0,00%	100,00%
9. Ingegneria industriale e dell'informazione	1,77%	5,32%	2,48%	2,84%	42,91%	44,33%	0,35%	100,00%
10. Scienze dell'antichita,filologico-letterarie e storico-artistiche	6,89%	14,84%	6,36%	4,35%	57,95%	9,60%	0,00%	100,00%
11. Scienze storiche, filosofiche, pedagogiche e psicologiche	7,43%	31,55%	5,69%	2,29%	48,13%	4,90%	0,00%	100,00%
12. Scienze giuridiche	5,39%	34,02%	3,69%	1,75%	53,85%	1,29%	0,00%	100,00%
13. Scienze economiche e statistiche	4,95%	11,10%	2,70%	11,74%	57,75%	11,76%	0,00%	100,00%
14. Scienze politiche e sociali	7,67%	18,12%	7,67%	5,80%	58,53%	2,20%	0,00%	100,00%
Total	**5,71%**	**18,03%**	**4,66%**	**4,45%**	**55,32%**	**11,72%**	**0,11%**	**100,00%**

Annex 3: Questionnaire in English

1. Please tell us all the forms of dissemination you can think of, in addition to those reported below.
2. Which are your usual dissemination channels?
3. Is there any reason why you are using this/these channel/s instead of others?
4. When do you think books are an appropriate way of dissemination?
5. Please quote all types of books you can think of.
6. What is a good publishing house?

   [Existence of a thorough peer-review procedure]
   [Well defined publishing profile]
   [Existence of specific series]
   [Good reputation of the series coordinators]
   [Good reputation of the scientific committees]
   [Distribution of books (presence in international libraries and book stores)]
   [Transparency of publishing procedure]
   [Open access facilities]

7. Could you list 5 publishers you consider good in your field?
8. Why do you choose a publisher as opposed to another?

   [National/international good reputation]
   [National/international distribution]
   [Facility to publish with the selected publisher]
   [Existence of a thorough peer-review procedure]
   [Correspondence between your topics and publisher's series]
   [Open access facilities]
   [Other, please specify]

9. Could you list the publishers you have published with?
10. How much feedback did you get during the publication process?
11. What would make your work easier when publishing a book?
12. What are the criteria adopted by your university /institution to value books?
13. In your department, how is a book weighed compared to an article or a book chapter?
14. In your university, how is a book weighed compared to an article or a book chapter?

References

Association of Universities in the Netherlands (VSNU), Netherlands Organisation for Scientific Research (NWO), Royal Netherlands Academy of Arts and Sciences (KNAW). (2014). *Standard evaluation protocol 2015–2021: Protocol for research assessments in the Netherlands*. https://www.knaw.nl/nl/actueel/publicaties/standard-evaluation-protocol-2015-2021

Dolan, C. (2007). *The evaluation and benchmarking of humanities research in Europe*. Arts and Humanities Research Council, UK. http://heranet.info/system/files/HERAJRPdocuments/Deliverables/d4.2.1_hera_final_report_on_evaluation_and_benchmarking_of_h.pdf

Engels, T., Ossenblok, T., & Spruyt, E. (2012). Changing publication patterns in the social sciences and humanities, 2000–2009. *Scientometrics*. https://doi.org/10.1007/s11192-012-0680-2.

Galleron, I. (2016). Enabling the SSH. *PEN: Science & Technology, 19*, 75–77.

Giménez-Toledo, E. (2016). Assessment of journal & book publishers in the humanities and social sciences in Spain. In M. Ochsner, S. Hug, & H. D. Daniel (Eds.), *Research assessment in the humanities. Towards criteria and procedures*. Cham: Springer. https://doi.org/10.1007/978-3-319-29016-4_8.

Giménez-Toledo, E., Mañana-Rodríguez, J., & Tejada-Artigas, C. M. (2015a). Review of national and international initiatives on books and book publishers assessment. *El profesional de la información, 24*(6), 705–716.

Giménez-Toledo, E., Mañana-Rodríguez, J., & Tejada-Artigas, C. M. (2015b). Scholarly publishers' indicators: Prestige, specialization, and review systems of scholarly book publishers. *El profesional de la información, 24*(6), 855–860.

Giménez-Toledo, E., Mañana-Rodríguez, J., Engels, T., Ingwersen, P., Polonen, J., Sivertsen, G., Verleysen, F., & Zuccala, A. (2016). Taking scholarly books into account: Current developments in five european countries. *Scientometrics*. https://doi.org/10.1007/s11192-016-1886-5.

ISBN Agency. (2012). *ISBN users' manual: International edition*. Sixth Edition. https://www.isbn-international.org/sites/default/files/ISBN%20Manual%202012%20-corr.pdf

Mutz, R., Bornmann, L., & Daniel, H. D. (2013). Types of research output profiles: A multilevel latent class analysis of the Austrian science fund's final project report data. *Research Evaluation, 22*, 118–133. https://doi.org/10.1093/reseval/rvs038.

Ochsner, M., Hug, S. E., & Daniel, H. D. (2013). Four types of research in the humanities: Setting the stage for research quality criteria in the humanities. *Research Evaluation, 22*(2), 79–92.

Ochsner, M., Hug, S., & Daniel, H. D. (Eds.). (2016). *Research assessment in the humanities. Towards criteria and procedures.* Cham: Springer. https://doi.org/10.1007/978-3-319-29016-4.

Office of the Government of the Czech Republic. (2013). *Methodology of evaluation of research organizations and evaluation of finished programmes (valid for years 2013–2015)*. http://vyzkum.cz/FrontClanek.aspx?idsekce=695512. Accessed 14 July 2017.

REF. (2012). *Panel criteria and working methods REF 01 12*. http://www.ref.ac.uk/media/ref/content/pub/panelcriteriaandworkingmethods/01_12.pdf. Accessed 14 July 2017.

Royal Dutch Academy (KNAW). (2011). *Quality indicators for research in the humanities*. https://www.knaw.nl/shared/resources/actueel/publicaties/pdf/quality-indicators-for-research-in-the-humanities. Accessed 14 July 2017.

Sivertsen, G. (2015). Patterns of internationalization and criteria for research assessment in the social sciences and humanities. *Scientometrics*. https://doi.org/10.1007/s11192-016-1845-1.

Vincent, N., & Wickham, C. (Eds.). (2013). *Debating open access*. London: The British Academy.

Williams, G., & Galleron, I. (2016). Bottom up from the bottom: A new outlook on research evaluation for the SSH in France. In M. Ochsner, S. Hug, & H. D. Daniel (Eds.), (2016) *Research assessment in the humanities. Towards criteria and procedures*. Cham: Springer. https://doi.org/10.1007/978-3-319-29016-4_14.

Wilsden, J. (2015). *The metric tide: Report of the independent review of the role of metrics in research assessment and management*. HEFCE. http://www.hefce.ac.uk/pubs/rereports/year/2015/metrictide/. Accessed 14 July 2017.

Wouters, P. (2014). The new Dutch research evaluation protocol. *The Citation Culture*. https://citationculture.wordpress.com/2014/05/08/the-new-dutch-research-evaluation-protocol/.

Zaccai, E., Timmermans, B., Hudon, M., Clerbaux, B., Leclercq, B., & Bersini, H. (2016). *Penser la science: L'évaluation de la recherche en question(s)*. Bruxelles: Académie royale des Sciences/des Lettres et des Beaux-Arts de Belgique.

Research Quality Criteria in the Evaluation of Books

Carla Basili and Luca Lanzillo

1 The Role of Books and Monographs in the Social Sciences and Humanities

In the Social Sciences and Humanities, particularly in the latter, books and monographs have long been considered as valuable research outputs, contributing to the scholarly discourse. Ochsner et al. (2016) and Giménez-Toledo et al. (2016) summarise the extensive literature in demonstration of this statement. However, in the context of research policy, books and monographs are also extensively discussed as products in crisis, due to the threats to their quality and production.

The specialised literature has in fact been increasingly concerned with the role of books and monographs given the current economic climate, the imperatives of the publishing market, and to a lesser extent as regards the issue of research evaluation. The current chapter is not the place for an exhaustive literature review of the many issues at play, and instead concentrates on the issue of the role of the book in the communication habits of the SSH communities.

Influential studies on the topic begin from the assumption that in the Humanities, books and monographs are research products of great value (Hicks 2004, 2006), but then highlight their progressive decline in sales, usage and quality (Williams et al. 2009) while identifying a set of causes. Much of the literature identifies technological factors as determinants both of the general decline in publishing (Bennett 1998; Elliott 2015; Joseph 2015) and the specific crisis in academic sphere (Thompson

C. Basili (✉)
Istituto di Ricerca sulla Crescita Economica Sostenibile (IRCRES), Consiglio Nazionale delle Ricerche (CNR), Rome, Italy
e-mail: carla.basili@ircres.cnr.it

L. Lanzillo
Dipartimento di Scienze documentarie, linguistico-filologiche e geografiche, Università di Roma La Sapienza, Rome, Italy
e-mail: lu.lanzillo@gmail.com

© Springer International Publishing AG 2018

A. Bonaccorsi (ed.), *The Evaluation of Research in Social Sciences and Humanities*, https://doi.org/10.1007/978-3-319-68554-0_7

2002; Williams et al. 2009). In fact, when limited to academia, the monograph has high production costs and is a niche product. Chodorow (1999) provides trenchant comment:

> Monographs are economically inefficient, because they are large, specialized, expensive works with small audiences.

Thompson (2002) wonders if the monograph is now dead, but then concludes that:

> Clearly the scholarly monograph in the humanities is not dead. […] Articles form an important aspect of literary research, but are not substitutes for monographs. […] Moreover, the university press monograph is still arguably the most significant vehicle for scholarly communication in this field.

In his report to the Higher Education Funding Council for England (HEFCE), Crossick (2015) takes a similar position:

> When we think about the monograph it is therefore important to avoid the danger of seeing it as an awkward outlier in relation to a mainstream framework of research communication defined by the journals and refereed conference proceedings that dominate the sciences.

In an observation of particular interest to the current chapter, Williams et al. (2009) identify the pressures of research assessment exercises as a specific constraint on the production of *quality* monographs:

> There was universal agreement that the RAE did not facilitate quality monograph writing. […] the system makes monograph production more difficult, because of the artificial time limit and consequent pressures of time and space – factors detrimental to the kind of scholarship that produces monographs.

Williams et al. (2009) address the problem of measuring the role and value of Arts and Humanities research monographs "in intellectual, personal fulfilment and career-related terms". Their conclusions confirm the common perception that in the relevant academic community, monographs are more valued than journal articles. Crossick's HEFCE report (2015) corroborates the point, extending the conclusions of value beyond monographs to "other longer research publications":

> Monographs, and other longer research publications such as edited books, scholarly editions and exhibition catalogues, make a fundamental contribution to scholarly communication within many disciplines. They offer the space needed to set out arguments and evidence in disciplines where that is necessary, complementing the importance of articles in learned journals. Monographs are not, however, simply a means of communicating research findings. The process of constructing and writing a book is often a core way to shape the ideas, structure the argument, and work out the relationship between these and the evidence that has emerged from the research process. At their best, monographs provoke debate, can shift paradigms, and provide a focal point for research.

Crossick's comments, above, are fundamental to the current chapter, and are quoted at length for series of reasons: these are the conclusions of a major consultative and data-gathering project, reporting to a national research funding agency; the project analysed the full spectrum of SSH disciplines, and takes a position based on very recent investigation. In addition, the conclusions were further confirmed in a summing-up article of 2016, as follows Crossik (2016):

> The monograph has a central place in the culture and ecology of research publication in the arts and humanities, and is important in most of the social sciences.

On the other hand, Chodorow (1999) suggests a highly provocative proposition:

> The decline of the monograph might be a sign that the academic environment has become *inhospitable for the humanities*.

This was written at a time when academic publishing was severely threatened by the prospects of technological development, and the crisis in publishing was a dominant feature in the research landscape. Since then the scenario has changed and the publishing situation has consolidated remarkably, yet Chodorow's caveat still seems worthy of careful consideration (see Lanzillo's chapter in this book).

Returning more directly to the question of value, one of the strategic areas of research concerns the disadvantaged position of books in the international databases used in evaluation procedures (Giménez-Toledo et al. 2016) and the exploration of alternative metrics and reputation mechanisms (Zuccala et al. 2015). Based on a study of the different national assessment systems for books, Giménez-Toledo et al. (2016) identify that all of these research lines deal with common considerations and elements: "the existence of peer review processes, the prestige of book publishers and the specific features of each discipline". The last element is discussed in the next section.

1.1 Knowledge Dissemination in SSH: Distinctive Features in Italy

As early as 1970, Derek de Solla Price had observed the fundamental importance of disciplinary variation in the dissemination of research results.

> Within common principles of publication and communication norms, scientific communities exhibit quite different behaviour depending on their type of research, their degree of application and the nature of their field. Even when sharing the same communication system, different disciplines do not publish with the same frequency, do not exhibit the same propensity to collaborate and co-author papers, nor have the same citing practices both in volume (the length of references list in the articles) and immediacy (the age of references they cite). These discrepancies were evidenced in early literature on publication and citation practices (Price 1970) as cited in Zitt and Bassecoulard (2008).

In general, and particularly in the Italian academic system, dissemination of research results in SSH shows characteristic features, not generally shared with the hard sciences: (a) journal articles are not the main publication channel; (b) the language used is not exclusively English; (c) the best articles are not necessarily published in *core* journals; (d) it is common to take a retrospective attitude in research, and as a consequence, long-term impact of publications must be expected and accounted for in the research evaluation processes.

Gimenez-Toledo et al. (2015) and Huang and Chang (2008) offer particularly important analyses of books as research output, and more specifically of research

output in SSH. A detailed description of the SSH disciplines in Italy is provided in other chapters of this volume. Here we note only aspects concerning the dissemination of research results, and more specifically, the forms of results that can be categorised as *books* or *book parts*.

1.2 The Monograph as Research Product

Gimenez-Toledo et al. (2015) and Huang and Chang (2008) offer particularly important analyses of books as research output, and more specifically of research output in SSH. In this and the following subsections we will see how *books* can vary in form and content, depending on the objectives of the underlying research and the intended uses of results. Within this category, the first type we consider is the *monograph*: a research product that is particularly difficult to assess given its remarkable complexity in structure and content.

The monograph is the product of an intense but wide-ranging, systematic and unified research examination of an area of study. Each element contributes to forming the complex of the work, which could not be successfully communicated through publication of separate parts. Such extensive works are generally the result of many years of research, and not infrequently of an entire career.

The monograph has its own editorial style, different from that of articles in journals, but within this there can also be variations, depending on factors such as the discipline, the contents requiring communication, and the intended audience.

At the same time, the research presented in monographs may have high levels of interdisciplinarity: Broadus (1971) stresses that books cite more widely than journals, indicating disciplines other than those of the author.

For the humanist, the monograph is typically the privileged form of study, one where the researcher is at ease. While the journal article is the place of trials, the monograph is that for establishing and strongly affirming ideas. The monograph then gives the investigator the time and space necessary to structure their research, strengthen the content, and show it in entirety.

Monographs resulting from SSH research often consist of combinations of core text and para-textual elements. The preface, introduction, indices, appendices, illustrations and tables are necessary for completion and consolidation of the work, and without them, much of the scientific value would be lost.

A true, single-author monograph is not comparable to a book of chapters written by individual authors. Even coauthorship of a monograph can be problematic, in hampering textual and conceptual unity. Participation by joint authors requires that the different contributions merge to create a product which is more than the simple sum of chapters. Academics may avoid the coauthorship approach not only because of practical difficulties, but also because of the emphasis on individual assessment.

Still, it is quite common for books to be written by many hands, with each author assigned the responsibility for one or more chapters. In Italy, this type of product is often a response to the needs of *recruitment* systems, given that the committees responsible for recruitment and career advancement generally require that the can-

didate demonstrate direct personal authorship of publications listed in their applications. By organising and sharing in the production of multi-author books, individual academics can communicate their research in a manner that bears on recruitment and advancement.

1.3 Research Product Classification in Italian National Evaluation Exercises and the SUA-RD Template

The Frascati Manual definition of research (OECD 2002, p. 30),[1] begins with the words "creative work undertaken on a systematic basis …". The definition provided allows association of a wide range of research products, in which the power of communication depends on both content and means of presentation. The product may be devised for different roles, depending on the initial purposes of the research and the kinds of results to be conveyed. In assessing the quality of a research product it is therefore essential to consider its communicative effectiveness in relation not only to the results it is intended to convey, but also in relation to the objectives of the communication and its intended recipients.

The classification of the products of research is therefore fundamental to the evaluation process, and must take account of the aims of the communication, such as: for advancement of science within the research system; for purposes outside the system, in the world of production (e.g. for innovation); for general benefit of the citizenry (e.g. popularization of science) (Basili 2003).

In the Italian context there have been several rounds of national evaluation exercises. As these progressed, there has been a significant increase in the classes and detailed specifications of the types of objects considered.

Ministry of University Education and Research (MIUR) Decree no. 2206/RIC of December 2003 regulated the first Italian evaluation exercise, covering scientific production over the 2001–2003 triennium (*Valutazione Triennale della Ricerca*, VTR 2001–2003) (CIVR 2007). Implementation of the exercise was entrusted to a Committee for Research Evaluation (CIVR). The CIVR directives specified the types of research products eligible for consideration (De Santis 2006, p. 24):

- Books and their chapters;
- Articles in journals [scientific];
- Patents;
- Projects, compositions, drawings and design;
- Performances, exhibitions and exhibitions;
- Artefacts and works of art.

Does not include purely editorial activities; software; texts of exclusive educational interest; abstracts of conferences.

[1] Research and experimental development (R&D) comprise creative work undertaken on a systematic basis in order to increase the stock of knowledge, including knowledge of man, culture and society, and the use of this stock of knowledge to devise new applications.

For the second national assessment exercise, responsibility was passed to the Italian National Agency for the Evaluation of Universities and Research Institutes (ANVUR). In November 2011, ANVUR published the call for participation in the Evaluation of Research Quality (*Valutazione della Qualità della Ricerca*, or VQR 2004–2010). The products eligible for evaluation were indicated as (categories including *books* and *book parts* emphasised) (ANVUR 2011, p. 3):

(a) journal articles;
(b) **books, chapters of books; conference proceedings (only if assigned ISBN);**
(c) **critical editions, translations and scholarly commentaries;**
(d) patents awarded over the 7-year evaluation period, for which the subject being evaluated is listed as patent holder/co-holder;
(e) musical compositions, drawings, performances, organised exhibitions, works of art and their plans; designs and prototypes; databases and software; thematic maps (only if accompanied by publications permitting adequate evaluation).

Subsequently, Legislative decree 19/2012 laid the foundations for a new university "Self-evaluation, Periodic evaluation, Accreditation" system (*Autovalutazione, Valutazione periodica, Accreditamento,* AVA) (ANVUR 2013), based on the standards and guidelines adopted in 2005 by the European Association for Quality Assurance in Higher Education (ENQA). The AVA guidelines are further specified in the form of the "Single Annual Report Form on Departmental Research" (*Scheda Unica Annuale delle Ricerca Dipartimentale*, or SUA-RD).

Since 2013, each university has been responsible for ensuring the compilation and submission of these forms, for every department, on an annual basis. The form includes Annex A (ANVUR 2014), which specifies the types of research products eligible for the evaluation procedure.

The SUA-RD form is the instrument through which the universities conduct internal self-assessment, evaluation and accreditation procedures, preparatory to the ANVUR evaluation process. The ANVUR Guidelines (ANVUR 2016a) describe the SUA-RD structure, consisting of three main parts:

1. Departmental objectives, resources and management;
2. Research results;
3. Third mission results.

In the first part, the university department reports its organisational structure, human resources and infrastructure, its implementation of multi-year programmes in keeping with policies and the university strategies, and procedures for verification relative to objectives.

The second part reports the departmental scientific activity in terms of research products, international collaboration and mobility, receipt of competitive funding for national and international projects, teaching and research abroad, and scientific awards received.

The third part gathers information on activities concerning the third mission, namely economic exploitation of research (intellectual property, spinoffs, third

party research, intermediaries), and the production of public and social good (cultural heritage, health, public engagement, continuing education).

The SUA-RD form also served as the basis of classification for research products considered under the third national research assessment exercise (VQR 2011–2014). Before proceeding, we should note certain structural features of the research evaluation process in Italy, relating to disciplinary groupings and their respective evaluation panels.

The VQR identifies 16 research areas, each with an Expert Evaluation Panel (*Gruppo di Esperti della Valutazione*, GEV) appointed by ANVUR and entrusted with assessing the relevant research products. In the Italian academic system, the Social Sciences and Humanities are divided into five disciplinary (non-bibliometric) areas, each of which is assigned a GEV:

– Antiquities, Philology, Literary Studies, Art History (Area 10)
– History, Philosophy, Pedagogy (Area 11a)
– Psychology (Area 11b)
– Law (Area 12)
– Economics and Statistics (Area 13)
– Political and Social Sciences (Area 14)

For purposes of the VQR, each discipline is further subdivided in fields, each with a "sub-GEV". For each field, the specific modes of disseminating research results are considered. By way of example, Area 10 (Antiquities, philology, literary studies, art history) is subdivided in the following fields:

– Sub-GEV 1: Classical and Oriental Studies
– Sub-GEV 2: Art, Cinema, Performance
– Sub-GEV 3: Foreign Languages and Literatures, Comparative Studies
– Sub-GEV 4: Italian Studies and Linguistics

Table 1 shows the correspondence between Area 10 and the relative European Research Council (ERC) fields.

Given the differences between research fields as widely distant as musicology, neurolinguistics and historiography, we expect analogous differences in their respective research outputs and dissemination strategies. This is the reason why the VQR 2011–2014 guidelines assign substantial responsibility to the individual disciplinary evaluation panels (GEVs) for selecting the types of research products eligible for evaluation.

ANVUR (2015a, p. 5–6) provided a full list of the products eligible for the VQR 2011–2014, from which the GEVs could select the ones suitable to their fields. The quotation below cites only the wording concerning types which coincide with our categories of *books* and *book parts*:

> […] Publication types described below represent the total set of eligible categories. Every GEV may, on the basis of the characteristics of research areas related to it, clarify or restrict the set of types eligible for assessment, motivating the decision in their document on evaluation criteria.

Table 1 Italian disciplinary area 10: correspondence to ERC panels and fields

ERC SSH panels	ERC Fields corresponding to those covered under Disciplinary Area 10
SH5. Cultures and Cultural Production: Literature and philosophy, visual and performing arts, music, cultural and comparatives studies	SH5_1 Classics, ancient Greek and Latin literature and art
	SH5_2 History of Literature
	SH5_3 Literary theory and comparative literature, literary styles
	SH5_4 Textual philology, palaeography and epigraphy
	SH5_5 Visual and performing arts, design
	SH5_6 Philosophy, history of philosophy
	SH5_7 Museums and exhibitions
	SH5_8 Music and musicology, history of music
	SH5_9 History of art and architecture
	SH5_10 Cultural studies, cultural diversity
	SH5_10 Cultural heritage, cultural memory
SH4. The Human Mind and its Complexity: Cognitive science, psychology, linguistics, education	SH4_6 Formal, cognitive, functional and computational linguistics
	SH4_7 Typological, historical and comparative linguistics
	SH4_8 Psycholinguistics and neurolinguistics: acquisition and knowledge of language, language pathologies
	SH4_9 Use of language: pragmatics, sociolinguistics, discourse analysis, second language teaching and learning, lexicography, terminology
SH6. The Study of the Human Past: Archaeology, history and memory	SH6_1 Archaeology, archaeometry, landscape archaeology
	SH6_2 Prehistory and protohistory
	SH6_3 Ancient history
	SH6_11 Cultural history, history of collective identities and memories
	SH6_12 Historiography, theory and methods in history

1. Scientific monographs and similar products:

   (a) Research monographs
   (b) Collections of own research articles (excluding articles published prior to 2011)
   (c) Systematic indices
   (d) Scientific comment, illustrating and providing innovative interpretation of aspects of meaning, language, style, historical-cultural context, techniques of composition or the history of reception of texts of previous knowledge
   (e) Annotated or critical bibliographies
   (f) Critical editions of texts by other authors

(g) Critical editions of previous excavations, based on scientifically conducted research, applying specified methods towards stated objectives, and concerning substantial areas of excavation
(h) Publications of unpublished sources with introduction and commentary
(i) Scholarly textbooks (not basic instruction/training manuals)
(j) Grammatical and scientific dictionaries
(k) Translation of books (by decision of the relative GEV), where these are interpretive, critical works on the part of the translator

2. Contributions to journals […]
3. Contributions to volumes

(a) Contributions to volumes (chapters and substantive sections)
(b) Scientific papers in conference proceedings subject to peer review
(c) Prefaces/appendices of substantive character
(d) Edited volumes, with editorship including substantive introductory essay
(e) Catalogue with substantive introductory essay
(f) Substantive and original commentaries in dictionaries or encyclopaedia
(g) Translation in volume, only for those disciplinary sectors where scientifically relevant (by decision of the relative GEV)
(h) Substantive sections of catalogues, repertories or corpora of works

4. Other types of scientific products […]
5. Patents […]

The GEVs shall establish criteria to determine whether submissions of the following publications provide significant innovations qualifying them as suitable for evaluation as research products:

1. Revised editions and translations of works published prior to 2011
2. Introductions and/or postfaces to revised editions of works published prior to 2011
3. Abstracts

The following are not eligible for VQR evaluation:

1. Manuals and training texts
2. Reviews of single works, lacking critical analysis of the related broader literature
3. Short entries in dictionary or encyclopaedia without character of originality
4. Brief editorial commentaries, without character of originality
5. Brief catalogue entries, not requiring independent scientific contribution.

From the above quotation we can observe the important responsibilities delegated to each disciplinary GEV concerning the admission of products to the VQR 2011–2014.

Also notable in this short history of the various evaluation exercises is the gradual extension of the classes of eligible products. The single category of *books and their chapters* initially indicated in the 2006 CIVR guidelines has now been separated into two classes, of *scientific monographs and similar products* and *contributions to volumes*, in turn subdivided into a further 11 and eight subclasses. The process of this gradual increment involved the intermediate steps of SUA-RD form for self-evaluation, which in 2013 indicated 23 *Book* and seven *Contributions to volume* subclasses.

What are the causes of these changes? Why has the Italian classification become so complex, and in doing so become distanced from international practice? A first reason could be a desire to respond to the peculiar needs of the Humanities, as

opposed to other disciplinary areas, given the type of categories added. A comparison of the SUA-RD form to the relevant international standards (see Sect. 3) seems to confirm this perception.

We can gain further understanding of these developments by considering another recent experience, occurring in the Italian academic system in the years between the VQR 2004–2010 exercise and the 2013 introduction of the AVA "auto-evaluation" system. This was the "National Scientific Habilitation" (*Abilitazione Scientifica Nazionale*, ASN). Established under the most recent legislative reforms of the Italian university system (Law no. 240/2010), the ASN is a nation-wide assessment for "pre-qualification" of potential candidates to university positions as full or associate professors. Candidates who achieve "habilitation" are then eligible to participate in the individual faculty competitions.

For purposes of the ASN, each scientific area was divided in *competition sectors* (*settori consorsuali*, SCs). Applicants under each sector were judged by a panel of five experts, including one expert from a non-Italian OECD country. The experts were selected from a list of full professors, who in turn competed as candidates for these roles. In the first ASN procedure, the panels evaluated the candidates' scientific expertise on the basis of their curricula and the papers they chose for submission.[2]

To be eligible for consideration, candidates had to have achieved the so-called "research product medians". In non-bibliometric areas these were specified in terms of: (a) numbers of books with ISBN code; (b) numbers of journal articles and chapters in books with ISBN code; (c) number of journal articles in top-rated journals. These metrics were applied to the candidate's scientific production over the 10 years immediately prior to the issue of the MIUR ministerial decree (no. 76/2012) which set out the criteria and parameters for habilitation. Each panel could choose a selection or adopt all the criteria and parameters, based on their considerations of the sectoral characteristics.

The fact that the 2012 administrative regulation (decree law) established criteria and parameters represent a very significant difference between the national assessments of institutions (VQR) and of scholars (ASN). Apart from the basic framework determined at central level, the VQR criteria are established by each GEV in consideration of the relative scientific communities, while in the ASN the criteria are established by legislation, in rather broad manner. But in the Italian context, legislation makes it possible to launch appeals before the court, and in fact hundreds of applicants in the first two ASN competitions registered complaints in administrative tribunals. A large part of these cases can be attributed to the ministerial classifications, which did not take due account of the disciplinary specificities of research outputs. Terms such as *book* and *book chapter*, used in defining the ASN *medians*, are very general: the ministerial decree did not define these terms or describe the scientific nature of their content. *Book* or *book chapter* could then include products as insignificant as a basic instruction manual, or a short narrative introduction to a volume

[2] Up to 12 or 18 publications in non-bibliometric areas, depending on whether the candidate tendered for role of associate professor or full professor.

by other authors. Another problem was that the lists of products indicated by the individual candidates were then simply counted, without submission to peer review.

This experience demonstrated that the communication of clear definitions was necessary, to permit common application by all actors: candidates, panel members, and even administrative tribunals. A more fully articulated classification also sets limits on what can and cannot be submitted for evaluation, in both qualitative and quantitative terms.

Another important aspect of the Italian context has been the lack of a single national repository of research products, common to all research institutions.[3] Historically, each university managed its own repository in isolation from others. Recently, CINECA (a non-profit consortium of 70 Italian universities) established an Institutional Research Information System (IRIS), which subsumed two previous national information systems, and a number of universities began migrating their repositories of research production information to this new standard. However, IRIS derived its classifications of research products from the previous U-Gov system, which is different from the one developed by ANVUR. The IRIS operators state that to now merge the two classifications would require a deep and complex re-set of the entire system. In addition, the IRIS system relies on the scholars themselves to enter the metadata for their research products. The process of data entry is not simple, and the information as entered by the scholars is in only a few cases monitored by the relative university or departmental librarians, or subject to any form of ex-post quality control.

The necessity for clear and articulated classifications and definitions is therefore even more urgent in Italy than in other countries, and the Italian experience can serve as a case study of particular interest in an international comparative perspective.

2 National and International Systems of Quality Criteria for Books

2.1 Objectives and Methodology

In the light of the Italian experience, the aim of this chapter is to outline a methodology for a structured and systematic classification of research products referable to the class of *book* and *its parts*, on the basis of a comparative analysis of national and international classification systems. We concentrate the analysis on international and national experiences of particular relevance to the Italian experience, in order to move towards an improved common approach, applicable in diverse contexts. However, proceeding from the conclusions of the current study, and applying the

[3] As occurs, for example, in Norway with CRIStin. In Italy a first important step has been made the census of research products operated by SUA-RD and the obligation for each scholar to provide an ORCID identifier within the VQR 2011–2014.

same methodology, other countries or international bodies could carry out further analyses, thus building towards a common European repository for SSH research. The current European repository, Zenodo,[4] evolved from international practices, and suffers from very limited classification for SSH, which could penalize the exceptional variety of production in this area.

Achievement of a broad, structured classification would be a step towards a true ontology of *books* and *book parts,* which at present does not exist. Among many purposes, this would support the specification of evaluation exercises, the choice of products for submission by scholars, and the evaluation processes conducted by referees and panels.

In view of the above, we begin our analysis by examining the classification systems commonly used at the international level.

The objective to consider all types of publication related to the category *book* led to the decision to include in the analysis any form of publication with an ISBN code, assigned as described in ISO standard 2108:2005:

> The purpose of this International Standard is to establish the specifications for the International Standard Book Number (ISBN) as a unique international identification system for each product form or edition of a monographic publication published or produced by a specific publisher. It specifies the construction of an ISBN, the rules for its assignment and use, the metadata to be associated with the ISBN allocation, and the administration of the ISBN system.

The presence of the ISBN code, however, proved to be a condition necessary, but not sufficient in itself, since the application criteria are imprecise for our purposes.

First, the ISBN system applies to the book as container, without entering into the merit of the content. Since the purpose and indeed the form of publication would be determined by the intended communication of contents, the latter is in fact of great interest to our classification. Secondly, the ISBN standard seems to consider the *books* and *monographs* without distinction, whereas for our purposes, classification of the monograph distinctly from the different forms of books is essential to the subsequent assessment procedures.

Indeed technological developments have made it possible to publish research products in multimedia forms, ranging from texts with simple attachments of experimental data to much more complex hypermedia. At the same time the increasing specialisation of the disciplines has broadened the concept of the *book* remarkably. In fact disciplinary specialisations and diversifications, particularly within the Humanities, have determined that scientific results are produced in many forms of *books.* These products then assume varying scientific relevance within the scholarly domains.

For this reason, we next proceed to examining other international and national guidelines, including Italian ones, to achieve greater articulation in the definition of different types of books and monographs.

[4] https://zenodo.org

The international sources chosen for examination are International Organization for Standardization (2001, 2004, 2005, 2014), those issued by the International Organization for Standardization (ISO), specifically:

- ISO 5127:2001 – *Information and documentation, vocabulary*, with proposed revision ISO/DIS 5127:2014 – *Information and documentation, foundation and vocabulary*. The standard defines agents, objects and processes of scientific communication (thus excluding any aspect of assessment). The aim is to facilitate international communication in the fields of information science and documentation, by providing relevant definitions, terms, and indications of their relationships.
- ISO 12615:2004 – *Bibliographic references and source identifiers for terminology work*. This standard applies to management and use of bibliographic information regarding citations, but is not designed for bibliographic description. One of the goals is to describe the essential identifying elements of an information resource and define the ways in which these should be placed in a citation, thus ensuring timely and unambiguous identification of the original product.

In the Italian context, two major universities (University of Bologna – UniBO; University of Turin – UniTO) have published classification systems for their institutional research products. These systems offer precise definitions of all *books* and *book chapters* categories, taking into account the formal, scientific, content and editorial requirements, as well as those identifying the research results. The following is a brief summary:

- University of Bologna Research Observatory (2013) – The classification system was developed and published as an internal aid to researchers and evaluators. Following an explanatory introduction, the document is divided into a *definition-oriented* section (referring to all individual research products, not limited to publications), and an *evaluation-oriented* section (setting out an internal evaluation system). In fact the UniBO Research Observatory was itself created for development of a unified evaluation system, serving in annual assessment of the university's scientific output. One of the aims was to achieve a balanced response to the characteristics of the technical-scientific and socio-humanistic subject areas.
- University of Turin Research Observatory (2013) – The classification system provides definitions of individual product types, some of which have been revised in successive editions. (The progress of these revisions was retraced for purposes of the current analysis.) In addition to the definitions, the document also provides practical notes for precise identification of the product category.

The UniTO document is purely methodological in approach, and provides relatively little insight on the assessment of specific products. The UniBO document instead proposes a system of five classes,[5] which serve in weighting research products by importance within the landscape of scholarly production. The UniBO system has the further advantage of attributing distinct values to individual products, distinguishing the humanities from technical-scientific areas, and depending on the

[5] The classes are: AA (= 1.5), A (= 1), B (= 0.6), C (= 0.3), D (= 0.05).

territory covered by each product in the landscape of macro-area scientific production. For example the same product could be assigned to a range of different classes, depending on whether the production was in the humanities or sciences.

The design of the UniBO system thus makes it possible to develop a *scalable* indicator of productivity, relating to individual researchers, departments, faculties and subject areas. Given that this model is "wired" to Italian university needs, it could serve as a starting point, integral to a national evaluation system.

The current analysis also examines systems from other national contexts:

- New Zealand (Tertiary Education Commission 2013) – The Performance-Based Research Fund, PBRF, exercise, currently in its third edition, applies only qualitative assessment. The defining documents specify the guidelines to be applied by evaluation panels, in particular the aspects to be accounted for in evaluation and the *weight* (in numerical terms) that each item should have in the overall assessment of the products.
- Australia (Higher Education Research Data Collection, HERDC, Australian Department of Education and Training 2015) – The introduction to the most recent document on research evaluation activities briefly explains the aims, procedures and scheduling of assessment activities. It also provides guidelines for participating actors concerning the selection and evaluation of products and research aspects. For the current analysis, the most interesting material is found in paragraph 6.8 (p. 22–25), which describes the elements characterising each research product (publication).
- Spain (National Association for Education and Peace, Spanish Foundation for Science and Technology; ANEP, FECYT 2007) – The document, on "Criteria for quality in humanities research", provides useful contributions to the definition of quantitative criteria for evaluation of research products. As stated in the introduction (p. 21), the report stems from a combination of three previous documents, on criteria for: (i) quality evaluation of researchers' curricula vitae; (ii) evaluation of research projects; (iii) classification of Spanish journals in the humanities. Each of the resulting sections suggests the weights for the objects subject to evaluation, within an overall rating. The section of most use to the current analysis is paragraph B2 (p. 26–28) of *Criteria for evaluating curricula vitae* (p. 25–29).

The above sources are selected from countries with a culture of evaluation and significant accumulated experience. Such sources are both more general and essential than Italian ones, particularly in providing guidelines on what is included in product categories (for current purposes, *books, book part, conference publications*) and the elements essential for acceptance of a product as *scientific* (Australian Department of Education and Training 2015, p. 19):

> Research publications are books, book chapters, journal articles and/or conference publications. Research publications may be produced in any appropriate format, such as print, publication online, or publication in digital form on separate media such as a CD. A publication is more than the release of a work. It implies quality control (such as peer review or in-house quality control) and enhancement through processes such as assessment or review, editing, copy-editing, design, and conversion of the work to an appropriate format. […] Scholarly editions and scholarly translations must have a major demonstrable original

research component in the edition or translation to be considered for inclusion in a HEP's research publications return. Unless otherwise specified, a reference to publication refers to research publications.

Unlike the systems created for internal use of Italian universities, the systems cited are designed for service in national evaluation exercises, and provide guidelines not only to authors in the identification of eligible products, but also in support of evaluators.

2.2 Framework for Comparison of Classification Systems

2.2.1 Items in the Single Annual Report Form on Departmental Research (SUA-RD form) Considered

In the Italian context, ANVUR (2015b, 2016b) has repeatedly stressed the importance of the SUA-RD form. Therefore, as a preliminary step to comparing the various classification systems, we prepare an analytical table derived from the form. The table considers 17 entries and their relative specifications, pertaining to three of the five main categories identified for the SUA-RD: *books* (9 entries), *contributions to volume* (7 entries), *contributions by conference proceedings* (1 entry):

- Book categories: Monograph or scientific treatise, Concordance, Index, Bibliography, Critical edition of text/excavation, Publication of unpublished sources, Scientific Comment, Book Translation, Edited book. The entry Monograph includes other three sub-entries: Research Monograph, Collection of own essays, Scholarly textbook.
- Contribution to volume: Contribution by volume (Chapter or Essay), Preface/Postface, Short Introduction, Entry in Dictionary/Encyclopaedia, Translation by volume, Book Review by volume, Entry in catalogue/directory/corpora.
- Contribution by conference proceedings: Contribution by conference proceedings.

These items serve as reference categories for our synoptic analysis, with the intention that the results could eventually serve in enriching the current ANVUR classification system.

2.2.2 Structure of the Synoptic Table

The synoptic table serves for comparison of the systems identified in the preceding subsections of the current chapter, also taking into consideration the literature on the subject. It consists of nine columns (Table 2). Each column serves for entry of information deriving from one of the selected classification systems (plus "notes"), as follows:

Table 2 Analytical table

A SUA-RD	B1-B2 ISO	C DEF. UNITO	D DEF. UNIBO	E EVAL. UNIBO	F N. ZEL.	G AUSTR.	H SPAIN	I NOTES.

(a) SUA-RD typology
(b) ISO definitions (B1-B2)
(c) UniTO definitions
(d) UniBO definitions
(e) UniBO evaluation criteria
(f) New Zealand definitions (PBFR)
(g) Australia definitions (HERDC)
(h) Spain definitions
(i) Notes

The columns of the table can be "partitioned", making it adaptable to multiple analytical purposes.

We identify a *defining* group of columns (A, B, C, D, G). Of these, column A contains entries from the SUA-RD classification system, which is the main reference for the analysis. A second group is termed *evaluative*: these columns contain information on the evaluation criteria for the different types of publications, as identified by Italian institutions (column E – UniBO), as well as agencies of other nations (columns F and H – NZ Tertiary Education Commission; Spain ANEP, FECYT).

For improved data analysis (as well as readability), we further disaggregated the table, subdividing the *defining* group into the following three tables:

– Table 1: comparison of SUA-RD product types to ISO definitions;
– Table 2: comparison of SUA-RD product types to definitions in documents of Italian universities and other national agencies;
– Table 3: comparison of SUA-RD product types to indications on product evaluation in Italian university and non-Italian national guidelines.

The following section summarises the results and critical aspects of the analyses deriving from the above tables.

3 Comparative Analysis of Classifications

A first macroscopic observation concerns the comparative presence or lack of typologies among the classification systems, the iteration of the typologies, as well discrepancies between definitions of the same research products and evaluation criteria. The following subsections focus on: comparison between SUA-RD and ISO

classifications (Sect. 3.1); comparative examination of definitions and evaluation criteria under the SUA-RD, Italian university and other national systems (Sect. 3.2).

3.1 Comparing SUA-RD Classification System vs. ISO Standards

Considering that our research aim is to define criteria and guidelines suited to wide application, a first, necessary step is to examine the existing standardised systems with strong international value, such as ISO standards. They are in fact developed through structured processes based on experience from many countries and applications, and, given the important differences between.

The SUA-RD classification system and ISO standards reflect significantly different aims: the SUA-RD system was developed in support of institutional *evaluation*, while ISO standards serve only *explanatory* purposes. This implies less fine-grained classification than seen in the other systems; in particular, the ISO standard lacks many entries seen in the SUA-RD classification. Indeed there are two ISO standards dealing with books, and these were themselves developed for different purposes and users: ISO 12615:2004) is a *bibliographic/citational* standard, with the main users being librarians in general; ISO/DIS 5127:2014 instead deals with objects and processes of scholarly communication, with the main users being scholars and academic librarians.

Our aim in comparing the SUA-RD classification system and ISO standards is to understand the main differences that could influence evaluation process and strategies, rather than simply comparing definitions. Our approach reflects two preliminary considerations: the identification of the "best" definitions would be ambitious, particularly considering that such identification would be strongly influenced by the different disciplinary context, and secondly, scarcely justifiable in case of minor differences among definitions. The analysis therefore concentrates on the presence and absence of particular types of research products, and on the level of detail (fine or coarse-grained) of the classification frameworks.

The SUA-RD group *Contributions to volumes* includes six entries which are totally absent in ISO 12615, although in some cases present in ISO/DIS 5127: *Chapters and substantive sections; Prefaces/appendices of substantive character*, *Catalogue with substantive introduction, Substantive commentary in dictionaries/encyclopaedia*, *Review in volume*, *Conference proceedings subject to peer review*. It is noteworthy that the ISO standards do not cover production in the form of single encyclopaedia entries, although these are clearly conceived as part of the wider work (*Encyclopaedia*), published as a complete object. In the same way, although ISO would not consider a single paper from *Conference proceedings*, it would include the proceedings as a complete book: however the single paper would be better linked to *book chapter,* and could even be autonomous product in digital environment.

Corresponding to our methodological categories of *books and book parts*, the SUA-RD type *Scientific monograph and similar products* includes four publication subtypes: *Research monographs*, *Collections of own research articles*, *Critical editions, Scholarly textbooks*.

The ISO standards indicate the sole class of *Monograph*, since the interest is in the material form of the work, whether in one or more volumes. ISO 5127 also indicates the class *Pamphlet*, which differs from a monograph only in the number of pages (maximum 48). There are no classes for *Collections of own research articles* and *Critical editions*. In the SUA-RD typology, the class of edited books is also seen under *Contributions to volumes*, as "*Editorship of volumes including substantive introductory essay*".

The SUA-RD subtype *Scholarly textbook* can be related to the ISO 5127 specification of *Handbook*, in terms of content type: "reference work presenting basic knowledge on specific subjects". In this case the link is the content, and not the material form of the work, as was the case for the monograph.

Our methodological category of *Research tools* includes SUA-RD types *Systematic indices* (not present in either ISO standard), *Index* (not present in ISO 12615) and *Annotated or critical bibliography*. ISO 5127 notes the *bibliography,* type, however in this case the concept refers to a system of access to a set of documents, while in ISO 12615 the emphasis on semantic description of documents, corresponding with our own conception.

In the class *Critical editions/Translations* are included, while the entries *Critical edition/Publication of unpublished sources*, *Scientific comment* are entirely absent from both ISO standards. *Book Translation* and *Translation by volume* categories are included as general definition of *Translation*.

Beyond the presence or absence of individual categories of research products in the different classification systems, and the specific definitions and their similarity or difference, the analysis allows a macro-observation concerning the degree of subdivision of the schemes: the "granular" detail of the Italian classification is very high compared to the others.

This suggests to deepen – with a special studio – the analysis of the Italian landscape for the Humanities, in order to fully represent this sector of excellence in Italy, systematising such a fine grain classification.

3.2 Comparing SUA-RD vs. Italian University and Other National Systems

The above comparisons of national and international sources reveal the complexity and level of detail of the current Italian classification of research products. Based on the analysis, and considering the evaluation criteria identified in the University of Bologna (UniBO) document, we now propose a conceptual grouping of product types in three sets: primary, secondary and tertiary.

*Primary (*or *fundamental*) *products* refer to publication types that convey highly elaborate and deeply scrutinised scientific results, often the result of long-life study. Especially in the Humanities, these are typically produced by a single scholar. They are generally considered the most important products, essential for gaining academic prestige and career advancement.

Primary products include three publication types: *monographs*, *edited volumes* (including *critical editions, critical editions of excavations*), and *book translations*. Under the UniBO system, these are the only *books* that can be assigned an "AA" rating (1.5 points), in the case of works of "undeniably great international impact". While this consideration could have value, at some point it would be necessary to further define the criteria for such a judgement. Some assistance can be gained by further examination of the UniBO document, where we find that the AA rating can also be applied to articles published in journals with impact factor > = 75° percentile, and the case of SSH articles, only those indexed in the Clarivate/Thomson Reuters Social Sciences Citation Index, for which the journal would be indexed in the Journal Citation Reports (JCR).

Depending on the object and complexity of study, the monographs containing the resulting knowledge could also be published as a series of several volumes. These would only be considered individually in the case of sufficiently discrete themes and editorial approaches.

In some cases a specialist university textbook (see *Tertiary products,* below) could be considered an educational monograph. As well, some humanities monographs intended for a "popular" audience can be virtually indistinguishable from what would be considered a true research monograph).

Translations of books can only be considered as scientific monographs within the field of Linguistics, in the case that the work involves well-founded skills and significant critical-editorial contribution. Simply transferring a work to another language, or translations of own works, cannot be considered as qualifying.

Monographs, edited volumes and book translations may eventually be issued as revised editions or reprints. The first edition of a work is "all copies of a resource produced from substantially the same original input and issued by the same agency or group of agencies or a person" (IFLA 2011, p. 324). A new or revised edition can be considered as a new product only if it has been updated by addition or revision of a substantial amount of content (Australian Department of Education and Training 2015), or at least 30% of the content is different from the original version (University of Turin 2013). A "reprint" is defined as: (i) a new edition reproducing the text of an earlier one as exactly as possible; (ii) a new impression made, or derived from, the same master as an earlier one" (IFLA 2011, p. 337). The reprint is not a new product, since "it does not involve substantial additions or revisions", but it could be seen as an indicator of the importance and success of the original product.

Secondary (or *ordinary*) *products* include the types of products considered as *current production* by socio-humanistic researchers. Relative to a monograph, they address narrower issues or individual aspects of broader topics. Secondary products include the types: *contributions to volumes*, *scientific papers in conference proceed-*

ings, most *translations of books*, and *publications of unpublished sources* (in the case of a single source or a small/partial corpus).

In the case of publications of unpublished sources, the publishing and distribution sphere has significance. Works with only brief critical input would not be considered as true scientific products. It should be noted that publications of original documentary sources are often of interest within a limited regional context, and so only regional or national journals would accept these products for publication.

Secondary products also include more complex and articulated publications, bringing together the work of a number of scholars linked by a unifying research stream. In this case the *editorship* of the work is a type of research product, which deserves more detailed examination in terms of features and role. By "editorship" we mean scientific responsibilities for selecting, coordinating and contributing to the unification of the work by several authors on an overall scientific theme. To achieve an "A" rating under the UniBO system, the editorship must be of an "international" volume and include a substantive introductory essay. If the introduction lacks substantive character, the maximum score is "B", while if the volume is only national in scope, the range of scores is C-B. Editorship without evidence of other scholarly input (introduction, etc.) can achieve only a "D" score.

Under the UniBO system, the potential scores for secondary products range from D to A (0.05 to 1.0 points), and only in the case of multiple contributions by the same author/editor in a single volume could the product reach AA value (University of Bologna 2013).

Tertiary (or *toolkit*) *products* generally receive lesser peer recognition, and serve mainly to complement primary and secondary publications. Some of the types included are prefaces, postfaces, introductions, reviews of books, concordances, systematic indices and bibliographies.

Tertiary products are generally supplements to more important works. In some cases (e.g. introductions, prefaces, postfaces) they can be products by influential scholars in the subject area, sometimes in friendship or superior-inferior relations (e.g. professor-student) with the main author or authors. Others are tools serving in support of a primary or secondary work (indices, systematic indices, annotated bibliographies), not necessarily produced by the main author.

The tertiary type also includes entries in larger works of encyclopaedia or thematic dictionaries. Such products are generally brief summaries or definitions, concerning restricted knowledge within a much broader field.

Collections of research articles (previously published or unpublished) can be produced for a range of reasons, such as: in celebration of prestigious scholar; to offer review of topics that have been dealt with over the years; to bring together an otherwise fragmentary series of texts. Such products can be part of advancing knowledge in the relevant field, therefore the preparation of such volumes can in some cases be considered as editorship (University of Turin 2013).

A last type of tertiary product is the *scholarly textbook*. These works synthesise and unify the knowledge pertaining to a specific field or topic, for reference by students or scholars operating in other disciplinary sectors. However in some cases

they can include advancements in knowledge, given the author's thorough mastery of the subject matter.

The UniBO system does not consider all of the tertiary products just identified, however where they are present, the potential scores range from D to B (0.05 to 0.6 points). Given their nature as supplements to more complex scientific production, and lesser contributions to advancement of knowledge, they receive lesser scores than primary and secondary products.

It should be noted that although in this subsection we have proceeded from the UniBO system for elaboration of the three types of scientific products, the University of Turin (2012) also proposed a classification and weighting system, based on a ratio of "main products" (articles in journals and monographs) to "minor products", with the latter described as "valid but 'off-peak' results of academic research". The UniTO system requires that the individual scientific sectors be carefully analysed, and that the weight assigned to products could then vary by sector (for example, greater weight for critical editions of previous excavations, in Archaeology).

3.3 Concluding Remarks

Currently, the scientific book faces two significant problems: high production costs and restricted commercial markets. Publication of a monograph requires investment in the initial research, analysis and writing, in procedures for ensuring adequate quality (expert committees, peer review, editorial preparation), and in the physical production, marketing and sales. Against this there is only a limited market, consisting mainly of university libraries and individual scholars: both segments that have significantly scaled back their purchases.

Nevertheless, the monograph has not disappeared, and indeed is experiencing a degree of recovery, thanks in part to the recent appearance of open-access monographs (Capaccioni 2014; Darnton 1999, 2009), and in part to the growing need of worldwide scholars to publish books for improvement of their academic status (Cronin and La Barre 2004). The monograph is not at all exclusive to the Humanities, but for this area in particular is considered the "cellulose-based engine of academic life" (Willinsky 2009).

The methodology of comparative analyses conducted in this chapter offers a tool in support of assessment procedures, in particular for definition of the spectrum of significant research products in the SSH communication processes. The methodology applies criteria derived from diverse perspectives, issued by different entities at national and international levels: global ISO standards, Italian academic and national evaluation initiatives, and other European and extra-European national initiatives.

The analytical process revealed the unique character of the Annual Report Form on Departmental Research (SUA-RD form), deriving from the national evaluation system for Italian universities, which has more numerous and precisely specified

product categories relative to those found in other classification systems. The analysis also led to broadening of the set of SUA-RD product types.

Among the types of products considered, the analysis revealed the critical nature of the *monographs* and *edited volumes (with substantial introductive essay,* also *critical editions, critical editions of excavations*) types. Both of these require very high scientific effort on the part of the author/editor, although of different nature. Both types also exhibit borderline cases where it can be difficult to clearly distinguish the defining traits.

The discussion of the monograph raises the question of the university "third mission". Since SSH research deals with subjects closely concerned with modern culture and society, of interest to the population at large, the research products (in our case monographs) can take what would be considered a "popular" more than academic form. Such dissemination activities require mastery of the issues, and scientific and methodological rigour, but also advanced skills in communication. Therefore, instead of the common consideration of works intended for public understanding as degrading the content and form of research products, assessment systems should be designed to avoid this prejudice. Given the trend of research systems towards development of social and economic benefits, popularisation of science has gained importance as necessary to wide communication of scientific insight, and stimulation of critical and constructive debate (EU High Level Expert Group (2015), p. 66–75).

Closely related to this is the question of the *scientific nature* of a research product. During the construction and use of our analytical tables, we were able to reflect on the potential criteria for designating one product as "scientific", and another not. Two readily apparent possibilities were the scientific criteria proposed by the Italian University Council (CUN 2013), or the choice of leaving the authors the responsibility of identifying the scientific nature of their works, depending on the individual context. This would conform with the practice of the SUA-RD form, in which each product can be indicated as either *scientific*, *educational* or *informative*. The preference of the current authors is to allow the individual researcher the facility of deciding the category.

Another particular case is the *catalogue* of a curated exhibition: when published at the local or national level, exhibition catalogues generally receive very little prior review outside of the persons most closely involved in their immediate production (artist, author, publisher) and almost certainly without rigorous peer review process. On the other hand, in the case of the *catalogue* for an internationally curated exhibition, the *introduction* and *preface* are likely to have substantive character, standing in themselves as true essays, and be subject to assessment by others. Yet in general, the role of these products is unclear. In the Italian experience they tend to be simple descriptions of the genesis of the content of the larger, more significant work, without offering a real original contribution, and therefore they cannot be considered as true scientific contributions. This consideration is supported by observation of the classification systems from Australia, New Zealand and Spain, where curatorship is not considered within the assessment exercises.

A further case is the *encyclopaedia entry* which, depending on the disciplinary sector, can vary considerably in status. In law and legal studies, for example, these products are true essays, while in other areas they would consist of short contributions providing little more than the most basic of information.

Assessment systems should also be designed to clearly identify *book reviews* in volumes or journals, often presented by the authors as journal articles or book chapters, without explicit declaration of their nature.

In general, a more detailed classification system permits clear identification of product types, with specific definitions, also gaining *a priori* exclusion of materials not admissible for assessment.

The choice among multiple forms for the presentation of research results descends from the context, goals and structure of the underlying work. The research object, target audience, rate of obsolescence of information conveyed and considerations of timeliness are just some of the elements that can influence the choice of communication channel. In our analysis of product types we see how some of these are minutely defined, reflecting the high value and weight attributed by the community of scholars, while others are only mentioned in a marginal way, suggesting a typically minor role in academic production.

The future of these "minor types" is difficult to predict. On the one hand, certain types of products may gradually decrease or be excluded altogether from scientific production, as scholars increasingly realise their irrelevance from an evaluative point of view. Or in contrast, certain SSH scholars could instead insist on the worth of such products, and increase their efforts to adapt them to the standards required by the assessment procedures. A third hypothesis is that even in the case of exclusion from assessment, products such as book reviews and prefaces would continue to persist in the SSH academic areas, because of their social role within the reference communities. Indeed, it is clear that even the minor production types have a party to play in constructing and consolidating networks and relationships among researchers, not necessarily revealed by coauthorship.

Our analysis demonstrates the need for careful development of evaluation systems, respectful of the specificities of different fields and the relative patterns of knowledge dissemination. At the same time, the designers of evaluation systems must work in the direction of *consistency* in assessment criteria, extending across the humanities, social and hard sciences.

Applying the current methodology to the analysis of classification systems applied in other national and international contexts, academics and decision-makers could make further steps towards a pan-European framework for classification of research products in the SSH, thus continuing to strengthen the strategic directions of the European Research Area.

References

ANEP, FECYT. (2007). *Criterios de calidad en la investigación en humanidades*. http://www.idi.mineco.gob.es/stfls/MICINN/Investigacion/FICHEROS/2007-criterios-hh.pdf. Accessed 4 Sept 2016.

ANVUR. (2011). *Valutazione della qualità della ricerca 2004–2010 (VQR 2004–2010): bando di partecipazione*. http://www.anvur.org/attachments/article/122/bando_vqr_def_07_11.pdf. Accessed 4 Sept 2016.

ANVUR. (2013). *Autovalutazione, valutazione e accreditamento del sistema universitario italiano*. www.anvur.org/attachments/article/26/documento_finale_28_01_13.pdf. Accessed 9 Sept 2016.

ANVUR. (2014). *Linee guida per la compilazione della SUA-RD. Allegato A: Ridefinizione delle tipologie di prodotti della ricerca*. http://www.anvur.org/attachments/article/26/Linee%20Guida%20SUA_RD%20ALLEGATO%20.pdf. Accessed 9 Sept 2016.

ANVUR. (2015a). *Valutazione della qualità della ricerca 2011–2014 (VQR 2011–2014): bando di partecipazione*. http://www.anvur.org/attachments/article/825/Bando%20VQR%202011-2014_secon~.pdf. Accessed 9 Sept 2016.

ANVUR. (2015b). *Funzionalità e utilità della SUA-RD*. http://www.anvur.org/attachments/article/26/SUA-RD20150121_29_1_2015.pdf. Accessed 4 Sept 2016.

ANVUR. (2016a). *Linee guida per la compilazione della Scheda Unica Annuale della Ricerca Dipartimentale 2014, 2015 e 2016.* http://www.anvur.org/attachments/article/26/Linee_GuidaSUA_RD_2014-20~.pdf. Accessed 4 Sept 2016.

ANVUR. (2016b). *Rapporto biennale sullo stato del sistema universitario e della ricerca 2016*. http://www.anvur.it/attachments/article/1045/ANVUR_Rapporto_INTEGRALE_~.pdf. Accessed 4 Sept 2016.

Australian Department of Education and Training. (2015). *Higher education research data collection: Specifications for the collection of 2014 data*. http://docs.education.gov.au/documents/2015-higher-education-research-data-collection-specifications. Accessed 4 Sept 2016.

Basili, C. (2003). La diffusione dei risultati scientifici. In A. M. Scarda (Ed.), *Rapporto sul sistema scientifico e tecnologico in Italia. Elementi per un'analisi* (pp. 191–202). Milano: Angeli.

Bennett, S. (1998). Reprint: Just-in-time scholarly monographs. *The Journal of Electronic Publishing, 4*(1). doi:https://doi.org/10.3998/3336451.0004.103.

Broadus, R. N. (1971). The literature of the social sciences: A survey of citation studies. *International Social Sciences Journal, 23*, 236–243.

Capaccioni, A. (2014). La monografia scientifica e le sfide dell'accesso aperto. *AIB Studi*, *54*, 201–211. doi:https://doi.org/10.2426/aibstudi-10084.

Chodorow, S. (1999). The pace of scholarship, the scholarly career, and the monograph: The once and future monograph. In M. C. Case, & M. M. Case (Eds.), *The specialized scholarly monograph in crisis, or, how can I get tenure if you won't publish my book? Proceedings of a conference sponsored by American Council of Learned Societies*, Association of American University Presses, and Association of Research Libraries. Washington, DC, September 11–12, 1997. Association of Research Libraries, Washington, DC. http://old.arl.org/resources/pubs/specscholmono/Chodorow.shtml. Accessed 10 July 2016.

CIVR. (2007). *Relazione finale VTR 2001–2003*. http://vtr2006.cineca.it/php5/relazione_civr/output/totale.pdf. Accessed 4 Sept 2016.

Cronin, B., & La Barre, K. (2004). Mickey Mouse and Milton: Book publishing in the humanities. *Learned Publishing, 17*, 85–98.

Crossick, G. (2015). *Monographs and open access: A report to HEFCE*. http://www.hefce.ac.uk/pubs/rereports/year/2015/monographs/. Accessed 4 Sept 2016.

Crossick, G. (2016). Monographs and open access. *Insights, 29*, 14–19. doi:https://doi.org/10.1629/uksg.280.

CUN. (2013). *Proposal «Criteri identificanti il carattere scientifico delle pubblicazioni e degli altri prodotti della ricerca»*. https://www.cun.it/uploads/4532/proposta_cun_criteri_scientificit%C3%A0.pdf*?* Accessed 4 Sept 2016.

Darnton, R. (1999). A program for reviving the monograph. *Perspectives on History, 37*. https://www.historians.org/publications-and-directories/perspectives-on-history/march-1999/a-program-for-reviving-the-monograph. Accessed 4 Sept 2016.

Darnton, R. (2009). *The case for books: Past, present, and future*. New York: Public Affairs.

De Santis, I. (ed) (2006). *Linee guida per la valutazione della ricerca*. http://vtr2006.cineca.it/documenti/linee_guida.pdf. Accessed 9 Sept 2016.

de Solla Price, D. J. (1970). Citation measures of hard science, soft science, technology, and non-science. In C. E. Nelson & D. K. Pollock (Eds.), *Communication among scientists and engineers* (pp. 3–22). Lexington: Heath Lexington Books.

Elliott, M. A. (2015). The future of the monograph in the digital era: A report to the Andrew W. Mellon foundation. *The Journal of Electronic Publishing, 18*(4). doi: https://doi.org/10.3998/3336451.0018.407.

EU High Level Expert Group. (ed) (2015). *Commitment and Coherence: Essential ingredients for success in science and innovation: Ex-Post-evaluation of the 7th EU framework programme (2007–2013)*. http://ec.europa.eu/research/evaluations/pdf/fp7_final_evaluation_expert_group_report.pdf. Accessed 4 Sept 2016.

Gimenez-Toledo, E., Mañana-Rodriguez, J., & Engels, T. et al. (2015). The evaluation of scholarly books as research output. Current developments in Europe. In A. A. Salah, Y. Tonta, A. A. Akdag Salah, C. Sugimoto, & U. Al (Eds.), *Proceedings of the 15th International Society for Scientometrics and Informetrics Conference, Istanbul, Turkey, 29th June to 4th July, 2015*. Bogazici University, Instambul, pp. 469–476. http://curis.ku.dk/ws/files/141056396/Giminez_Toledo_etal.pdf.

Giménez-Toledo, E., Mañana-Rodríguez, J., Engels, T., et al. (2016). Taking scholarly books into account: Current developments in five European countries. *Scientometrics, 107*, 685–699.

Hicks, D. (2004). The four literatures of social science. In H. Moed, W. Glanzel, & U. Schmoch (Eds.), *Handbook of quantitative science and technology research* (pp. 473–496). Dordrecht: Kluwer Academic Publisher. http://works.bepress.com/diana_hicks/16/.

Hicks, D. (2006). The dangers of partial bibliometric evaluation in the social sciences. *Economia Politica, 23*(2), 145–162. https://doi.org/10.1428/22461.

Huang, M., & Chang, Y. (2008). Characteristics of research output in social sciences and humanities: From a research evaluation perspective. *Journal of the American Society for Information Science and Technology, 59*(11), 1819–1828. https://doi.org/10.1002/asi.20885.

IFLA. (2011). *ISBD: International standard bibliographic description, consolidated edition*. http://www.ifla.org/files/assets/cataloguing/isbd/isbd-cons_20110321.pdf. Accessed 4 Sept 2016.

International Organization for Standardization. (2001). *ISO 5127: Information and documentation – Vocabulary*. Geneva: ISO.

International Organization for Standardization. (2004). *ISO 12615: Bibliographic references and source identifiers for terminology work*. Geneva: ISO.

International Organization for Standardization. (2005). *ISO 2108: Information and documentation – International standard book number (ISBN)*. Geneva: ISO.

International Organization for Standardization. (2014). *ISO/DIS 5127: Information and documentation – Foundation and vocabulary*. Geneva: ISO.

Joseph, R. P. (2015). Higher education book publishing, from print to digital: A review of the literature. *Publishing Research Quarterly, 31*, 264–274. https://doi.org/10.1007/s12109-015-9429-0.

Ochsner, M., Hug, S. E., & Daniel, H. D. (2016). Humanities scholars' conceptions of research quality. In M. Ochsner, S. E. Hug, & H. D. Daniel (Eds.), *Research assessment in the humanities: Towards criteria and procedures* (pp. 43–69). Cham: Springer Open. https://doi.org/10.1007/978-3-319-29016-4.

OECD. (2002). *Frascati manual: Proposed standard practice for surveys on research and experimental development*. Paris: OECD. http://www.oecd-ilibrary.org/science-and-technology/frascati-manual-2002_9789264199040-en.

Tertiary Education Commission. (2013). *Performance-based research fund: Quality evaluation guidelines 2012*. https://www.tec.govt.nz/Documents/Publications/PBRF-Quality-Evaluation-Guidelines-2012.pdf

Thompson, J. W. (2002). The death of the scholarly monograph in the humanities? Citation patterns in literary scholarship. *Libri, 52*, 121–136.

University of Bologna. (2013). *Definizione e principali criteri di valutazione dei prodotti della ricerca*. http://www.catalogo-ugov.unibo.it/. Accessed 4 Sept 2016.

University of Turin. (2012). *Rapporto sull'attribuzione dei pesi tonali ai prodotti della ricerca (Rev. 04)*. Accessed 4 Sept 2016.

University of Turin. (2013). *Nuove tipologie U-GOV catalogo (rev. 03)*. Accessed 4 Sept 2016.

Williams, P., Stevenson, I., Nicholas, D., et al. (2009). The role and future of the monograph in arts & humanities research. *ASLIB Proceedings, 61*, 67–82.

Willinsky, J. (2009). Toward the design of an open monograph press. *The Journal of Electronic Publishing, 12*(1). http://quod.lib.umich.edu/j/jep/3336451.0012.103?view=text;rgn=main.

Zitt, M., & Bassecoulard, E. (2008). Challenges for scientometric indicators: Data demining, knowledge-flow measurements and diversity issues. *Ethics in Science and Environmental Politics, 8*(1), 49–60. https://doi.org/10.3354/esep00092.

Zuccala, A. A., Verleysen, F. T., Cornacchia, R., & Engels, T. C. E. (2015). Altmetrics for the humanities: Comparing Goodreads reader ratings with citations to history books. *Aslib Journal of Information Management, 67*, 320–336.

Quality Evaluation of Online Library Catalogues, Advanced Discovery Tools and Linked Data Technologies

Maria Teresa Biagetti, Antonella Iacono, and Antonella Trombone

1 Limitations in Using Online Library Catalogues

Monographs constitute the main part of the scientific production in the Human and Social Sciences (SSH). During the assessment of research, it is essential for evaluators to have reliable sources with which to verify the bibliographical data of the monographs presented by scholars who must be evaluated. The aim of this part is to verify the possibility of effective use of online public access catalogues to check the bibliographical data of the monographs published and presented by scholars in the course of the evaluation process.

First, it is necessary to define the features of monographs in the SSH and their treatment in cataloguing. Scientific publications in the Human and Social Sciences exist in great variety, unknown in other scientific fields. The list includes monographs by an individual author or several co-authors, with merged or distinct responsibility; collected works; the publication of unpublished historical sources and of monographs offering a new explanation of sources; indexes and bibliographies; monographs devoted to the description of older editions; collections of unpublished works; sociological essays.

Considering the Italian environment, the document drawn up by the Osservatorio della ricerca dell'Università degli studi di Bologna, "Definizione e principali criteri

M.T. Biagetti (✉)
Sapienza University of Rome, Rome, Italy
e-mail: mariateresa.biagetti@uniroma1.it

A. Iacono
Biella Public Library, Biella, Italy
e-mail: antonella.iacono@fastwebnet.it

A. Trombone
University of Basilicata, Central University Library, Potenza, Italy
e-mail: antonella.trombone@gmail.com

© Springer International Publishing AG 2018

A. Bonaccorsi (ed.), *The Evaluation of Research in Social Sciences and Humanities*, https://doi.org/10.1007/978-3-319-68554-0_8

di valutazione dei prodotti della ricerca" (Bologna, June 2013),[1] is the most useful in order to define the concept of monograph.[2] It clearly defines different kinds of publications in the SSH area and provides the main features and peculiarities for each of them. The document addresses quantitative and qualitative elements. In particular, it underlines that monographs must show a significant extension and must offer a deepened study, characterised by a critical approach. Moreover, it points out that authors must have direct responsibility for the overall content of the work.

The last is a very important point. It involves the necessity of clearly recognising the direct responsibility of authors and co-authors in chapters of books during an evaluation procedure. This point is directly connected with specific problems of library cataloguing. The cataloguing rules provide broad categories in order to establish a standardised way to attribute authorship, but they cannot foresee all borderline cases. Besides this basic difficulty, the personal interpretations of the single cataloguers actively contribute to create an environment corrupted by different cataloguing choices. As a result, bibliographic data are inconsistent.

In addition to these remarks, we can note that in SSH it is not unusual to find books formally attributed to a single author highlighted on the title page, who in reality was the author of the introduction alone, but rather written by several authors who are not formally mentioned on the title page. Moreover, an author may be often presented on the title page as the only person responsible for a book in which are issued or reprinted many works by other authors. The author responsible for the publication selected the works considering a particular scientific point of view, that is, he or she was the *editor* even if he or she claimed the role of author for the whole content. In these cases it is difficult to establish if the responsibility is for the overall work.

Now let us consider the quality and reliability of the data offered by library catalogues. The main purpose of this part of the chapter is to highlight caveats and limitations in the use of online library catalogues.

The assessment of the quality of catalogues can be carried out by considering the following factors:

- Identify authors and disambiguate personal names.
- Distinguish the different editions of the monographs and possible reprints.
- Mistakes or personal view in attributions of authorship.
- Distinguish co-authorship roles.
- Distinguish purchased books from donated books.

---

[1] http://www.catalogougov.unibo.it/_docs/Definizione%20e%20principali%20criteri%20di%20valutazione%20dei%20prodotti%20della%20ricerca_08.pdf

[2] Other notable definitions of monograph have been offered in Italy by the National Agency for the evaluation of University and research (ANVUR): VQR 2004–2010 http://www.anvur.org/attachments/article/122/vqr_d.m._n._17_del_15_07_2011_firmato.pdf; VQR 2011–2014 http://www.anvur.org/attachments/article/825/Bando%20VQR%202011-2014_secon~.pdf and SUA-RD for university research http://www.anvur.org/attachments/article/26/Linee%20Guida%20SUA_RD%20ALLEGATO%20A.pdf

1.1 The Identification and the Disambiguation of Authors

To identify authors and disambiguate those having the same names, a reliable tool for authority control is critical. The *Functional Requirements for Authority Data: a Conceptual Model* (IFLA 2009a) is the conceptual model for authority data.[3] The model defines the requirements that authority data should meet to satisfy the need of consistency and the necessities of users. As an example of *Authority file*, let us consider the *Authorities* service offered by the Library of Congress,[4] a tool that enables cataloguers to use uniform accesses to names, titles, subjects, and to disambiguate authors of the same name. It is a tool for use by technical services and an authority source of bibliographical information for users supplied on a separate website available free of charge. The Library of Congress authority records adopt MARC 21 *Format for Authority Data*.[5]

The Library of Congress developed its *Authority file* in collaboration with several libraries that participated in a cooperative program, *Name Authority Cooperative Project* (NACO), a project for authority control on both theoretical and practical levels. The libraries that participated in NACO must follow the standards established to create uniform accesses (Byrum Jr. 2003; Ilik 2015).

As regards Italy, Casalini Press, the most important seller of Italian books abroad, participated in the *Program for Cooperative Cataloging*, the largest project of participated cataloguing with bibliographic control, launched in 1995.[6] After its participation in the *Shelf-ready Project*,[7] authority data provided by Casalini were made available to participants in the *Program for Cooperative Cataloging*. Casalini Press is the provider of the authority records for Italian editorial works.

WorldCat[8] is the cooperative union catalogue developed and managed since 1996 by Online Computer Library Center (OCLC). WorldCat is the result of the control and editing of records produced by tens of thousands of libraries, OCLC members (OCLC Annual Report 2014–2015). Due to its size, WorldCat offers a great opportunity to find bibliographic information. However, its main limit is its lack of disambiguation of authors of the same name. Without a suitable authority file, authors' identification is impossible and data retrieval is imprecise. WorldCat

[3] http://www.ifla.org/files/assets/cataloguing/frad/frad_2013.pdf

[4] https://catalog.loc.gov/

[5] http://www.loc.gov/marc/authority/ecadhome.html

[6] http://www.loc.gov/catdir/pcc/ Starting from 2005, Casalini contributes to both *Authority control* projects, descriptive and semantic: *Name Authority Cooperative Program* (NACO), which starts in 1977, and *Subject Authority Cooperative Program* (SACO), launched in 1993.

[7] The *Shelf-ready Project* prescribed the delivery of books along with catalographic records, Table of Contents and the LoC control number. Since 2006, Casalini has annually furnished the Library of Congress with about 4000 *shelf-ready* books, almost half of the acquired Italian publications, ready to be put on the shelves, already processed and provided with an anti-theft device. (Genetasio and Terravecchia 2009).

[8] www.worldcat.org, online since 2006. WorldCat offers more than 347 million records, representing 2.3 billion items owned by libraries, in 480 languages and dialects.

Identities,[9] in fact, is the simple list of headings of authors and subjects, not completely disambiguated. Moreover, author searching is allowed using single terms of headings, which involves the retrieval of all the records in which the term is present.

Focusing now on the Italian environment, it is worth mentioning that the Servizio Bibliotecario Nazionale (SBN) managed by ICCU[10] since 2001 has assured uniform access for personal authors, corporate bodies, and uniform titles thanks to a reliable authority file, organised according to the *Guidelines for authority records and references* (GARR)[11] and UNIMARC/A.

The *authority file* made by ICCU[12] is therefore the essential tool to improve quality and retrieval of information present in the SBN collective catalogue and is developed in cooperation with the SBN's partners, either through normal cataloguing activities or through the more specialised work of defining the authority entries. The methodological guidelines and rules for implementing the authors' archive using a controlled and uniform procedure are provided in the document recently published by ICCU: Linee guida (2015). The OPAC of SBN provides a reliable and authoritative tool of *authority control*, with disambiguation through the date of birth and creation of authoritative cards for one part of the personal authors, often supplemented with biographical data. All authority entries can be consulted in the SBN OPAC. SBN contributes with its authority records (about 50,000 authority records relative to personal authors) to the Virtual International Authority File (VIAF),[13] the large OCLC collaborative project that collects the authority records of 30 countries.

In Italy there is another online tool with public access that makes it possible to verify monographs.[14] The *I libri* database created by Casalini Press[15] is a bibliographic database that offers a good level of coverage of commercial publishers and of academic publishing from 1985 to date, and includes updated data. It also shows bibliographic advertisements of publications that are in the publishers' programme,

[9] http://worldcat.org/identities/

[10] http://www.iccu.sbn.it/opencms/opencms/it/main/attivita/naz/pagina_335.html. The Central Institute for the Union Catalogue of Italian Libraries and Bibliographic Information (ICCU) is a central institute coordinated by the General Department for libraries and cultural institutes of the Italian ministry for cultural heritage and tourism (MIBACT). It was established through Decree of the president of the republic no. 805 dated 3 December 1975, assuming the functions of the National Centre for the Union Catalogue and bibliographical information (established in 1951). ICCU coordinates and manages SBN and national and international cataloguing projects, also using new technologies. http://www.iccu.sbn.it/opencms/opencms/it/

[11] http://www.ifla.org/VII/s13/garr/garr.pdf

[12] http://www.iccu.sbn.it/opencms/opencms/it/main/attivita/naz/pagina_335.html

[13] http://viaf.org

[14] It is to be noted that, by contrast, the Bibliografia Nazionale Italiana, elaborated by the Florence national central library on the basis of legal deposit by law, has been online for public access only since 30 May 2016, limited to years 2012–2014. Previously, it was accessible by purchasing a license for viewing.

[15] http://www.casalini.it/ilibri/index.asp.

with preliminary information, but it does not declare that its state is not yet definitive. Data are updated at the moment of effective publication of the monograph. It does not seem to have a robust system of *authority control*. However, even though it does not systematically disambiguate authors of the same name, it often provides indications of the author's university of affiliation in the notes field.[16]

1.2 Different Editions and Reprints

As regards identification of different editions of monographs, for the purpose of building an effective information source useful for evaluators in the activity of scientific evaluation, the catalogue must provide an easy and intuitive way for identification of different editions of a monograph, of different authorial roles and relationships among publications. Yet, still today, catalogues do not efficiently carry out this important function, with an evident loss in quality when retrieving bibliographic information.

In fact, research interfaces of online catalogues do not adequately apply the model outlined in *Functional requirements for bibliographic records* (FRBR). This study by IFLA, (IFLA 1998; IFLA, FRBR Final Report 2009b), which played an important role in the development and renewal of cataloguing theory in the past 15 years, has the merit of modelling the bibliographic universe, clearly distinguishing between different entities, ideal (works and their expressions) and material (editions and copies).

The catalogue organised according to the FRBR model should account for the identity of a publication considering its various titles, differences of its editions (manifestations), of the relationships between publications (e.g., between derivative works and works whose subject is constituted by other publications). In most cases it concerns a simple FRBR-*isation*, a procedure to realise groups based on the *works*, using specific algorithms. These groups are realised *a posteriori*, without restructuring the original records.

Considering the Italian environment and the SBN OPAC, it is to be noted that it can display the records of the different editions held by the libraries, which cooperate in the realisation of the national online catalogue. However, the Italian online catalogue also presents some critical elements. In addition to the records describing editions, at times we can have records describing reprints. A specific procedure makes this possible. The *Guida alla catalogazione SBN*, published in 1995 (ICCU 1995, p. 50–51) indicates to identify reprints as occurrences in the central index without creating a new record. Nevertheless, the *Guida* allows the cataloguer to create a new record for a reprint when one is not able to find the data of the edition in the index. On the other hand, the Casalini Press *I libri* database does not record

[16] The mentioned features refer to years 2014–2016. Starting from the year 2017, Casalini Press changed the layout of the database.

reprints; this is one important and appreciable element, very important when checking authors' monographs for scientific evaluations.

It is worth mentioning here that the new Italian cataloguing rules, REICAT (ICCU 2009), comply with FRBR, and that, starting from the moment of their application in Italy (2010), it will be able to improve the identification of works, expressions and manifestations. Despite the fact that the FRBR model was not applied in the Italian records before the publication of the Italian REICAT, and the different editions were not precisely differentiated, Italian public catalogues online can in any case be used to discriminate new editions from the unchanged reprints. They can also be used to identify false new editions, even when the title page declares it as a new edition. Among bibliographic data in records, it is possible to verify the ISBN number (if present), which changes with every new edition and is the same in case of reprints. One can also verify the pagination and the change of formats. One can check this information in the material description area, along with the presence of illustrations and accompanying material not present in the previous edition.

The procedure to verify bibliographic data can be automated through the development of an *ad hoc* software. The system, for instance, could take the authors' name and titles of works as input and acquire from the SBN OPAC the UNIMARC records, and in particular the following fields: 200, 205, 210, 215, 410, 700, 010$a. These fields concern the title and responsibility block, the place and date of publication, the name of the publisher, the material description data, the editorial series, the author's access and the ISBN number. The next step might consist in analysing the correspondence between the 10-digit ISBN format and the 13-digit format using the ISBN converter[17] tool. In the conversion from the 10-digit format to 13-digit, in fact, the last digit, a check-digit, is replaced using an algorithm. Thanks to this comparison, apparently different ISBNs can turn out to be simply ISBNs converted into the new format and therefore identify a reprint of the same edition and not a new edition.

1.3 Mistakes in Authorial Attributions

With regard to the verification of mistakes in attributions to authors and entities, we can remark that in the OPAC of the Italian Servizio Bibliotecario Nazionale, in particular in the case of publications devoted to the reconstruction of ancient book collections and the identification of editions, generally carried out on the basis of handwritten inventories, it frequently happened that cataloguers attributed authorial responsibility to the library that owned the book collection and not to the author who reconstructed it, and identified, described and indexed editions.

The Casalini Press *I libri* database, however, presents bibliographic information gathered from title pages without the cataloguers' interpretation. This places the Casalini's database in a particular condition. On the one hand, it avoids the always

[17] http://www.isbn.org/ISBN_converter

possible error of authorial responsibility and the wrong attribution of principal responsibility to an entity on the basis of the personal cataloguer's interpretation. On the other hand, the loss of the cataloguer's working according to specific cataloguing rules, REICAT in Italy, prevents roles and responsibilities from being correctly redefined, if these have been mistakenly reported on the title pages intentionally, e.g., in the case in which an editor, or a translator, is explicitly presented as author.

1.4 Co-author Roles

As regards the field of bibliographic description, both in the national standard (REICAT) and in the international ones (RDA, ISBD) the monograph is defined as a resource in only one part complete, or subject to be completed in a defined number of distinct parts (ICCU 2009; JSC RDA 2010; IFLA 2011).

The activity of bibliographic description assigns the authorial responsibility to publications on the basis of specific rules, defined by national and international standards mostly considering the main part of the publication and the mode of formal presentation of responsibilities in the primary information sources. Thus, it can happen, for example, that an author presented with absolute prominence on the title page could be given the paternity of a publication of which he has in reality written only a brief introductory part consisting in a modest number of pages.

With regard to co-authors, under the national REICAT rules or RDA indications examinations are restricted to the formal presentation of responsibilities on the prescribed information source. In neither case is it therefore possible to detect the responsibility of each cooperating authors, each one of them having written a part, or the responsibility for only one of the volumes published under the responsibility of all authors present in the sources. It is impossible to identify authorial responsibilities in the case of co-authors presented as being at the same level in the publication, and each one of them having written one part, declaring, e.g., personal responsibility inside the publication. Responsibilities that are attributed inside the publication cannot be detected through cataloguing data.

1.5 Acquisition of Books and Gifts

One of the factors relevant to the purpose of assessing catalogue quality is constituted by the possibility to verify, using online public catalogues, the responsibility of libraries in buying monographs and in accepting donations. Oftentimes university teachers and researchers donate a copy of their publications to their department library, as well as to the libraries of other Italian or foreign universities that own collections congruent with the topics treated in the volumes, and also publishing

companies often donate a copy of a book published by them to the library where it is presented.

It is important to take into account that many libraries are equipped with regulations that govern the acceptance and management of donations. However, the acceptance of book donations entails the assessment of consistency of gifts with topics and scientific level of the recipient library. This is very different from the decision to buy a book, to spend money, especially at a moment in which public budgets are particularly reduced. Such purchases are indicative of the will of the library, reflecting the library's guidelines for the development of its collection, to offer its users that specific book as it is considered significant for them. It is to be noted, moreover, that university libraries, in particular, adopt automatic procedures of purchase and return of books – *approval plans* – in agreement with publishers. Libraries decide the criteria to select books, often using the classes from DDC or LC, in accordance with their scientific sphere. However, vendors and providers choose single items, and the returns may not exceed 5% of the books sent (but this can vary, according to the agreements) (Nardini 2003; Morriello 2006).

To conclude, it is worthwhile to consider that in online public access catalogues information about the type of acquisition, gift or legal deposit by law are not shown, as this information is not considered relevant to users. However, MARC formats permit the inclusion of information about the type of acquisition, and the Library of Congress OPAC, for instance, in the 925 and 955 MARC fields shows step-by-step information about the book, from the acquisition to the cataloguing procedures. Besides, UNIMARC (IFLA *UNIMARC* 2007) and MARC21 formats permit the inclusion of information about the type of acquisition, by gift, bequest, loan, purchase, or deposit.[18]

The Italian SBN OPAC can display the records in UNIMARC and MARC21 formats, but in both cases it does not enable users to check information regarding the type of acquisitions.

After shedding some light on the advantages and drawbacks of using online library catalogues, we can say that the Italian SBN OPAC offers the best coverage for Italian libraries and a good level of accuracy of bibliographic data to verify the information of monographs published in Italy. Moreover, it presents a rather good authority control and can count on the participative cataloguing of thousands of libraries of a good level. It can be proposed as the most convenient tool to verify the bibliographic data of Italian monographs when performing assessment. It can be used along with the bibliographical database produced by Casalini Press, which is useful to control the edition of works when it does not register the reprints.

[18] http://www.loc.gov/marc/holdings/hd541.html

2 Discovery Tools: Hybrid Research Tools Based on the Web

Planning an analysis of data based on search tools available through libraries means necessarily taking into account the profound changes that have occurred – and are still in progress – in the search interfaces of electronic catalogues. Moreover, it is deemed important to highlight the increasing spread of research tools defined as "web scale discovery tools" that coexist with traditional online library catalogues. The analysis here proposed is aimed at examining the nature and operation of these research tools in order to understand if the new discovery tools, together with the OPAC or in their place, especially in university libraries, can affect the qualitative and quantitative outcomes of research into monographs.

Since the late 1990s, on the heels of the diffusion of electronic catalogues and OPAC, we have seen constant innovation in the online search interfaces of catalogues. The difficulties in using these tools in the beginning were mainly due to the need to structure research using exact terms or keywords, also putting them in the correct fields (Borgman 1996; Marchitelli and Frigimelica 2012). These interfaces are suitable for the "known item search" function, which implies the knowledge of at least one of the basic bibliographic elements – the author or the title of the publication – if not all the bibliographic details of the specific resource. Less expert users might have needed help from the librarian to formulate research in terms relevant to the system, just as for researches in specialised databases during the same period.[19]

Initially, the online catalogues sought to simplify research techniques and also allow the user to carry out an exploratory search in the catalogue using natural languages and allowing greater discovery of the publications possessed by a library. A primitive form of the discovery function, realised through catalogues, was offered by the "find" research function. This possibility was an alternative to the scrolling of lists of indexes and, combined with the Boolean operators, allowed the search for terms within all of the catalogue's description fields.[20] Alongside these innovations, the use of search engines as a source of information caused them to become the model for the elaboration of the next generation of catalogue search interfaces. These latter, conventionally defined "next generation catalogues", offer the possibility of simultaneously interrogating all the databases the library has access to, including the catalogue. These research systems interrogate the entire bibliographic collection of a library, including the institutional archives, the collections of online resources and the specialised databases subscribed to by the library (Christensen 2013; Nagy 2011; NISO 2013; Marchitelli 2015). The newest discovery tools add to the federated search system[21] consisting of next generation catalogues, in local and remote databases, a centralised index of scientific contents, a package of online

[19] Regarding search methods see Sweet (2012), Petrucciani (1984), p. 1–12.

[20] On the loss of professional and scientific identity of librarian instruments, see Petrucciani (2006).

[21] This expression means a simultaneous search in multiple databases, in this regard see Wang, Mi (2012).

resources to which the library has access through an indirect subscription included in the discovery system's licence.

The discovery tools are technically defined "web scale discovery systems" (Way 2010; Raieli 2015) and allow, therefore, the querying of the different resources that the library makes available to users through a single search box. The three distinctive components of a discovery tool are: the search interface, the central index and the local index (Gong 2012; Breeding 2015). The interface recalls that of search engines, it is navigable without the need for special instructions, starting from a unique Google like string, or with the possibility to articulate more search keys in an advanced search box. Search results are then presented in a short or complete format.

The sorting of results by relevance is a feature common to the various discovery systems and is accompanied by other ordering criteria, which, however, vary depending on the software chosen by the library. The algorithms that allow sorting by relevance of search results are the property of discovery systems and they are not always made public. Due to the huge amount of resources available in the various indexes interrogated by a discovery, including its internal index, sorting by relevance should enable the results that best match the search criteria to be placed at the top of the list, or that the most interesting and important resources appear first. The factors that determine the relevance of a publication can have a library science origin, therefore can be connected to the correspondence between the search terms and the results coming from the catalogue. Moreover, they can be influenced by the data of the library loans. Even user habits, the frequency of access and the number of connections to electronic resources identify the search paths and help establish rankings based on user behaviours that influence the discovery relevance criterion (Biagetti 2010; Breeding 2013). Consequently, for the purpose of an conscious organisation of search results, the other sorting criteria offered by the discovery tools besides relevance take on a fundamental importance. The possibility of sorting results by author or title, as well as by date, allows the user to get a list that is verifiable through objective criteria. In this regard it is necessary to point out that some producers of discovery systems are also suppliers of electronic publishing, and that without clear, objective and verifiable sorting criteria they might be believed to be favouring their own content through the discovery ranking (Kelley 2012).

Another feature of the discovery web interface is the use of facets and limiters to restrict or define the search results. The categories of terms used as facets are derived from cataloguing fields, from metadata and document formats.

The second part and distinctive element of the discovery tool is the central index, whose contents are not part of the library collections but are connected to the license of use, owned by the company that produces it. The central index contains metadata and full text resources that are the result of agreements with commercial publishers, to which are also added metadata and open access publications contained in research repositories.

Local indexes of discovery tools, the third part of which they are composed, contain very different documents that are located in separate databases different from the catalogue, such as digital or digitised library collections and the research institutional repository. These discovery systems allow, in fact, to index and also search specialised metadata formats like those of archives and museum materials.

The discovery tools are, therefore, index-based systems: the contents of all databases, local and remote, are re-indexed by the discovery system, including those subscribed to by the library with commercial agreements and the catalogue data. In theory, during the indexing process the system should treat all contents equally. In actuality, it is unclear how the contents are indexed and therefore retrieved in the research phase. There are no standards regulating this process, which remains, along with the relevance criteria, completely out of the librarians' control (Breeding 2015). Consequently, the results displayed by the interface of discovery tools are a secondary source, as they are the result of an activity of re-indexing of cataloguing data and of online resources' metadata searched by the discovery tool. To view the original descriptions, a connection must be made to the data source, always through the discovery interface, so to the catalogue or to the publisher's site, for example.

The problems that can be caused by the re-indexing method also derive from variations in the quality and completeness of metadata that these discovery systems receive from different sources. Metadata can vary in terms of quality and quantity (Somerville 2013) starting from the origin.

From what has been said so far, we can deduce that, in view of the possibility of simultaneous search in multiple, heterogeneous data silos, including proprietary indices of discovery tools, the critical issues detectable in these systems may be due to the lack of clarity about sorting criteria by relevance of search results. Another critical point is the re-indexing of all metadata originally attributed to resources and publications, an issue further complicated by the extreme variability and lack of clarity about the content of central indexes of discovery tools. The resources searchable through these systems are multiple, but, according to the considerations above it is clear that the accuracy cannot be the same as the precision achievable by querying the search interfaces of the original individual silos connected to the discovery tool. The discovery system offers less accuracy even compared to online catalogues, and a known item search carried out for a resource whose bibliographic data are partly known can be problematic. Similarly, researchers may prefer to query the native indexes of specialised databases directly because they need accurate answers, which do not require an exploratory research (Breeding 2015; Ellero 2013; Han 2012; Frederiksen 2015).

2.1 Checking the Availability of Monographs Through Discovery Tools

The effectiveness of discovery tools' search algorithms could encourage their application for quantitative and qualitative analysis of the presence of monographs in libraries. The limits currently recognised for such an application are the following:

- The re-indexing criteria of the resources implemented by the discovery producers are not made publicly available.
- The queried resources are heterogeneous and subject to ongoing negotiations, so the catalogue is only one of them.

- In almost all cases, the algorithms organising the search results cannot be negotiated with software houses. Therefore you can choose to query only the catalogues linked to the discovery tool, but the sorting order of the results is almost always not comparable to that of online library catalogues.

Moreover, it is necessary to examine carefully the websites that contain library search instruments, since many online catalogues have already replaced their OPAC (Online Public Access Catalogue), namely the search interface of the catalogue, with the discovery instrument adopted. If such a complete replacement has not occurred, there will be a coexistence of two search tools: on the one side the online library catalogue and on the other the discovery tool. The first stage of analysis must therefore include a distinction between websites wherein the catalogue and discovery tool coexist and those where the discovery tool completely replaces the catalogue search interface.

A second stage of analysis has to consider the search interface of the discovery tool, and, in particular, the sorting criteria of search results that it allows, and the requirements of facets and limiters applicable to search results. To search library catalogue data using discovery tools there must necessarily be a function that limits the query only to the catalogue. Moreover, success in terms of accuracy of this kind of search is linked to the type of classification of the results, or to the possibility of organisation of bibliographic records in short format that the discovery provides, beyond the relevance criterion. The possibility of organising the results through criteria characteristic of cataloguing indexing, in particular by author, title and date, allows data retrieval suitable for an activity of qualitative and quantitative verification of the presence of publications in a library.

For the purposes of verification to be carried out for our research project, a sample of monographs of 20% of the constructed database was selected for both Italian scientific areas surveyed.[22] Among the catalogues of libraries selected to test the presence of the monographs, for the purposes of this project the ones also queried through a discovery tool were chosen. The subsequent step was to choose two academic library systems as standard, an Italian one and a foreign one, which had not adopted the same discovery system.

Given these guidelines, the choice was the following: the library system of "Sapienza" University of Rome and the Oxford University Libraries.

The research therefore analysed only a sample of the monographs of both databases and the results are directly comparable with the searches for the same titles in their respective online catalogues.

The Oxford library discovery tool, SOLO (Search Oxford Libraries Online), offers two search options: "all libraries/collections" for a search in the local index of the discovery tool; "search everything" to search its local and central index.

[22] The survey was sponsored by the Italian Agency for the evaluation of the University and the research (ANVUR), and carried out in 2014 and 2015. See Section 4, Chapter 14 *Is the diffusion of books in library holdings a reliable indicator in research assessment?*, in particular Part 2.

Wanting to query only the catalogues of the Oxford libraries, the limiter "all libraries/collections" can always be selected before entering the search terms, namely the author and title of the monograph. In SOLO it is possible to select the following sorting criteria for the list of short results: relevance, date-newest, date-oldest, author, title, popularity.

The uniformly applied criteria for checking the presence of the monographs in SOLO (Search Oxford Libraries Online) are as follows:

- Search in "all libraries/collections".
- Search string composed by the author and the title of the monograph.
- Sorting of the list of short results by title or by author.

To perform a search in Discovery Sapienza it is possible to select the following sorting criteria of the short results: relevance, date-newest and date-oldest. The uniformly applied criteria to check the presence of the monographs in Discovery Sapienza are the following:

- Search using a keyword.
- Search string composed of the author and the title of the monograph.
- Use of the limiter "Available in library collections".
- Sorting of the list of short results by relevance, since the only alternative is a chronological order, ascending or descending.

The results recorded in the database show a perfect match of searches carried out first in the online catalogue and next in the discovery tool of Oxford University libraries: the search of the monographs selected as a sample provides, therefore, the same results in the catalogue and in the discovery tool.

The situation is different for the search done in "Discovery Sapienza": for monographs selected as a sample, and for both scientific areas, the searches in the online catalogue differ from those performed in the discovery system. The percentage of presence of monographs in the discovery tool amounts to about 30% less than the presence of these monographs in the online catalogue, for both scientific areas. However, removing the limiter "Available in library collections" when searching the monographs falling into cases described above, we find in 100% of cases the presence of a digital copy of the printed publications, resident in another local index or in central index owned by the discovery tool. In none of cases the digital copy is accompanied by the bibliographic record of the printed version, actually present in the online catalogue of the Sapienza libraries.

The case presented demonstrates the differences of use between the selected search systems and the differences in the results obtained in the case of a known item search carried out with the precise input of keywords, aimed to verify the presence of bibliographic resources in a database, not for an exploratory bibliographic research. In the case of "Discovery Sapienza", it is presumed that the presence in the indexes of a digital copy of the requested monograph prevails in the display settings over the presence of the same monograph in print format, partially hiding its presence in our search results.

2.2 Conclusions

The considered cases show that, with regard to the University of Oxford, there is a perfect match between the search done using the discovery tool and the search done through the library's online catalogue. As far as the results of Sapienza University library are concerned, there is a certain percentage of misalignment. Such a situation could depend on the settings adopted for the attribution of the relevance criteria to publications, for example. Actually, in the absence of selection criteria, or standard guidelines for the implementation of such discovery instruments, it can be stated that at present they are not suitable instruments for a quantitative analysis of the publications. They allow a federated search by querying all databases accessible through the library; they also make possible an exploratory research of resources indexed by the discovery tool but not owned by the library. However, searches through the online catalogue are to be preferred for quantitative and qualitative assessments, both for the clarity of the structure of the data and the comprehensibility of the document retrieval techniques, entirely managed in the field of Library and Information Science.

3 Application Perspectives of Linked Open Data in Research Assessment

At the end of this contribution, we felt it was appropriate to give an overview of the new opportunities that the application of *linked open data* to library catalogues could offer to scientific evaluation. In fact, the adoption of this technology to online library catalogues opens new and exciting perspectives that may also be of interest to this sector of inquiry.

First, we need to clarify what is meant by *linked data* and how this new approach can be used to enhance searches in online library catalogues.

As is known, Linked data, in the definition provided by the founder of the W3C[23] Tim Berners Lee, refers to a set of recommended best practices for publishing and connecting structured data on the Web, favouring the creation of a global information network whose contents are mainly exchanged and interpreted by machines, forming the basis for the realisation of the Semantic Web (Berners-Lee 2006).

Linked data can best be interpreted as a major paradigm shift in the way of understanding data, even of a bibliographic nature. The adoption of this technology offers the possibility to create "data" or structured information that is connected, interoperable and integrated with any other information found on the Web. It offers the possibility to integrate and connect the knowledge of the network in a global network of data connected to each other (the so-called "linked data cloud").

[23] Tim Berners-Lee invented the World Wide Web in 1989. He founded and directs the World Wide Consortium (W3C), the forum for technical development of the Web.

In the field of library OPACs this is a turning point: for years, libraries have entrusted their data to closed and exclusively *library-centric* bibliographic formats like MAchine Readable Cataloguing (MARC),[24] the main format used to store and exchange data for over 40 years. The adoption of this format has been identified by many scholars as the main cause of the slow evolution of library OPACs, since it directed their development towards solutions that have in fact prevented the use, exchange and discovery of bibliographic information on the Web (Yee 2009).

For decades bibliographic information was trapped in library OPACs and in various bibliographical archives, considered to be non-communicating "silos" (Naun 2010).

The most important paradigm shift that affects libraries is certainly that of the turn "from records to data". Creating "connected" bibliographic data in the new form of *linked data* means, in fact, first adopting the Resource description framework (RDF) as a new data model or a new logical model to express bibliographic data, modifying the concept of record as it has traditionally been conceived. The RDF model "breaks up" the information into "statements" or "triples" linking the data through qualified relationships.

In the Semantic Web it is expected that bibliographic data would adopt a form characterised by an increased granularity. As is the case for the Web, where the adoption of the LOD implies the turn from a network of HTML documents connected by links and designed as unique blocks of information to a new Web of linked data, in the field of libraries embracing the logic of linked data means building simple relationships between the elements of a bibliographic record and not between individual bibliographic records.

While the bibliographic record is currently formed by an indivisible block of bibliographic data, i.e. the traditional textual and monolithic record, with the application of linked data the record is deconstructed and broken down into a set of triples linked together and connected to other data on the Web.

This need for granularity of the bibliographic data is not new. With the formulation of the FRBR model (IFLA 1998) an irreversible process towards the conception of a more granular bibliographic record was begun. In this direction, the Library of Congress report *On The Record* (Library of Congress Working Group on the Future of Bibliographic Control 2008) was a turning point, and today an effective end point is represented by the drafting of a new BIBFRAME model (US Library of Congress and Zepheira 2016), currently under development, which should definitely replace the old MARC formats (Kroeger 2013).

This transition towards a new format of bibliographic data exchange will make the data fully compatible with the Web. The adoption of BIBFRAME is aimed to replacing MARC21 with a format fully suited to the Web, where there are both FRBR model entities, but also semantic annotations of various types and sources. Without going into the details of the new record structure predicted by BIBFRAME, it should be noted that this model aims to create a "bibliographic environment"

[24] https://www.loc.gov/marc/

(Miller et al. 2012)[25] that is built into the Web and may also contain information added by non-bibliographic sources.

Along with BIBFRAME as the primary means to encode a new widespread bibliographical environment in the Semantic Web, within the context of cataloguing theory the new RDA cataloguing rules (Resource Description and Access) published in 2010 created a new standard for access and description of information resources, "specifically designed for the digital world".

The cataloguing theory is oriented in this new direction: recent RDA (Joint Steering Committee for Development of RDA eds. 2010) cataloguing standards, thanks to the fruitful cooperation being developed with the Semantic Web community and following the logic of *linked data*, introduced in cataloguing a new scenario where every bibliographic record is formed by the extraction and orderly recomposition of "data" coming from various sources and made up by archives of names, works, expressions, events, places, concepts, etc. (Coyle 2010).

The new rules are clearly inspired by the Semantic Web, stating the need that bibliographic data be geared more closely to the data (being more "data-friendly"). As is known, the RDA is based on two fundamental objectives of identifying and connecting the resources. These goals come directly from user functions established in the FRBR/FRAD (IFLA 2009a) and ICP (IFLA 2008, 2009a, b) and are reflected in the articulation of the standards content. The growing role of authority records in the resource description clearly emerges in the new code.

The true new development of the RDA, as is clear from reading the guidelines, is its primary focus on content. It is concerned with the registration of the entity attributes and the recording of the relationships between the entities, i.e., the choice of the size of the catalogue and their attributes, and no longer the display or presentation of the elements. In this way RDA marks a sharp break from all previous codes, almost a Copernican revolution: from records management to entity management. We no longer produce a catalogue consisting of records, but rather define individual data formulated with terms extracted from ontologies and vocabularies on the Web.

The process of "identification" of the bibliographic data is conceived as a process in which each bibliographic authority data is accurately and uniquely identified and linked with other data. This systematic process of "precise identification of a resource on the Web' via a URI (Uniform Resource Identifier) allows you to build "dynamic links" to the "Web of data" that from the bibliographic data easily reach other types of data.

3.1 The Benefit for Scientific Research Evaluation

The evaluation of scientific research can greatly benefit from the new perspectives opened up by the application of linked data to bibliographic data and with the new cataloguing scenario we briefly mentioned. In this context it is clear that, in the

[25] http://bibframe.org/

research work of testing the availability of an author's monographs in library catalogues, the use of systems that adopt this technology will have a strong impact in some key aspects that here we can only briefly list:

3.1.1 Authority Control

The authority control is exercised to accurately identify bibliographic entities (a person, a corporation, a work), through a set of useful information in order to avoid the formal inconsistences of the catalogue entries (IFLA 2009a).

The control of authority data is a crucial element in the evaluation of scientific research, and allows the catalogue user to clearly and precisely identify each entity of interest (the author and his/her publication). A critical aspect of bibliographic data created and currently present in the online catalogues is certainly the inaccurate management of identifiers of the entities registered in the bibliographic records.

In these archives, entities' identifiers (e.g., an author) are created and maintained locally and can hardly be of any use to the library software, which normally does not handle them but simply records their presence. In the future – with the application of RDA and Linked Open Data – it will be possible to use external repositories to identify these data and "qualify" the relationships between them, using attributes defined in specific domain ontologies (e.g., FRBR, RDA etc.).

In the context of bibliographic control, this process will become increasingly important and widespread.

RDA puts into large account the authority data, obtained by special controlled vocabularies, with a high presence of open vocabularies that can be enriched with new terms by the communities that will use them.

The use of dereferenceable URIs for bibliographic data ensures the correct identification of persons and uses the connection to international authority services such as VIAF, the Virtual International Authority File. VIAF[26] is a service implemented and maintained by OCLC in cooperation with 20 national libraries, which virtually includes the authority files in the catalogues in a single authority service. The aim of the project is the reduction of costs and language barriers and availability on the Web of authorities that are exposed in linked open data.

For a long time catalogues have kept their authority files not always shared or accessible. Cataloguers have created the authors names following the rules laid down in their own cataloguing guidelines. The ability to link a given authority to VIAF, which contains all forms (REICAT, RAK, RDA etc.) created by cataloguers, besides simplifying cataloguing work, avoiding errors and duplications, creates an advantage in the search process. Someone carrying out a search in an OPAC could find a bibliographic description starting from the desired form for the name and find the link to all the variant forms in other languages.

[26] http://viaf.org/viaf/data

3.1.2 Application of FRBR and Improvement of Bibliographic Information Retrieval

As part of the studies made in the last two decades, Functional Requirements for Bibliographic Records (FRBR) is undoubtedly the most important theoretical reflection of the objects of cataloguing and bibliographic record functions in light of new technological changes, but yet still suffers from the failure of the model's cataloguing application.

The application of the FRBR model to catalogues has a direct and profound impact on searching and retrieving information. The full application of the model would allow the user to perform "significant navigation" in the catalogue, making it possible to clearly understand the retrieved entities.

The catalogue structured according to the FRBR model should give an account of the identity of a work under its various titles, the differences between the editions (manifestations), between the works (e.g., reports, including derivative works and the works that have other works as subject).

It should also allow the proper identification of the authors (especially useful in cases of same names), and the different roles of responsibilities (viewing the works and expressions in which an author has responsibilities of different types e.g., all the works in which he/she has a primary responsibility, all those in which instead he/she has a secondary responsibility). In particular the implementation of FRBR is essential to allowing the catalogue to perform its aggregating function, or grouping all the expressions and manifestations of a work, thus presenting to the user the works and related editions and links with other works in an orderly manner. It is a natural consequence of a higher *bibliographic control* or the possibility to better account for all the variations of names and titles.

Although there is now a consistent set of cataloguing rules that are inspired by the model outlined in the FRBR study (including the Italian cataloguing rules REICAT), the views of the conceptual model are applied with poor results in the library OPACs, reducing the quality of the information retrieved in the catalogue. To date, the only library catalogues that apply the FRBR model, are the "new generation OPACs" or discovery tools. In these instruments, however, we speak of "FRBR-isation" as a technique to obtain groupings according to the work, made possible through special algorithms employed retrospectively by the software without changing the structure of the original bibliographic records. Precisely for this reason, these groups often prove ineffective. In conclusion, even when doing research in OPACs that declare explicitly to apply the FRBR model (e.g., WorldCat), you cannot always easily retrieve the individual works, editions and reprints, as you find yourself having to further analyse the results retrieved.

Searching and browsing in online catalogues today requires special attention, especially for the verification of the different editions of a work and the rebuilding of links between works that are not always adequately made explicit. This happens because we do not intervene directly on the bibliographic record that, as we said earlier, is expressed in a format (MARC) that is not suitable to express the richness of bibliographic relationships provided by the model.

The deconstruction of the bibliographic records in data that are linked through LOD technology and the ability to qualify the links between the data by reconstructing the relationships required by the IFLA[27] Functional Requirements conceptual models will allow for a hierarchical navigation in the future that groups related works with the individual works, expressions and manifestations, and the items linked to them. This will effectively perform "bibliographic functions" for the benefit of the user when using the catalogue, making it possible to identify the "significant" contents of the catalogue. There are already significant applications in electronic and digital library catalogues.

As is clear from what has been presented so far, the application of LOD increases the chance that the information retrieved from online library catalogues will provide some elements and indicators to assess the activity and value of the scientific production of an author.

These new elements can be added to others, providing effective feedback in the evaluation process, improving the quality and amount of data retrieved when querying an OPAC.

3.1.3 Convergence of Search Tools

The new perspective of the Semantic Web directly involves the development of information retrieval systems such as electronic catalogues and means of carrying out searches, moving beyond traditional means of information retrieval, facilitating the discovery of the contents sought as well as supporting documents at the same time, and integration with other documents in the Web by creating a widespread network of knowledge. There will be, therefore, an increasing number of large platforms where the bibliographic information will be integrated with other information from other areas of cultural heritage, such as archives and museums. At that point from a given piece of bibliographic data it will be possible to find other data that enrich the information found, drawn from cultural heritage data available on the Web.

3.1.4 Improved Search Interfaces and New Ways of Cataloguing Research

Some concrete projects for implementing linked open data in library catalogues – such as the recent case of the OPAC of the Swedish National Library LIBRIS,[28] or data.bnf.fr project, the great platform of the Bibliothèque Nationale de France,

[27] This is referring to the models that make up the so-called "family" of IFLA functional requirements FRBR, FRAD, FRSAD, now in the process of consolidation in the new FRBR-Library Reference Model (IFLA 2016).

[28] The OPAC of the Swedish National Library LIBRIS is available at URL: http://libris.kb.se. Since 2008, the library displays the data as Linked data using ontologies known as Dublin Core, Bibliontology and SKOS, and is preparing to adopt RDF (Resource Description Framework) as its

which brings together bibliographic data, archival and museum (then incorporating the OPAC data of BNF) – show the potential of linked open data applications in the field of bibliographic research.

The Swedish OPAC LIBRIS features an innovative interface that connects data from multiple sources and displays them while maintaining the logical sequence that leads from the "works" to their "expressions", to the individual "manifestations" as provided for by FRBR models. This resolves a problem that affects today's electronic catalogues, yet still based on the single "manifestation" and providing the user with an interface based on the works, the authors, on the subjects. They exploit the spaces to explore the data related to every entity of interest in the catalogue, greatly improving the retrieval. You can get richer results and smarter groupings (groupings of works with related issues or events) with great benefit for research.

In the French platform data.bnf.fr the query interface offers three main channels of access to the catalogue: the work, the author, the subject. For each entity is shown a special page made of linked and updated data. The platform has a page for each work, topic, and author of the catalogue, structured according to the levels provided by FRBR, and created dynamically using the authority record for the work provided by the traditional catalogue of the Bibliothèque Nationale de France. For each work the key information about the content, the related works and their relationships with its parts are provided. The query by author responds with a page for each author containing biographical information, a list of works and an indication of other works connected with the indication of the type of relationship (Wenz 2013). The query by subject, on the other hand, responds to those carrying out a semantic search with a page for each subject. These subject pages contain all the variant and associated forms and shows the authors who have treated the subject, providing the works of the catalogue on that subject. All pages make use of further information from external datasets, in particular from DBpedia, the largest interdisciplinary datasets and benchmark for the extraction of data in Linked Open Data. External connections are imported into the platform via additional information about the works, titles and subjects. All data is queried by search engines through the application of RDFa standard markup and the use of shared vocabularies like Schema.org, recognised by the major search engines.

The Italian Share Catalogue project[29] shows how it is possible to create an interface in LOD for unified access to the catalogues of a group of academic libraries, through the creation of an LOD substrate of people-works that greatly simplifies the research process, exploiting the enrichment coming from the Web of data and the modelling coming from the BIBframe specifications.

native format of the catalogue records. The introduction of the new version of the catalogue Libris XL expressed in the new linked data format includes both the new and the previous records, the purpose of replacing MARC 21 with a data format as close to the data model outlined in Bibframe. (Forssblad 2013). The data are published in JSON-LD (JSON for linking data, http://json-ld.org). Today, thanks to the transformation of data in RDF, in response to a question the OPAC is able to provide a list sorted according to and within the FRBR model and will be able to provide a page for each title, author and subject of the database.

[29] http://www.sharecampus.it/1/share_catalogue_969557.html

The main and most enlightening perspectives, which at the moment, however, do not find any real application – come from the ability to enrich the bibliographic authority data with data coming from universities and research centres.

Of particular interest – in terms of richness of content obtained from a search in the catalogue – are the contributions of research institutions and their data that will be made available and could usefully enrich the bibliographic data.

This data may include:

- Universities' registers appropriately exposed in linked open data.
- Research data conveyed and made available in open format.[30]
- The data captured by ontologies linking scholars and research products, products in the context of cooperation projects among universities[31] (see recent projects that connect the scientific production of universities like, for example, *Linked data for Libraries project*[32] and *Linked universities project*[33]).
- Data from the institutional repositories that will be exhibited in the form of linked open data (Konstantinou et al. 2014).

The connection of all these data could provide a network of information or a kind of mapping of scientific research, which could be exploited as a useful outline to bibliographic environment.

3.1.5 Interoperability with Web: Quantity and Quality of Bibliographic Data in the Open Web

What has been stated so far focuses on the importance of interoperability. This theme has always been at the heart of library issues: libraries and IFLA have long been concerned with ensuring interoperability between archives, libraries, and other components of the bibliographic circuit. Today, though, the question becomes crucial: it is no longer sufficient to produce data that is interoperable between the actors of the cultural sector, but it must be interoperable with the Web i.e., to make library data freely accessible online.

---

[30] On the issue of the scientific data in open format see Cassella (2013).

[31] Among these VIVO is an ontology standard that connects researchers and communities using LOD, but also a data set and an open community with strong international participation that connects researchers, publications, resources for research, financing, required courses, scientific activities. The ontology is based on other known ontologies (SIOC and FOAF). The major classes of this ontology are people, organisations and information resources.

[32] *Linked data for Libraries* is a project of Cornell University Library, Harvard Library Innovation Lab and Stanford to create a LOD platform to extract data from various existing silos, to improve access to and discovery of academic information in the three libraries

[33] *Linked universities* is a project to connect the university library resources in LOD format involving various universities in the UK, Germany, Finland, Greece, Turkey, Czech Republic. The project aims to create and support specific vocabularies for universities for common concepts like courses, qualifications, teaching materials and sharing experience, the tools to format data of the University as linked data.

The application of linked data to library catalogues will deliver the bibliographic data to the entire Web. Bibliographic information can then be retrieved by search engines (both traditional and semantic). OCLC has been moving in this direction, releasing a huge amount of data of the world's largest single catalogue, WorldCat, in the form of linked open data and incorporating into their pages the marking provided by RDFa (RDF in Attributes) which makes it possible to embed semantic annotations pages in the Web following a special scheme[34] or ontology that is recognised by the major search engines.[35]

In the emerging future, through exchange formats compatible with the Semantic Web, these data may be freely available on the Web in the new structure of the bibliographic records designed in BIBFRAME, which is destined to host new and valuable information. The new bibliographic model developed by the Library of Congress for the Semantic Web and the world of linked data, BIBFRAME also includes some important new characteristics that are worth noting.

Along with data on the work, its editions, and authority data, the model also includes the Annotation class. This class is designed to include information about a resource, such as administrative and management information of the resource's life cycle (as traceability and provenance are part of the workflow that characterises the creation of a bibliographic record), but also other data that are added by others (users, commercial actors, the web community).[36] Not to mention metadata created by cataloguers: data resource locations and holdings, and access policies (Mitchell 2013).

This will involve the online presence of management data for the item that are today "trapped" in library management systems (ILS) and are not present on the Web. We know how complex it is to extrapolate a number of management data for assessment purposes, such as whether a monograph was purchased or donated. In fact there is no trace of this data in the library OPAC, but they remain stored in the ILS, together with other data about the copy (the FRBR item).

Along with this information, annotations can also be created by third parties and library users, who will enrich the record with images, reviews and ratings obtained from reliable sources.

With the application of linked data bibliographic information disintegrates into the Semantic Web, therefore drawing upon data from different sources. It will be of primary importance to ensure the quality of cataloguing data, a problem that is both theoretical (what does quality of data mean) and practical (how to ensure quality

[34] https://schema.org/

[35] The OCLC data policy is presented in http://www.oclc.org/data.en.html. See as an example of enriched records with linked open data record in WorldCat. https://www.worldcat.org/title/information-a-very-short-introduction/oclc/743804876&referer=brief_results where you can see the RDFa encoding based on Schema.org (www.schema.org)

[36] The purpose of the Annotation class is to "Express opinions about a resource, for example a review. Attach institution specific information, for example holdings. Contribute enhancements to a resource description, for example cover art or summary descriptions". See BIBFRAME Annotation Model (2013).

through the certification of data provenances) and that greatly impacts the assessment practices. The commitment of the libraries, so it is hoped, will be to build certified and reliable data networks.

References

Berners-Lee, T. (2006). Linked data: Design issues [updated version 2009 06 18]. http://www.w3.org/DesignIssues/LinkedData.html. Accessed 17 July 2016.

Biagetti, M. T. (2010). Nuove funzionalità degli OPAC e relevance ranking. *Bollettino AIB, 50*(4), 339–356.

BIBFRAME Annotation Model. (2013). *BIBFRAME Community Draft, 26 August*. http://bibframe.org/documentation/annotations/#whatIsAnAnnotation. Accessed 17 July 2016.

Borgman, C. L. (1996). Why are online catalogs still hard to use? *Journal of the American Society for Information Science, 47*(7), 493–503.

Breeding, M. (2013). Next-generation discovery: An overview of the European scene. In S. Chambers (Ed.), *Catalogue 2.0. The future of library catalogue* (pp. 37–64). London: Facet Publishing.

Breeding, M. (2015). *The future of library resource discovery*. Baltimore: NISO.

Byrum, J. D., Jr. (2003). NACO: un modello di cooperazione per l'allestimento e la manutenzione di un authority database condiviso per i nomi. In *Authority control: definizione ed esperienze internazionali. Atti del Convegno internazionale (Firenze, 10–12 febbraio 2003)* (pp. 187–198). Firenze/Roma: Firenze University Press/Associazione Italiana Biblioteche.

Cassella, M. (2013). Dati aperti e ricerca scientifica: aspetti gestionali e normativi nel contesto dell'e-science. *AIB Studi, 53*(3) doi:https://doi.org/10.2426/aibstudi-9131.

Christensen, A. (2013). Next-generation catalogues: what do users think? In S. Chambers (Ed.), *Catalogue 2.0. The future of library catalogue* (pp. 1–15). London: Facet Publishing.

Coyle, K. (2010). RDA vocabularies for a twenty-first-century data environment. *Library Technology Reports, 46*(2) (monographic issue).

Ellero, N. P. (2013). Integration or disintegration: Where is discovery headed? *Journal of Library Metadata, 13*(4), 311–329.

Forssblad, H. (2013). *Uppdaterad information om Libris XL och nya katalogiseringsverktyget*, Librisbloggen.kb.se (blog).

Frederiksen, L. (2015). Exploring discovery. *Public Services Quarterly, 11*(1), 43–47.

Genetasio, G., & Terravecchia, E. (2009). Authority control at Casalini libri. *Biblioteche oggi, 27*(10), 42–52.

Gong, R. (2012). A review of planning and implementing resource discovery tools in academic libraries. *Journal of Electronic Resources Librarianship, 24*(4), 340–341.

Han, M. (2012). New discovery services and library bibliographic control. *Library Trends, 61*(1), 162–172.

ICCU. (1995). *Guida alla catalogazione in SBN. Pubblicazioni monografiche, pubblicazioni in serie* (2nd ed.). Roma: ICCU.

ICCU. (2009). *Regole italiane di catalogazione REICAT*. A cura della Commissione permanente per la revisione delle regole italiane di catalogazione. Roma: ICCU. http://www.iccu.sbn.it/opencms/export/sites/iccu/documenti/2015/REICAT-giugno2009.pdf.

ICCU. (2015). *Linee guida per la compilazione dell'Authority Control-Autore personale in SBN*. http://www.iccu.sbn.it/opencms/export/sites/iccu/documenti/2015/AF_Indicazioni_e_regole_riv_23_luglio2015.pdf. Accessed 18 Sept. 2016.

IFLA. (1998). *Functional requirements for bibliographic records: Final report*. München: K. G. Saur. http://www.ifla.org/VII/s13/frbr/frbr.pdf. Accessed 20 Sept 2016.

IFLA. (2007). *UNIMARC Manual for Holdings Format. Version 1. Recommended by the working group on the UNIMARC holdings format*. Approved by the Permanent UNIMARC Committee (PUC), pp. 36–86. http://www.ifla.org/files/assets/uca/unimarc-holdingsformat.pdf. Accessed 30 Dec. 2015.

IFLA. (2008). *Cataloguing principles: Statement of international cataloguing principles (ICP) and its glossary*. Edited by Barbara Tillett and Ana Lupe Cristán, München, K.G. Saur http://www.ifla.org/VII/s13/icc/imeicc-statement_of_principles-2008.pdf Italian version. IFLA (2009) Dichiarazione di principi internazionali di catalogazione, Roma, ICCU http://www.ifla.org/files/cataloguing/icp/icp_2009-it.pdf

IFLA. (2009a). In G. E. Patton (Ed.), *Functional requirements for authority data. A conceptual model*. München: K.G. Saur. Online version, as amended and corrected through July 2013, http://www.ifla.org/files/assets/cataloguing/frad/frad_2013.pdf.

IFLA. (2009b). *Study group on the functional requirements for bibliographic records, functional requirements for bibliographic records: Final report*. Approved by the Standing Committee of the IFLA Section on Cataloguing. September 1997. As amended and corrected through February 2009, http://www.ifla.org/VII/s13/frbr/frbr_2008.pdf.

IFLA. (2011). *ISBD International standard bibliographic description*, Consolidated edition. Edizione italiana, a cura dell'Istituto centrale per il catalogo unico delle biblioteche italiane e per le informazioni bibliografiche, Roma: ICCU 2012, p. 241. http://www.iccu.sbn.it/opencms/export/sites/iccu/documenti/2012/ISBD_NOV2012_online.pdf.

IFLA. (2016). *Library-reference model*. Pat Riva, Patrick Le Bœuf, and Maja Žumer Consolidation Editorial Group of the IFLA FRBR Review Group Draft for World-Wide Review 2016–02-21. Not yet endorsed by the IFLA Professional Committee or Governing Board, http://www.ifla.org/files/assets/cataloguing/frbr-lrm/frbr-lrm_20160225.pdf. Accessed 17 July 2016.

Ilik, V. (2015). Cataloger makeover: Creating non-MARC name authorities. *Cataloging & Classification Quarterly, 53*(3–4), 382–398.

Joint Steering Committee for Development of RDA (ed.) (2010). *Resource description & access*. RDA. American library association; Canadian library association; CILIP.

Kelley, M. (2012). Coming into focus: Web-scale discovery services face growing need for best practices. *Library Journal, 137*(17), 34–40.

Konstantinou, N., et al. (2014). Exposing scholarly information as linked open data: RDFizing DSpace contents. *Electronic Library, 32*(6), 834–851. Accessed 17 July 2016.

Kroeger, A. (2013). The road to BIBFRAME: The evolution of the idea of bibliographic transition into a post-MARC Future. *Cataloging & Classification Quarterly, 51*(8), 1–18.

Library of Congress Working Group On The Future Of Bibliographic Control. (2008). *On the record: Report of the library of congress working group on the future of bibliographic control*. Washington, DC: Library of Congress. http://www.loc.gov/bibliographic-future/news/lcwg-ontherecord-jan08-final.pdf. Accessed 17 July 2016.

Marchitelli, A. (2015). *Orientarsi tra le informazioni in biblioteca. Cataloghi, banche dati, motori di ricerca*. Milano: Editrice Bibliografica.

Marchitelli, A., & Frigimelica, G. (2012). *OPAC*. Roma: Associazione Italiana Biblioteche.

Miller, E., et al. (2012). *Bibliographic framework as a web of data: Linked data model and supporting services*. Washington, DC: Library of Congress.

Mitchell, E. (2013). Assessing the value of metadata in information services. *Technical Services Quarterly, 30*(2), 187–200.

Morriello, R. (2006). Acquisizioni tramite approval plan. Outsourcing o nuova opportunità per i bibliotecari? *Biblioteche oggi, 24*(1), 22–31.

Nagy, A. (2011). *Analyzing the next-generation catalog*. Chicago: American Library Association.

Nardini, R. F. (2003). Approval plans. In M. A. Drake (ed.), *Encyclopedia of library and information science*. New York: Dekker. 1, 133–138. Accessed 28 June 2016. doi:https://doi.org/10.1081/E-ELIS 120008874.

Naun, C. C. (2010). Next generation OPACs: A cataloging viewpoint. *Cataloging & Classification Quarterly, 48*(4), 330–342.

NISO. (2013). *ODI survey report: Reflections and perspectives on discovery services*. Available at http://www.niso.org/apps/group_public/download.php/9977/NISO%20ODI%20Survey%20Report%20Final.pdf. Accessed 16 July 2016.

OCLC Annual Report. (2014–2015). http://www.oclc.org/etc/designs/oclc/oclc-annualreport-15/images/215510-2014-2015OCLCAnnualReport.pdf.

Patton G (ed.) (2009) Functional requirements for authority data: A conceptual Model. Ed. by Glenn E. Patton. IFLA working group on functional requirements and numbering of authority records, final report December 2008. IFLA series on bibliographic control, 34. K.G. Saur, Munich.

Petrucciani, A. (1984). *Funzione e struttura del catalogo per autore*. Firenze: La Nuova Italia.

Petrucciani, A. (2006). La catalogazione, il mercato e la fiera dei luoghi comuni. *Bollettino AIB, 46*(3), 177–185.

Raieli, R. (2015). Limiti dell'information discovery e necessità dell'information literacy. *Nuovi annali della Scuola speciale per archivisti e bibliotecari, 29*, 179–194.

Somerville, M. M. (2013). Digital age discoverability: A collaborative organizational approach. *Serials Review, 39*(4), 234–239.

Sweet M (2012) There's nothing wrong with discovery services that can't be fixed by the reference layer. In: *The Fiesole collection development retreat*, April 12 2012. Available at http://www.casalini.it/retreat/2012_docs/sweet.pdf. Accessed 16 July 2016.

US Library of Congress and Zepheira. (2016). *Overview of the BIBFRAME 2.0 Model*. https://www.loc.gov/bibframe/docs/bibframe2-model.html. Accessed 17 July 2016.

Wang, Y., & Mi, J. (2012). Searchability and discoverability of library resources: Federated search and beyond. *College & Undergraduate Libraries, 19*(2–4), 229–245.

Way, D. (2010). The impact of web-scale discovery on the use of a library collection. *Serials Review, 36*(4), 214–220.

Wenz, R. (2013). Linked open data for new library services: The example of data.bnf.fr. *JLIS.it, 4*(1), 403–416. doi:https://doi.org/10.4403/jlis.it-5509. Accessed 17 July 2016.

Yee, M. (2009). Can bibliographic data be put directly onto the Semantic Web? *Information Technology and Libraries, 28*(2), 55–80. also available online: http://escholarship.org/uc/item/91b1830k.

A Survey on Legal Research Monograph Evaluation in Italy

Ginevra Peruginelli, Sebastiano Faro, and Tommaso Agnoloni

1 Introduction

Within the general context of scientific research assessment exercises, evaluation of legal research has specific characteristics due to the peculiarity of legal science [see the Chapter "Research Quality Evaluation: The Case of Legal Studies"]. These characteristics, partly shared by humanities and social sciences, have raised a lively debate about this topic. The debate is focused in particular on the notion of research quality in legal studies and on the way to evaluate the quality of research once quality has been specified.

The discussions are very lively in Italy, in particular since 2011 when the Agency for the Evaluation of University and Research System (*Agenzia Nazionale per la Valutazione del sistema Universitario e della Ricerca – ANVUR*)[1] launched national exercises on evaluation of quality research (VQR 2004–2010[2] and VQR 2011–2014).[3]

The effectiveness and significance of legal research assessment as well as the tools to be used to evaluate research outputs are at the core of the debate in Italy (Conte 2015a; Carinci and Brollo 2013). On the latter, in particular, there is a widely held opinion that bibliometric evaluation techniques cannot and should not be used in legal sciences. On this matter, the Conference of the Scientific Association of Legal Area (*Conferenza delle Associazioni Scientifiche di Area Giuridica – CASAG*)

[1] www.anvur.org

[2] Call of the *Valutazione della Qualità della Ricerca 2004–2010* (VQR 2004–2010) http://www.anvur.org/attachments/article/122/bando_vqr_def_07_11.pdf

[3] Call of the *Valutazione della Qualità della Ricerca* 2011–2014 (VQR 2011–2014) http://www.anvur.org/attachments/article/825/Bando%20VQR%202011-2014_secon~.pdf

G. Peruginelli (✉) • S. Faro • T. Agnoloni
Institute of Legal Information Theory and Techniques of the National Research Council of Italy (ITTIG-CNR), Florence, Italy
e-mail: ginevra.peruginelli@ittig.cnr.it; sebastiano.faro@ittig.cnr.it; tommaso.agnoloni@ittig.cnr.it

© Springer International Publishing AG 2018

A. Bonaccorsi (ed.), *The Evaluation of Research in Social Sciences and Humanities*, https://doi.org/10.1007/978-3-319-68554-0_9

has clearly confirmed this view (CASAG 2014). Peer review has been considered the only reliable method for assessing research quality in the field of law.

This position has been confirmed also by practice: the first national assessment exercise for the years 2004–2010 was based on peer review and the same goes for the latest exercise related to research outputs of the 2011–2014 period.

Despite this attitude and practice, the topic of bibliometric indicators to support peer review cannot be ignored. This issue is in fact at the core of the current international debate on research quality criteria for the assessment of legal scholarship (van Gestel and Vranken 2011). Bibliometric evaluation techniques specifically defined for the area of legal studies can become a valuable tool to address the limitations of peer review (i.e., its time consuming process, scarce availability of expert reviewers and the increasing demand for research output evaluation).

Within this context, in 2015, the Institute of Legal Information Theory and Techniques of the National Research Council of Italy (ITTIG-CNR) conducted a research study on the assessment of legal research monographs in Italy (hereafter ITTIG Project). The ITTIG Project's objective was to verify the opportunity to identify bibliometric indicators that can support the peer review of legal monographs. For this purpose two surveys at a national and international level were carried out, as well as a comparative analysis of research evaluation systems adopted in three foreign countries (the Netherlands, United Kingdom and France).[4]

In particular, this chapter presents one of the results of the ITTIG Project: the survey undertaken on Italian legal academia. The survey aimed at discovering legal scholars' view on the adoption of quality and impact indicators of legal monographs in support of peer review. The national survey, as discussed later on, achieved a high level of participation: this makes its results particularly significant for understanding the dominant feeling among Italian legal scholars. It also shows, in a sufficiently realistic way, the framework and spirit in which possible proposed indicators in the area of legal research assessment could be applied.

2 The Scope of the ITTIG Project: The Monograph

The evaluation of research in the humanities, social sciences and law is a particularly delicate issue. The heterogeneity of the research and scientific products, the specific nature of the communication channels, the absence of objective data to support the assessment – due also to poor coverage in the citation databases – are some of the factors that are fuelling the actual debate.

In the fields of social sciences, including both law and economics, it is important to note that the most relevant channel in the scholarly communication process is represented by books: this literary genre is chosen, among others like journal

[4] The first two countries have been chosen for their long-standing experience in the science of the quality of research evaluation, France for the similarity with the Italian context.

articles, for its ability to fulfil specific research needs. A monographic work offers the room to properly describe the context under study and to analyse all the variables, quoting all the sources, thus providing evidence of what is claimed.

Hicks (2004) observes the book as one of the four literary genres particularly fitted to social sciences. In particular, Hicks notes that a book reflects a humanities-type rather than a scientific approach to scholarship and it is most often based on qualitative rather than quantitative evidence, often being the result of the work of more experienced researchers.

To properly frame the scenario in which the ITTIG Project is placed, it is worth pausing to clarify our definition of monograph. Several definition are proposed by the literature: those that capture the fundamental aspects of the genre are outlined below.

"A large, specialized work of scholarship that treats a narrow topic in great detail. Size is a critical characteristic, because it distinguishes the monograph from the article, which has the same purpose, but is small. It presents what the scholar has concluded is the truth about some set of historical events, the characteristics of some work of art or literature or the biography of a historical figure, an artist or a writer" (Chodorow 1999).

"A work of scholarship on a particular topic or theme. It is written by a scholar (or scholars) and intended for use primarily by other scholars" (Thompson 2002, 2005).

"Books, which are records of primary research intended for other researchers and bought mainly by libraries.... In (certain) disciplines such work represents the main channel for communication of research and is recognized as such for purposes of tenure and promotion" (Watkinson 2001, p. 15 and 70).

"A monograph is a long academic book on a single research topic, normally written by one or on occasions two authors" (Higher Education Funding Council for England 2015).

"Opera di consistente estensione, da imputare per intero alla responsabilità scientifica di uno o più autori, e che si propone quale studio approfondito e caratterizzato da un approccio critico" (Consiglio Universitario Nazionale 2013) [English translation: Extensive work, whose scientific responsibility is of one or more authors, setting itself as an in-depth study characterised by a critical approach].

"Contributo scientifico su un singolo, ben definito argomento la cui trattazione è dettagliata con impostazione sistematica ed apertura critica al complessivo dibattito accademico sui temi trattati. L'autore/gli autori deve/devono avere responsabilità intellettuale diretta sull'intero contenuto dell'opera, pubblicata sotto forma di libro. La tipologia non comprende la manualistica scolastica o universitaria, ad eccezione di opere con contenuti scientifici originali e innovativi" (Osservatorio della ricerca dell'Alma Mater Studiorum – Università di Bologna 2013) [English translation: Scientific contribution on a single, well-defined topic covered in detail with a systematic and critical approach to the complex academic debate on the subject. The author(s) should have direct responsibility over the whole intellectual content of the work, published as a book. Handbooks, with the exception of works with original and innovative scientific content, are not considered scientific monographs].

According to these definitions, the monograph proves to be an original contribution on a single, well-defined topic, whose treatment is carried out with a systematic approach and critical capabilities to contribute to the overall academic debate on the subject matter. The work must have an adequate extension and structure, and the author(s) should have direct responsibility for the entire intellectual content of the work, published in book form. However, the concept of monograph does not coincide with that of a book. In the field of law, for instance, a work presented in the form of a book may contain a mere review of legal literature and case law. This work does not necessarily constitute an original work, as in this case a critical and constructive approach required for scientific monographs is missing.

By contrast, it may happen that scientific relevance and organic unity of a monograph are present in studies dedicated to treaties, commentaries or encyclopaedic entries whose format is not one of a monographic publication. These studies could have their own independence and be published as a monograph. Furthermore, a volume intended for teaching, as long as it represents a synthesis of original thought and the cultural experience of its author could be considered a scientific monograph.

As regards the value of monographs, there is no doubt that in the field of legal studies the monograph is the principal "coin of the realm", used as a paramount criterion for promotion decisions by universities. Monographs confirm themselves as the fundamental tools to provide documentary evidence of research results in law (Di Raimo 2015).

In particular, in contrast with scientific legal journals for which evaluation systems exist based on the classification of journals, no structured methods of legal research assessment for monographs have been developed. Still today in major Western countries, there are very few experiences of classification of publishers and of editorial series.[5] Evaluating legal monographs through the creation of lists and classifications of different types of books, publishers or editorial series is much more difficult than dealing with journals (van Gestel 2011). This is due to the many factors of distortion of the evaluation, such as the presence of editorial series by the same publisher having different weight and prestige, the extreme specialisation of some works that, despite their value, do not catch the attention of any major publisher.

It is evident that the process of evaluation of this particular scientific product in the legal domain is very delicate. Peer review is definitely an expensive activity in terms of intellectual effort and time. However, it remains the most suitable method, hopefully supported by an editorial system based on transparency and verification of the quality of content independent of any connection with the author.

[5] An example is the Scholarly Publishers' Indicators (SPI) initiative of the Spanish National Research Council (CSIC). It is an information system that tracks a ranking of publishers (Spanish and otherwise), of scientific books in the areas of humanities and social sciences. http://ilia.cchs.csic.es/SPI

3 The ITTIG Project Survey: Structure, Target Audience, Respondent's Profile

The idea behind the ITTIG Project was that any proposal on the use of indicators for the evaluation of legal monographs must give sufficient consideration to legal scholars' opinion. This idea relies on the fundamental epistemological principle that scholars of a specific research area are best suited to recommend relevant indicators for the evaluation of publications containing research results. Based on this, the ITTIG Project envisaged a detailed survey (questionnaire) involving Italian legal scholars.

Two investigations inspired the Italian survey: one national and one conducted in Switzerland. The "Public consultation for the establishment of the National Registry of Professors and Researchers and Scientific Publications" (*Consultazione pubblica per l'istituzione dell'Anagrafe nazionale nominativa dei professori e dei ricercatori e delle pubblicazioni scientifiche – ANPRePS)* was prepared by the National University Council (*Consiglio Universitario Nazionale – CUN*) and proposed to Italian scholars in 2013. The investigation concerned the criteria for the definition of the scientific nature of research outputs. The Swiss investigation was based on the Swiss questionnaire undertaken within the project "*Évaluation de la recherche en droit*",[6] part of the wider research programme "*Performances de la recherche en sciences humaines et sociales*", coordinated by the *Conférence des recteurs des Universités Suisses* (CRUS). The Swiss survey, conducted in 2014, intended to define the concept of quality of legal research and to prepare an inventory of standards and existing procedures. The survey was sent to various stakeholders (such as law professors, scientific committees of law reviewers, a selection of lawyers practicing in Switzerland).

Both surveys have been very helpful for the drafting of a number of questions proposed in ITTIG's questionnaire.

The ITTIG questionnaire was structured in three sections. The first was dedicated to the experience and knowledge of respondents in relation to the questionnaire's subject matter. The second specifically addressed the indicators for the assessment of legal monographs. This section was divided in three parts: (i) definition of a legal monograph; (ii) quality and impact indicators for legal monographs; (iii) focus on specific aspects of indicators (ranking of publishers and series, usefulness of citation databases, risks related to the use of indicators). The last section of the questionnaire covered specific information on respondents.

Questions were posed with the choice between default options or allocation of a Likert scale to investigate and measure respondents' opinions. For each question, except those relating to respondents' profile, the chance to enter a free comment was offered, with no character limit.

In May 2015 4729 professors and researchers of Italian universities were identified using the public database of the Ministry of Education, Universities and

[6] http://www.unige.ch/droit/cetel/recherches/evalRecherche.html

Research (MIUR).[7] Scholars were grouped in the 21 subject areas that make up the area of legal sciences. The list of the academic personnel in the field of law was identified and an email address associated with each name, where possible. 4645 invitations to participate were sent via email. The invitations arrived to 4501 respondents (target population) representing 95.2% of the workforce resulting in the Ministry database. On 8 July 2015 the questionnaire was made available online. On 15 September 2015, the day when online access to the questionnaire was closed, a total of 1241 answers had been received, representing about 26% of the target population.

Below some significant data are reported in order to better describe the respondents of the survey and their profiles. This makes it possible to understand not only who the respondents of the questions are but also to make interesting statistical data analysis and calculations that will be partly discussed in Paragraph 4.

If we compare the number of invited scholars with the number of respondents, a different interest of scholars to the questionnaire emerges according to their affiliation to areas of law, as indicated by the respondents. Philosophers of law were very interested (42.73% of the invited responded) compared to scholars in public law (only 14% of the invited responded) and in criminal procedure law (19% of the invited answered).

Table 1 shows the distribution of the invited scholars by areas of law, the number of respondents and the percentage of the latter compared to the invited scholars (the difference between the number of MIUR academic personnel and the invited respondents depends on the fact that it was not possible to find all the email addresses of Italian legal scholars).

The breakdown of respondents of the four categories of scholars (professor, associate professor, researcher with fixed-term contract, researcher with short-term contract) is presented in Table 2. In particular, this table shows that the distribution of the different type of scholars who have responded was balanced. All of them were quite interested in expressing their ideas on legal research evaluation.

As regards respondents' age, the vast majority were over 40: 64.3% were between 41 and 60 and 22.85% were over 60.

To better define the profile of respondents and verify the extent to which some special knowledge and experience could orient their answers, information was collected on four different aspects.

1. The first aspect concerns the experience of respondents in editorial roles (Table 3). A large majority of them have experience as scientific and/or editorial committee members of Italian journals and as reviewers of manuscripts (journal articles) on behalf of Italian journals. 140 respondents are editorial board chairs of Italian journals and 124 are editorial board chairs of Italian series. By contrast, the number of editorial board chairs of foreign journals (17) and of editorial board chairs of foreign series (13) is quite low.

[7] Professors' website – loginmiur: https://loginmiur.cineca.it/

Table 1 Number of respondents compared to MIUR academic personnel and the invited scholars

Areas of law	MIURstaff	Invited	Respondents	Response rate (%)
IUS01 – Private Law	688	677	143	21.12
IUS02 – Comparative Private Law	162	161	43	26.71
IUS03 – Agri-food Law	41	40	12	30.00
IUS04 – Business law	408	390	113	28.97
IUS05 – Economics Law	66	65	22	33.85
IUS06 – Navigation and Air Law	47	45	9	20.00
IUS07 – Labour Law	295	294	77	26.19
IUS08 – Constitutional Law	234	230	73	31.74
IUS09 – Public Law	285	277	39	14.08
IUS10 – Administrative law	422	410	90	21.95
IUS11 – Ecclesiastical and Canon Law	111	109	33	30.28
IUS12 – Tax Law	186	184	40	21.74
IUS13 – International Law	265	260	72	27.69
IUS14 – European Union Law	106	103	33	32.04
IUS15 – Civil Procedural Law	199	196	45	22.96
IUS16 – Criminal Procedure	185	181	35	19.34
IUS17 – Criminal Law	274	273	66	24.18
IUS18 – Roman and Ancient law	235	235	75	31.91
IUS19 – History of Medieval and Modern Law	159	157	58	36.94
IUS20 – Philosophy of Law	229	227	97	42.73
IUS21 – Comparative Public Law	132	131	41	31.30

Table 2 Breakdown of respondents of the four categories of scholars

	MIUR	Respondents	% respect to MIUR personnel
Tenured	1455	385	26.46
Associate	1303	388	29.78
Researcher fixed-term	1648	360	21.84
Researcher short-term	257	55	21.40
Other	61	4	6.56

2. The second aspect concerns the involvement in evaluation activities (Table 4). It is noted that a large number of respondents have evaluation experiences: the predominant activity is the ex-ante evaluation of journal articles to be published. Moreover, about 40% of the respondents carried out ex-ante evaluation of monographs.
3. The third aspect regards the international experience of the respondents, defined in terms of experience in publishing with foreign publishers (Fig. 1) and in conducting research abroad (Fig. 2). On the first point, the results show that about 41% of respondents have never published with a foreign publisher. Of the

Table 3 Number and percentage of respondents with experiences in editorial roles

Editorial role		
Editorial board chair of Italian journal	140	13.2%
Editorial board chair of foreign journal	17	1.6%
Editorial board chair of Italian series	124	11.7%
Editorial board chair of foreign series	13	1.2%
Scientific and/or editorial committee member of Italian journal	809	76.4%
Scientific and/or editorial committee member of foreign journal	176	16.6%
Scientific and/or editorial committee member of Italian series	240	22.7%
Scientific and/or editorial committee member of foreign series	48	4.5%
Reviewer of manuscripts (journal articles) on behalf of Italian journal	657	62%
Reviewer of manuscripts (journal articles) on behalf of foreign journals	204	19.3%
Reviewer of manuscripts (monographs) on behalf of Italian publisher	255	24.1%
Reviewer of manuscripts (monographs) on behalf of foreign publisher	50	4.7%
Other	88	8.3%

Table 4 Number and percentage of respondents involved in evaluation activities

Evaluation activities		
PhD thesis evaluation	709	67.3%
Ex-ante evaluation of monographs	391	37.1%
Ex-ante evaluation of journal articles	796	75.5%
Participation in commissions for researchers professors appointments	262	24.9%
Research project evaluation (ex ante)	364	34.5%
Research project evaluation (in progress or ex post)	105	10%
Awards or recognition award	141	13.4%
Ex-post evaluation of monographs already published (for example, within competition or evaluation exercises)	295	28%
Ex-post evaluation of journal articles already published (for example, within competition or evaluation exercises)	328	31.1%

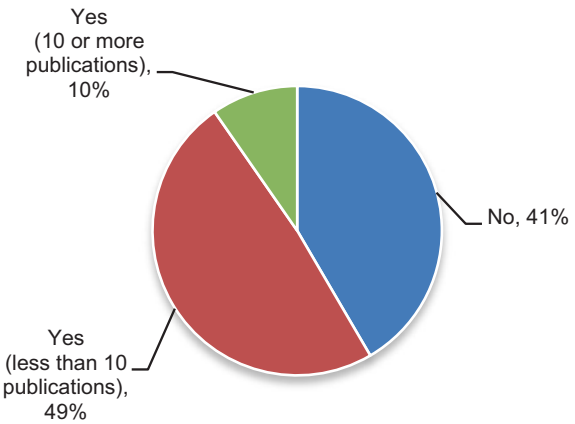


Fig. 1 Experience in publishing with foreign publishers

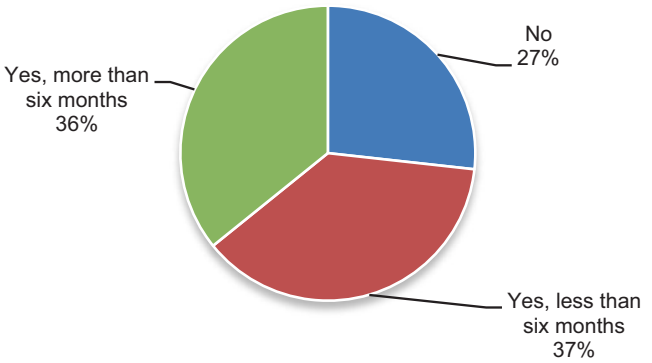


Fig. 2 Research experience abroad

Fig. 3 Confidence in bibliometrics and associated tools

remaining portion of respondents who have published with foreign publishers, almost 10% have significant experience (over 10 publications each). On the second point, a total of over 70% of respondents have conducted research activities abroad (almost half of them for more than 6 months).

4. Finally, the fourth characteristic of the respondents was related to their confidence in bibliometrics and associated tools (Fig. 3). Only about 10% of respondents declared a good or excellent knowledge of bibliometrics, whereas 16% had any knowledge.

4 Perceptions of Legal Scholars: Some Key Results of ITTIG Survey

Four major issues have been raised from the questionnaire. Below, data obtained from answers and comments given by respondents are presented. In some cases the comments received have been very numerous and are as important as the answers to grasp the approach of the Italian legal academia.

4.1 The Definition of Monograph in Legal Research

For the purposes of the ITTIG Project a specific definition of legal monograph was adopted using the following notion established by the National University Council (*Consiglio Universitario Nazionale* – CUN): "a work of considerable extent, entirely under the scientific responsibility of one or more authors, that is an in-depth study, characterised by a critical approach that, in addition, presents original results, supported by references reporting documentary, bibliographic and methodological foundations".

The great majority of respondents (87.5%) agreed on this definition. Those who did not agree with this definition were invited to comment on it and to indicate the reasons for their disagreement. The opportunity to leave a comment was also offered to those who agreed with the definition.

Some recurrent observations in the more than 150 comments received – in some cases very long and articulated – follow:

- Monographs should refer to a single author (or at least to a limited number of authors). This argument is often associated with the idea that monographs are an expression of full scientific maturity reached by the author, and are presented as the result of a long and outstanding research. Even those who admit that monographs can be the work of many authors tend to emphasise that it is necessary to clearly identify the specific parts of the text attributed to each author.
- The expression "considerable extent" is considered too ambiguous. Suggested remedies in the comments include a specific reference to a minimum number of pages or characters.
- The concept of "original results" is considered too vague (as alternative some respondents propose the concept of newness). Similarly, the concept of "critical approach" is considered too indefinite.
- Reference to the breadth of the research topic is also missing in the definition. It was noted that the broader the research topic, the more difficult it is to achieve original and critical results.

Two more questions were strictly linked to the definition of monograph: one related to the characteristics that a research output should have to be considered a monograph; the other concerning the possibility of considering as monographs some specific types of publications (such as manuals, handbooks, volumes for the professional world, for dissemination purposes…). With regard to the first question, the vast majority of respondents focused on the necessity that monographs (i) have an ISBN code, (ii) contain references to bibliographic sources, (iii) include explanatory and/or integrative notes to the research content, (iv) have a bibliography, and (v) be organised into chapters and sections. However, an index of authors and of the cited sources is not considered fundamental. With regard to the second question on types of publication to be considered as monographs, about 70% of respondents consider that manuals and handbooks can be considered as monographs. This percentage is composed of the sum of respondents who believe that these types of publications are always considered monographs in every case (19%) and of those

who believe that they are monographs only if they present original results (50%). The number of respondents who consider that volumes for professionals can be regarded as monographs only when presenting original results is just over 50%. On the other hand, 57% of respondents do not agree that volumes for dissemination purposes can be considered as monographs.

4.2 Indicators for Assessing the Quality of Legal Monographs

The core of the questionnaire concerned the opinion of respondents on a number of indicators to support the evaluation of the quality of legal monographs.

According to Baccini (Baccini 2010) three different dimensions of quality were identified and proposed to the respondents:

(a) Quality *tout court*, defined as compliance with the standards required by the scientific community for research be considered of high quality (hereafter "quality").
(b) Impact and dissemination of a monograph within the scientific community, to be intended as the ability to influence research conducted by others and to produce knowledge for use by others (hereafter "impact on scientific community").
(c) Societal impact of a monograph.

The list of proposed indicators is the following:

- Publisher
- Inclusion in a particular series with special characteristics (review process…)
- Availability in the catalogues of national and international libraries
- Indexing in existing citation databases (Web of Science, Scopus, SSRN …)
- Indexing in citation databases specifically created for the legal domain
- Number of citations received
- Reviews in scientific journals
- Reviews in specific scientific journals (only highly ranked journals)
- Reporting in scientific journals
- Reporting in specific scientific journals (only highly ranked journals)
- Comments and citations in no scientific sites (e.g., newspapers)
- Number of copies sold
- Number of downloads in case of monograph available online
- Publishing cost covered by a research project financed on the basis of an evaluation process
- Quotes in policy documents

Respondents were asked to indicate the usefulness of indicators so as to define the quality, the impact on the scientific community and the societal impact of a monograph. Four possible rating (high, fair, low, none) were proposed to respondents.

Table 5 First five positions in the ranking of indicators according to quality, impact on scientific community, societal impact (all respondents)

Rank	Quality	Impact on scientific community	Societal impact
1	Inclusion in a particular series with special characteristics (review process…)	Inclusion in a particular series with special characteristics (review process…)	Availability in the catalogues of national and international libraries
2	Publisher	Availability in the catalogues of national and international libraries	Publisher
3	Availability in the catalogues of national and international libraries	Publisher	Inclusion in a particular series with special characteristics (review process…)
4	Reviews in scientific journals	Reviews in scientific journals	Number of downloads in case of monograph available online
5	Reporting in scientific journals	Reviews in specific scientific journals (only highly ranked journals)	Number of copies sold

On the basis of respondents' answers a ranking of indicators was obtained. If we refer to the first five positions – summarised in Table 5 – it emerges that for the quality and the impact on scientific community there is consensus on indicators considered as more reliable. On the other hand, as regards societal impact it is not surprising that different indicators are ranked in the fourth and fifth position. However, this ranking should be read as an overall sceptical view of the concept of societal impact referred to legal monographs, which is very difficult to define and also very hard to assess (as clearly evidenced in the comments to the question related to the dimension of quality).

In general terms and with regard to the most commonly used indicator in Science, Technology, Engineering and Medicine (STEM), neither indexing in existing citation databases or specifically created for the legal domain nor the number of citations received is considered robust indicators.

The Fig. 4 shows the ranking of the indicators with regard to each quality dimension.

In detail, it is interesting to verify how the answers to questions on indicators were affected (i) by experience in editorial roles, (ii) by international experience of respondents defined in terms of publishing with foreign publishers and of conducting research abroad, (iii) by the scholars' affiliation with a specific area of law, and (iv) by the level of knowledge of bibliometrics. It is interesting to notice that profiles of respondents do not significantly affect the ranking of the indicators. There are no notable differences in the ranking of the indicators when considering the answers of subgroups of respondents with international experience and with the profiles of editorial board chairs of Italian and foreign journals or series.

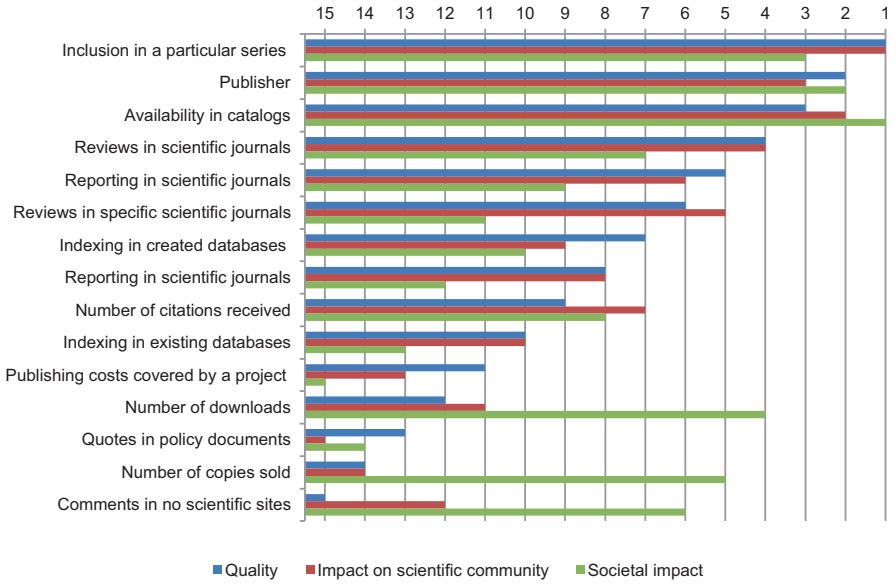


Fig. 4 Ranking of all indicators compared to quality, impact on scientific community, societal impact

With respect to the dimension of quality, more significant data are obtained by establishing 21 rankings, one for each area of law, based on the answers given by the respondents who declared their affiliation with a specific area of law. As evidenced in Fig. 5 it emerges that the inclusion in a particular series with special characteristics (review process, etc.) is always considered the most relevant indicator by the scholars of all areas of law. Another indicator of maximum importance for all scholars is the availability of monographs in the catalogues of national and international libraries. The other three indicators present in the top five of the general ranking are not of equal importance so that they are not always present in the top positions of the ranking for single areas of law. In particular, for example, reviews in scientific journals and reporting in scientific journals shift from the second and third positions in the ranking for the area of Roman law to the seventh and the tenth positions in the ranking for the area of criminal procedure.

In many areas of law (private comparative law, navigation and air law, constitutional law, tax law, European Union law, criminal procedure law, comparative public law) indexing in citation databases specifically created for the legal domain is considered particularly important and is included in the first five positions of the ranking. Indexing in existing citation databases (such as Web of Science Scopus, SSRN) receives less attention. As shown in Fig. 6 this indicator is placed in the top positions of the ranking only by scholars of a specific area of law (navigation and air law).

Based on the rankings of each area of law, different positions on indicators are recorded with respect to the overall average on indicators. This is the empirical confirmation of how the research assessment in legal domain must necessarily take

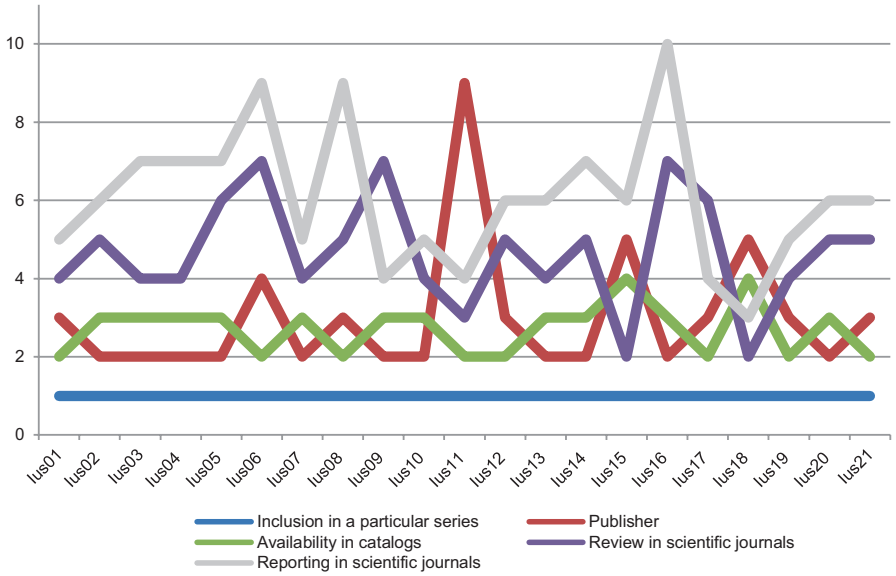


Fig. 5 Position of top five quality indicators (Table 5) according to each area of law identified by the codes listed in Table 1

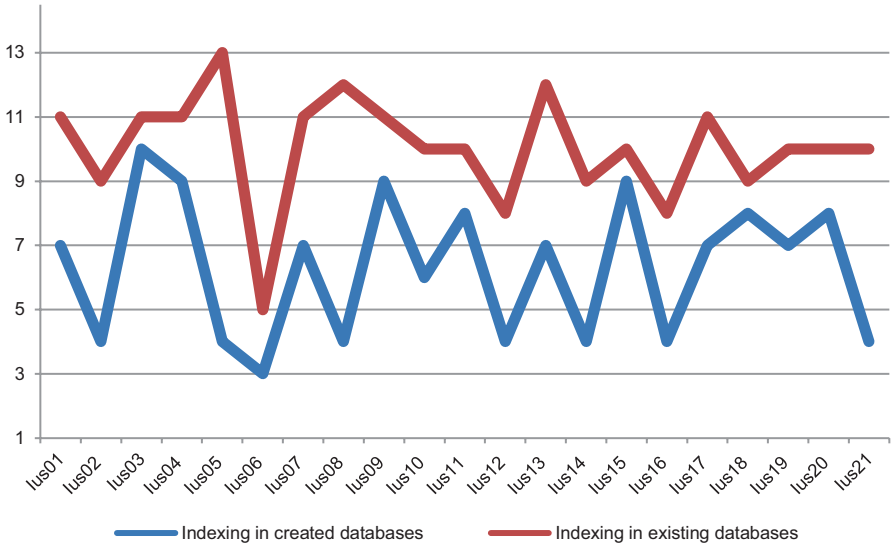


Fig. 6 Position in the ranking of the two indicators "Indexing in existing citation databases" and "Indexing in citation databases specifically created for the legal domain" according to each area of law identified by the codes listed in Table 1.

Table 6 First five positions in the ranking of indicators according to three different areas of law with reference to quality

Ranking	Comparative private law	Roman law	Criminal procedure
1	Inclusion in a particular series with special characteristics (review process…)	Inclusion in a particular series with special characteristics (review process…)	Inclusion in a particular series with special characteristics (review process…)
2	Publisher	Reviews in scientific journals	Publisher
3	Availability in the catalogues of national and international libraries	Reporting in scientific journals	Availability in the catalogues of national and international libraries
4	Presence in citation databases created specifically for the legal domain	Availability in the catalogues of national and international libraries	Presence in citation databases created specifically for the legal domain
5	Reviews in scientific journals	Publisher	Reviews in specific scientific journals (only highly ranked journals)

into account the fact that law is not monolithic: there are disciplines with very different characteristics affecting not only the methodology of legal research but also the methods of their evaluation. Just as an example, Table 6 shows the first five positions of three different areas of law.

Finally, with regard to the knowledge declared by respondents of bibliometrics, those who declared a deep knowledge only partially departed from the general ranking by giving particular weight to indicators such as indexing in citation database specifically created for legal domain and the number of downloads in case the monograph is available online.

4.3 Measuring Quality: Approaches for Legal Evaluation

On the basis of the data analysis, three main aspects have emerged that can provide the foundation for a constructive and stimulating debate that can also extend beyond national borders.

1. The role of peer review, with particular reference to the review process of the manuscripts before their publication in a specific series. Respondents were asked to indicate if the peer review process declared by a series is actually performed. 23% of respondents believe that the review process is always carried out, while 55% think that it is conducted in most cases. About 20% have little confidence that the process is actually performed. Some comments show critical observations on how the procedure is performed (e.g., in some scientific areas where the number of experts is low it is difficult to guarantee the anonymity of the author or of reviewers) and on the absence of transparency of procedures (e.g., lack of

information about the percentage of works not accepted or the percentage of requests for changes or additions requested by reviewers).
2. The usefulness of a ranking of publishers and of series to support the purposes of the evaluation of monographs. Almost 80% of respondents said they did not find it useful. In some comments specific reasons in support of this negative position are indicated: the belief that publishers often publish books without an effective evaluation of the work; the practical difficulty of defining a possible ranking according to fair and objective criteria; the risk of crystallising the market in favour of publishers who would remain in the top positions of the ranking; the idea that the value of the work alone should be assessed.

In particular, with regard to these two types of rankings it was significant to verify if those who gave a negative opinion on the opportunity of a formal ranking of publishers and series then actually implicitly admitted the existence of different weights of publishers and series. For this purpose a comparison was made of the answers to questions about the usefulness of the rankings and the question on the impact of the indicator "Publisher" and the indicator "Inclusion in a particular series with special characteristics" to be used to support the evaluation of quality.

As shown in Figs. 7 and 8, in a significant number of cases those denying the usefulness of rankings of publishers and/or series in contrast note their belief that the indicator "Publisher" and the indicator "Inclusion in a particular series with special characteristics" are fairly or very relevant for the assessment of the quality of monographs, and then, as a matter of fact, implicitly recognise the existence of rankings of publishers and of series but do not want them (or do not consider that it is possible for them) to be formalised. To compare the answers to questions about the usefulness of these two rankings in which respondents have been asked to choose only between "yes" and "no", with those on the impact of the indicators, "high" and "fair" have been considered as a positive (yes) answer and "low" and "none" as a negative (no) answer.

3. A final issue concerns the usefulness of a citation database that includes legal monographs. In this case, 52% of respondents believe that citation databases are key tools to support the evaluation process. In the comments of respondents who do not agree on the use of these kinds of databases, doubts related to the special significance of citations in the field of legal research have been outlined. Legal scholars mainly cite everything that exists on a certain topic and not only that which is considered useful or interesting by the author.

4.4 Perceived Risks Associated with the Use of Indicators

According to the questionnaire, respondents consider the risks usually associated with the use of indicators quite realistic. Given this response, it is not surprising that the Italian debate is characterised by a clear scepticism about the use of indicators to support the quality assessment of monographs.A large majority of respondents

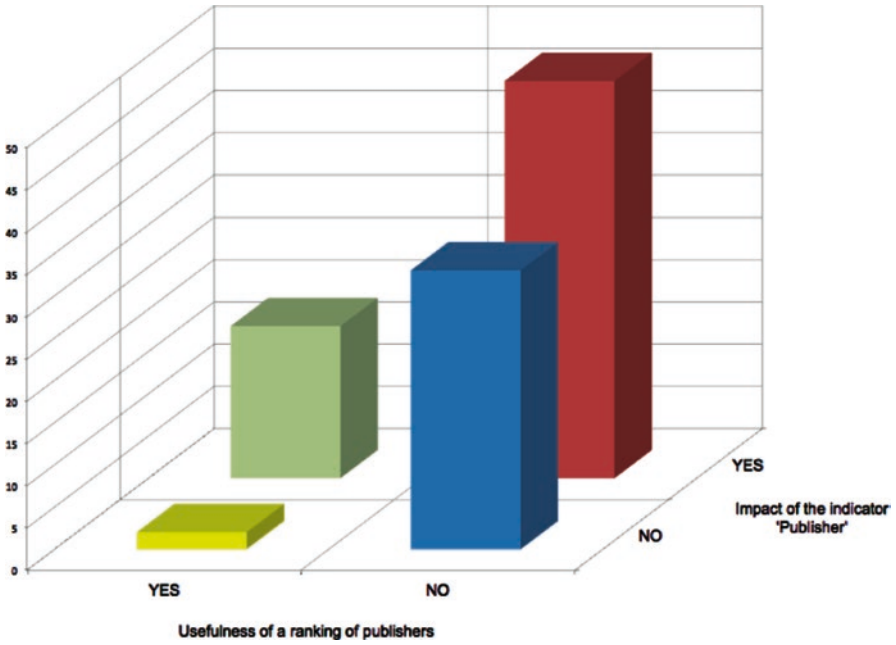


Fig. 7 Comparison between the answers to questions about the usefulness of a ranking of publishers and answers on the impact of the indicator "Publisher" with reference to quality

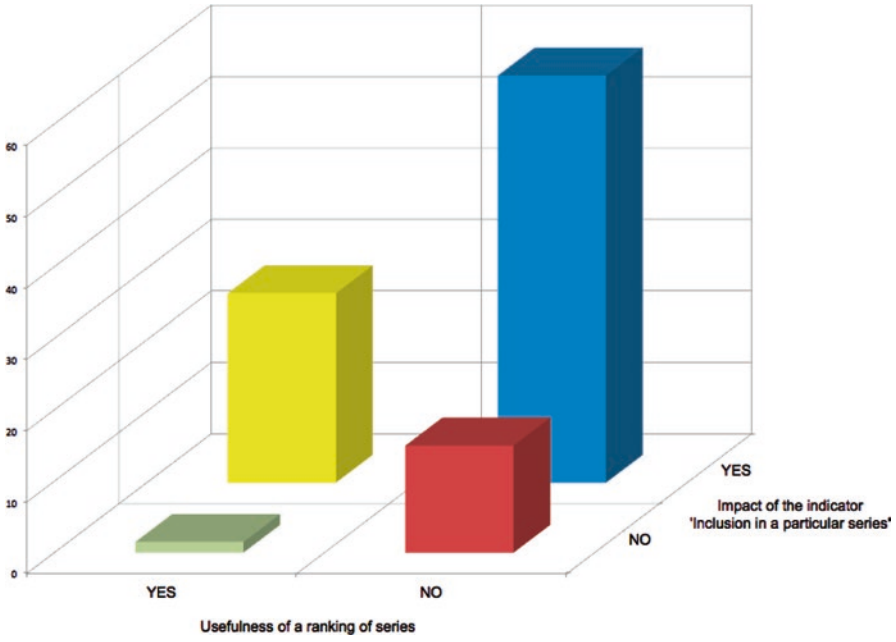


Fig. 8 Comparison between the answers to questions about the usefulness of a ranking of series and answers on the impact of the indicator "Inclusion in a particular series with special characteristics" with reference to quality

believe that the use of indicators produces negative effects on the production of research outputs. Some of these risks are normally the following (Van Gestel and Micklitz 2011a, b; Van Gestel and Micklitz 2014; Conte 2015b) (i) "salami tactics" (fragmentation of scientific contributions to the least publishable unit in order to produce a large number of publications); (ii) the privilege of a language and of a certain publishing environment (publishers or/and series); (iii) citation strategies; (iv) the choice of "trendy" research subjects.

In particular, the greatest risk is represented by citation strategies: according to respondents' comments, their view is that citations are made not because of the outputs merits but just to increase the score of research groups and of schools of interpretation. Many comments were also dedicated to the risk associated with the privilege of a language: the effect is that scholars choose a common language like English just because it allows research outputs to be more cited.

5 Conclusion and Prospects

The ITTIG survey described provides a framework that empirically confirms the state of the Italian debate on the evaluation of scientific legal monographs. Italian legal scholars have assessed the quality of their scientific production using qualitative methods rather than bibliometric. Under the influence of the exact sciences, they are now encouraged to follow the evaluation model used for such sciences. However, this process is not obvious. It is possible to affirm, based upon the answers to the national questionnaire and on the basis of the Italian and foreign literature, that the research quality assessment in legal sciences must take into account a number of important factors.

In the field of law, monographs are definitely works with greater scope. They have a high scientific value that is often – or is conceived as being – the result of several years of study on a specific topic, not necessarily innovative, but certainly important as a point of reference on the subject. The work should include sound proposals of interpretation on issues involved. The discussion should be structured, systematically and consistently focused on a well-defined unitary theme: the monographic nature is provided by the logical concatenation of chapters, which cannot rely on a mere "similarity of topics". Concerning the assessment of monographs, the central role (if not exclusive) of peer review is undeniable. However, measures should be taken to avoid the risk of subjectivity of judgement. Indicators are certainly useful to orient the evaluation but cannot be strictly applied as the result of mere mathematical and statistical analysis, easily exposed to erroneous and arbitrary use.

The formal criteria for the evaluation of monographs, let alone an independent analysis of the content, tend to rely on traditional elements such as the publisher, the inclusion in editorial series (and their characteristics in terms of ex-ante evaluation), reviews, awards, dissemination within the scientific community, frequency of citations. Indicators must indubitably be applied retrospectively, therefore great caution is required as regards their definition and application.

Passing successfully through a peer review and having writings published in series or editorial spaces used by the scientific community for debate are constituent elements of a "scientific" work. Nevertheless, the weight of the container (editorial series, publishers) should not be emphasised. Instead, the debate should focus on the transparency of procedures in the selection and review of research outputs.

The societal impact of legal monographs remains a barely addressed issue. According to the Italian respondents, there is no real societal impact of legal monographs. However, some monographs may indeed influence judges or lawmakers' opinion, mainly by virtue of the authors' prestige and their argumentative capacity.

The use of different languages and the value of international cooperation must not be held as an indicator of quality, but must hold a very important place in certain areas of law. On the other hand, it should be noted that in legal research the intensification of this trend (encouraged, for example, by the use of criteria that reward the number of international citations) could lead to paradoxical and pernicious effects. It is unacceptable for a monograph be penalised for its in-depth analysis of the interpretation and application of rules presented in the language used for their application.

The lack of qualification of reviewers is a critical issue that must be addressed both at a national and international level. In Italy, for example, it is very important to make a selection of peers based on their research background and competence with respect to candidates, rather than the affiliation of experts to a specific area of law. Adapting methods used in other sciences (including the social sciences) to evaluate legal research does not seem to be a good solution. Rather, it would be appropriate to take actions at an international level to create transnational standards for assessing the quality of legal research. This is in line with the internationalisation of research, which is underway in this area, and the increasing mobility of students, courses and international law schools. The establishment of harmonised standards or generally accepted quality indicators allowing jurists to work in different legal systems is a challenge to be conquered despite the differences among national evaluation methods, different publishing cultures and the many academic traditions.

In conclusion, it appears that in Italy peer review is a well-established methodology, but the "popularity" of bibliometrics is gaining. However, serious shortcomings of both peer review and use of metrics have been observed. Is it necessary to choose alternatives, such as stronger focus on the methodological justification for legal research, better clarity and transparency of editorial boards in the application of quality criteria for approving or rejecting research outputs, and finally a serious and fruitful European debate on the benefits and disadvantages of the various research evaluation systems, rather than letting each country reinvent its strategy (van Gestel 2011). Of course, these brief considerations are placed in a cultural context of ceaseless change, with the consequent need for constant repositioning of this study, future surveys and experiments in the field of evaluation of legal science, in particular with regard to the opportunity to agree on a common approach beyond national legal boundaries.

References

Baccini, A. (2010). *Valutare la ricerca scientifica. Uso e abuso degli indicatori bibliometrici*. Bologna: Il Mulino.

Carinci, F., & Brollo, M. (Eds.). (2013). *Abilitazione scientifica per i professori universitari: Legge n. 240/2010 e decretazione attuativa*. Milan: IPSOA.

CASAG- Conferenza italiana delle associazioni scientifiche di area giuridica. (2014). Parere sull'uso degli indicatori bibliometrici per la valutazione in ambito giuridico. http://www.associazione-deicostituzionalisti.it/download/Jvy0AePS84owLDwij3bSn07t0n6ZcqEYKVqDd4Ljuxk/casagparere-indicatori-bibliometrici-a.pdf

Chodorow, S. (1999). *Specialized scholarly monograph in crisis or how can I get tenure if you won't publish my book?* Washington, DC: Association of Research Libraries.

Consiglio Universitario Nazionale. (2013). Proposta 'Criteri identificanti il carattere scientifico delle pubblicazioni e degli altri prodotti della ricerca', Prot. 22011 dated 24/10/2013. https://www.cun.it/uploads/4532/proposta_cun_criteri_scientificit%C3%A0.pdf?

Conte, G. (Ed.). (2015a). *Evoluzione e valutazione della ricerca giuridica*. Naples: ESI.

Conte, G. (2015b). Le reciproche interferenze tra evoluzione della ricerca giuridica e valutazione della medesima. In G. Conte (Ed.), *Evoluzione e valutazione della ricerca giuridica* (pp. 9–16). Naples: ESI.

Di Raimo, R. (2015). la valutazione delle monografie nelle scienze sociali, secondo il sistema attuale. In G. Conte (Ed.), *Evoluzione e valutazione della ricerca giuridica* (pp. 241–252). Naples: ESI.

Hicks, D. (2004). The four literatures of social science. In F. Moed et al. (Eds.), Handbook of quantitative scienceand technology research (pp. 473–496). Kluwer Academic Publishers.

Osservatorio della Ricerca dell'Alma Mater Studiorum – Università di Bologna (OR). (2013). Definizione e principali criteri di valutazione dei prodotti della ricerca, giugno.

Thompson, J. W. (2002). The death of the scholarly monograph in the humanities? Citation patterns in literary scholarship. *Libri, 52*(3), 121–136.

Thompson, J. (2005). *Books in the digital age: The transformation of academic and higher education publishing in Britain and the United States*. Cambridge/Malden: Polity Press.

van Gestel, R., & Micklitz, H.-W. (2011a). Revitalising doctrinal legal research in Europe: What about methodology? In U. Neergard & R. Nielsen (Eds.), *European legal method-paradoxes and revitalisation*. Copenhagen: Djof Publishing.

van Gestel, R., & Micklitz, H.-W. (2011b). Why methods matter in European legal scholarship. *European Law Journal, 20*(3), 292–316.

van Gestel, R., & Vranken, J. (2011). Assessing legal research: Sense and nonsense of peer review versus bibliometrics and the need for a European approach. *German Law Journal, 12*(3), 901–929.

Watkinson, A. (2001). *Electronic solutions to the problems of monograph publishing*. London: The Council for Museums, Archives and Libraries.

Part III
Journal Classification and Rating

Publication Strategies in SSH: Empirical Evidence and Policy Implications

Domenica Fioredistella Iezzi

1 Introduction

In past decades, universities have undergone institutional changes that have significantly affected their objectives, sources of funding and financing methods, as well as their mode of action. A relevant change is evident in the increase in the share of university budget represented by competitive grants and third-party funding, as opposed to the traditional and institutional block grant funding (Bonaccorsi et al. 2014). United Kingdom, Spain, Slovak Republic, Poland, Portugal, Italy, Belgium (Flemish Community), Norway, Sweden, Denmark and Finland in Europe, as well as Hong Kong, China, Australia and New Zealand conduct regular comparative performance evaluations of universities and use results in performance-based funding schemes. Research is evaluated in many places, by many types of organisations, for many different purposes (Hicks 2012). In the science and technology disciplines (STEM) there is a long tradition of quantitative analysis based on the bibliometric approach, whereas in social sciences and humanities (SSH) this evaluative practice is not widespread. Bibliometric indicators have increasingly become of interest because they can be helpful in providing some measurement of visibility of scientific publications, as well as in evaluating research success and the impact on the scientific community (Alves et al. 2016).

In Italy, on 16 July 2013, the Ministry of Education, University and Research (MIUR) closed the second research evaluation exercise (VQR, or Research Quality Evaluation 2004–2010), with the publication of university performance ranking lists. The results determined the allocation of an important share of funding for each institution. The 2004–2010 VQR is aimed at evaluating the results of scientific research of both public and private universities carried out during the period ranging

D.F. Iezzi (✉)
Dip. Ingegneria dell'Impresa "Mario Lucertini", Università di Roma Tor Vergata, Rome, Italy
e-mail: stella.iezzi@uniroma2.it

© Springer International Publishing AG 2018

A. Bonaccorsi (ed.), *The Evaluation of Research in Social Sciences and Humanities*, https://doi.org/10.1007/978-3-319-68554-0_10

from 2004 to 2010. The VQR was formalised with the approval of Ministerial Decree no. 17 on 15 July 2011.

From 2004 to 2010, Italian universities and other public research bodies submitted almost 185,000 articles, books, patents, and other scientific products. In most cases, scientific publications were peer reviewed; however, bibliometric indicators were also used in hard sciences, medicine, engineering, and economics (Cicero et al. 2013). The VQR was based on a hybrid model (bibliometric and peer review approach). The MIUR entrusted the implementation of the VQR to the Agency for the Evaluation of University and Research Systems (ANVUR), which commenced the evaluation process at the end of 2011.

The goal of this article is to analyse the strategies of publication in SSH. The research questions are: (1) Do researchers in SSH have a strategy in the choice of dissemination channels for their research products? (2) Does the strategy (if any) depend on the disciplinary area? (3) Which factors explain the choices of publication channels?

For this purpose, we analyse the scientific production in the SSH area submitted to the VQR 2004–2010, as well as the journal articles collected in the Loginmiur database managed by the University Consortium Cineca in the period 2004–2013. As in other chapters in this volume, we examine five macro-areas as defined by the National University Council (CUN): Area 10 (Antiquities, Philology, Literary Studies, Art History), Area 11 (History, Philosophy, Pedagogy and Psychology), Area 12 (Law Studies), Area 13 (Economics and Statistics), and Area 14 (Political and Social Sciences).

This paper is structured as follows. Section 2 (Bibliometric versus peer review methods for research evaluation in SSH) discusses the literature in social sciences and humanities about the use of instruments of bibliometric and/or the peer review approach. Section 3 (The main results of 2004–2010 VQR) illustrates the results of the VQR exercise. Section 4 (Discussions and conclusions) compares the results of VQR across disciplinary areas and examines the journal articles published by scholars in SSH from 2004 to 2013 in DB CINECA, suggesting implications for future evaluation projects.

2 Bibliometric Versus Peer Review Methods for Research Evaluation in SSH

The measurement approach in bibliometrics makes use of *ad hoc* indicators. The adoption of a number of *conventions of equivalence relation* has permitted to achieve an agreed definition of entities such as “journal article”, “author”, “affiliation”, “citation”, or “subject category”. These conventions do not allow evaluation of scientific books because these are heterogeneous products. The bibliometric approach was born in the mid-1900s in the United States on the initiative of the Institute for Scientific Information (ISI) founded by Eugene Garfield, father of the well-known Impact Factor (IF). Garfield, however, had at the time recommended

the use of bibliometric indices primarily with the aim of orienting purchase decisions of libraries (WoS 2016).

The supporters of the use of bibliometrics for the evaluation of research products believe that this is the only "objective" procedure for classifying a journal (the container), from which can then be deduced, by transitive property, the assessment of an article (the content), with obvious advantages in terms of elimination of any biased external interventions, and reduction of the time and cost of the evaluation.

Among the unintended negative consequences of the adoption of bibliometrics, the literature often mentions the reduction of academic freedom in the choice of research topics and in publishing strategies, the criminalisation of interdisciplinary studies, the marginalisation of topics of niche or national interest. These problems are amplified in SSH, because many products of research are published in journals without an Impact Factor. According to updated data at the beginning of 2015 from the loginmiur database managed by Cineca, the top 10 articles written by Italian statisticians from 2002 to 2014 (about one-fifth of all articles published in the period), have been published in 5 Italian journals without IF.

It is also worth mentioning that for several years the evaluation of research organisations in the UK and Australia have openly abandoned the use of pure bibliometrics, which has also been criticised in the Declaration On Research Assessment sponsored by the American Society for Cell Biology (ASCB 2012) and signed by a large group of researchers, organisations and publishers of scientific journals of international standing. More recently, *Nature* published the Leiden Manifesto for Research Metrics (NAT 2015).

Given the problems of the bibliometric approach, in recent years it has been developed an approach that many experts consider the most promising. While this model does not exclude the rational use of bibliometric information, it relies on a large and regular consultation of experts to also consider the reputation of the journals within the relevant scientific community: there are mentioned in particular the experience of France for the area of Economics and Management with the National Committee of Scientific Research (CNRS 2016), the three Nordic countries (Denmark, Finland and Norway) with the Norwegian Centre for Research Date (NSD 2016) and Australia with the Australian Research Council (ARC 2015). At the international level, EvalHum is a network that allows a stimulating comparison between European scholars in SSH.[1] It is registered as an association in France under the French 1901 law.

Peer-review methodologies are based on the assessment of research outputs of research organisations, by panels of assessors selected by the authority presiding over the assessment. In this approach, the judgement of quality depend on experts. Time and cost of assessment exercises are also a critical element. The methodology, furthermore, presents no universality, as the mechanisms of appraisal are defined independently by the assessing panel, and are therefore open to possible distortions (Abramo and D'Angelo 2014).

[1] See http://www.evalhum.eu/

3 The Main Results of the VQR 2004–2010

The 2004–2010 VQR is divided into 14 research areas, as defined by CUN. For each of them, ANVUR has identified a group of experts (GEV) and its president. The composition of each GEV, the curriculum vitae of the president, and the approved evaluation criteria of GEV were published on the web site of ANVUR, section VQR. The quality assessment is based on the criteria published in Ministerial Degree no. 17 on 15 July 2011.

As discussed elsewhere in this volume, three research quality criteria were defined as important: (a) relevance, to be understood as added value for the advancement of knowledge in the field and for science in general, even in terms of appropriateness, effectiveness, timeliness and duration of impact; (b) originality/innovation, intended as a contribution to the advancement of knowledge and/or new acquisitions in the field; (c) internationalisation and/or international competitive potential, intended as positioning in the international arena in terms of relevance, competitiveness, editorial dissemination and appreciation of the scientific community, including explicit collaboration with researchers and research teams from other countries.

Moreover, judgements expressed by referees are summarised in a score from 1 to −2 (Table 1). The GEV together with ANVUR defined the criteria used for the evaluation. The products are not evaluable if they belong to types excluded from the list of admissible products or if the documentation provided is inadequate. There is a penalty for confirmed cases of plagiarism or fraud. The list of scientific journals, published on the ANVUR website, is the result of the work of their respective GEV and have been approved by the executive board after consultation with the scientific societies, journal editors, and scholars. This list is not a necessary condition for calculating bibliometric indicators: the Italian journals in SSH present in the ISI and Scopus databases are still few and the database offered by Google Scholar is not entirely reliable. For this reason, ANVUR considered it necessary to establish an *ad hoc* working group in order to develop criteria, parameters and methods of action for the future. The rating of journals built in the last exercise of the 2004–2010 VQR had a significant impact in the following years on the whole Italian scientific community. In fact, the results of VQR have been used, in addition to the mentioned performance-based funding scheme at the level of universities, for the evaluation of the composition of the teaching colleges of PhD programmes. Journal ratings have

Table 1 Judgements associated with scores expressed by the peer review

Judgement	Score
Excellent	1.0
Good	0.8
Fair	0.5
Poor	0.0
Not evaluable	−1.0
Fraud or plagiarism	−2.0

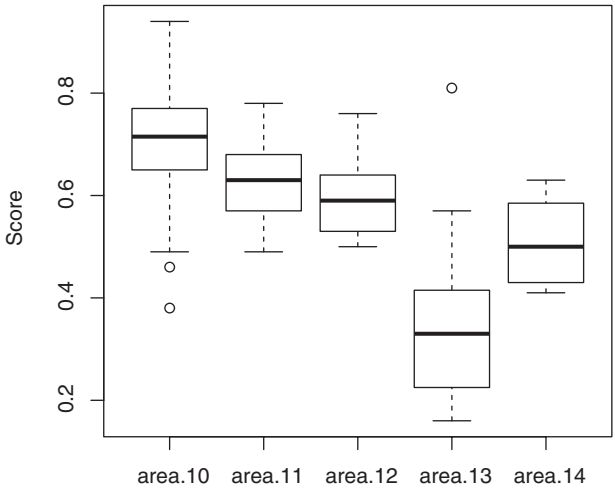


Fig. 1 Score by CUN AREA in SSH

been used as the basis for the calculation of indicators for National Scientific Habilitation. In some universities, although this use has been strongly discouraged by the Agency itself, VQR data have also been used for the allocation of individual research funds. In some cases the procedures decided by Evaluators Expert Groups (GEV) have been criticised (see for example for the Area 13 Economics and Statistics: SIS 2014).[2]

The new VQR 2011–2014 is currently in progress and the results will be delivered by the end of 2016. Also, the new VQR lists of journals were created using indexes derived from bibliometrics, so it is possible to automate the processes of journal classification. Let us examine more closely the results of VQR in the SSH areas.

Boxplots of scores in SSH show very different patterns. In the areas 10, 11, 12, and 14 the range of scores is from 0.5 to 1, the median value is from 0.5 to 0.7 (overall, a good judgement). In the area 13 the range is from zero to 0.5, indicating that at least half of the scholars submitted poor production. In the area 10 the values 0.38 ("Portuguese and Brazilian literatures"), and 0.48 ("Portuguese and Brazilian translations") are outliers; in the area 13 0.81 ("Econometrics") is an outlier.

These differences are amplified when we look at the average score by discipline and by product types (Table 2). In CUN areas 10 and 11, the highest scores are assigned to the monograph or scientific treatise (monographs received have an average score of 0.72 and 0.64, respectively); in area 12, the highest score is assigned to chapters in book with an average value of 0.63; in area 13, the journal article has an

[2] See website: SIS (2014. Available at www.new.sis-statistica.org/wp-content/uploads/2014/12/statsoc-3_2014.pdf

Table 2 Product types (average) by CUN areas 10–14

	Area 10	Area 11	Area 12	Area 13	Area 14
Journal article	0.70	0.62	0.59	0.51	0.51
Monograph	0.72	0.64	0.56	0.11	0.51
Edited book	0.63	0.55	0.33	0.04	0.51
Book chapter	0.69	0.58	0.63	0.13	0.41
Proceedings	0.66	0.59	0.26	0.07	0.38
Others[a]	0.65	0.43	0.56	−0.13	0.22

Source: Based on DB ANVUR 2014 (97%)

[a]This category includes: scientific review; composition; design; critical exposure; exposure; software; show; artefact; note to judgement; preface; afterword; art prototype and related projects; publication of unpublished sources; translation; voice (dictionary or encyclopaedia); correlation; performance.

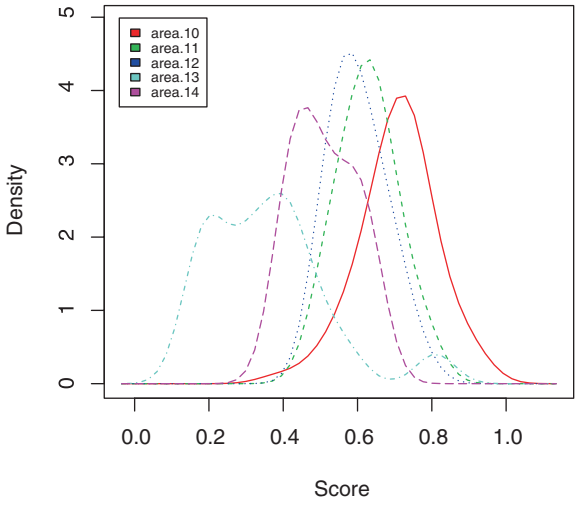


Fig. 2 Distributions of the scores by CUN AREA (Source: Based on DB ANVUR 2014 (97%))

average score of 0.51; in area 14, the journal article, the monograph and the edited book receive the highest score at 0.51.

In area 10, there are no marked differences between the assessments of the various products. Consequently, producing a monograph rather than an article or another product has led to an average score above 0.5. On the contrary, in area 13 those who have presented products different from journal articles received a very low average score. Figure 2 shows a general idea of the shape. The areas 13, and 14 have a multimodal distributions with more modes. The area 13 presents three distinct peaks (local maxima) close to 0.2, 0.4, and 0.8; the area 14 two peaks close to 0.4, and 0.6. These peaks highlight the existence of a disparity of scores within the same areas. To compare the variability of scores in the areas 10–14, we calculate the coefficient of variation (CV[3]), a standardised measure of dispersion of frequency distribution.

[3] CV is the ratio of the standard deviation to the mean.

It is clear that variability greatly differs across areas. Taking a closer look at score peculiarities, we examined the distribution by discipline, or Settori Scientifico Disciplinari (SSD). The analysis of the scores, aggregated for 154 scientific disciplines, shows a distribution that differs from the aggregate area.

The SSDs with a CV lower than 20 are: L-OR/14 "Philology, Religions and History of Iran" (CV = 12.8), L-OR/01 (CV = 14.1) "History of the Ancient Near East", L-OR/03 "Assyriology"(CV = 14.4), L-OR/07 "Ethiopian Languages and Literatures" (CV = 18.2), L-OR/13 "Armenian Studies, Caucasology, Mongol Studies and Turkology" (18.6), and L-OR/19 "Modern Languages and Literatures of the Indian subcontinent" (CV = 19.5). They belong to the same CUN AREA (area 10), and the sub-group "Cultures of East and Africa", and represent "small fields which are dying", composed by a minimum of 6 to a maximum of 9 scholars, super specialised in peculiar aspects of African and East culture. Their products are always evaluated with high scores.

The first quartile includes all the SFs with very high scores and low variability. They belong to area no.10, except for Agricultural Law (IUS/03 –area 12), Econometrics (SECS-P/05 – area 13) and Maritime Law (IUS/06 – area 12). In the second quartile, the SAs are assessed with medium scores and low variability, mainly including SAs of the areas 11 and 12. The third quartile includes SAs evaluated with low scores and high variability, primarily SAs of the areas 13 and 14. Figure 3 points out that there are some outliers with CV greater than 130. Table 3 consists in a list of SAs with denomination and CUN area.

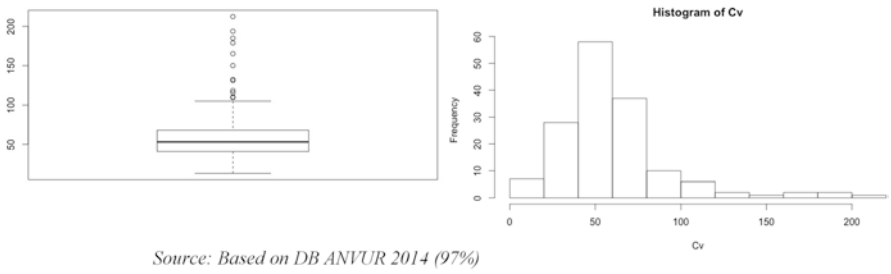


Fig. 3 Distribution of CV scores by CUN area (Source: Based on DB ANVUR 2014 (97%))

Table 3 Outliers in the coefficient of variation by SA

CUN AREA	SA	Denomination	CV
Area 13	SECS-S/03	*Economic Statistics*	131,0
Area 13	SECS-S/04	*Demography*	132,3
Area 13	SECS-P/13	*Commodity Science*	150,0
Area 13	SECS-P/08	*Economics and Business Management*	165,2
Area 13	SECS-S/05	*Social Statistics*	178,9
Area 13	SECS-P/09	*Business Finance*	185,0
Area 13	SECS-P/11	*Financial intermediaries*	193,8
Area 13	SECS-P/07	*Business Administration*	212,5

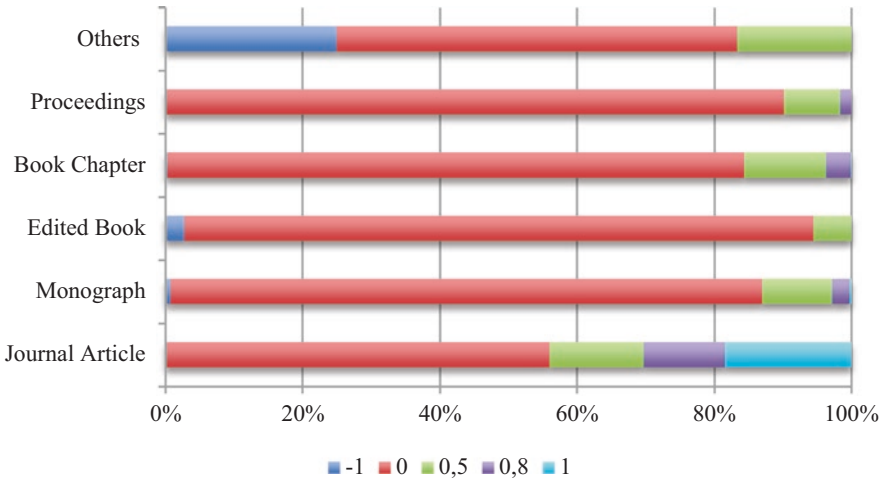


Fig. 4 Score by products for the outlier group

In this class, there are 8 disciplines that belong to area 13, with a very low score and high variability. In those sectors, almost all scholars have low scores (about zero), and few have high grades (from 0.5 and 1).

Figure 4 shows that 56% of the papers have a score equal to zero, 14% of 0.5, 12%, equal to 0.8 and 18%, equal to 1. Those who have submitted a journal article have had a higher average rating than other products. In 87% of cases monographs were evaluated with a score of zero, as were book chapters in 84% of cases. There are statistically significant differences in the scores and the type of product (X-squared = 732.52, df = 9, p-value <2.2e-16). Those who wrote a journal article had more chance of receiving a good evaluation. Assessment of the VQR product is based on three parameters: originality, relevance and internationalisation measured in scores from zero to 9. The distribution of the parameters shows positive skewness[4] in all distributions and a median value of 6. The first quartile is 3 for internationalisation assessment, while a value of 4 for originality and importance (Fig. 5). So at least half of the work has a sufficient evaluation for the three parameters, while the first 25% has a low rating up to 4 or 3 in the case of internationalisation.

These parameters are strongly correlated (Table 4). So, if the work is of good quality for originality, it will be the same value for importance and internationalisation. The average by product type has values close to 5, so there are no large differences between types of publications. The analysis becomes more interesting if the analysis is broken down by CUN area. In area 10, the average rating is between 6.5 and 6.3 for all parameters with a very low standard deviation (SD); in areas 11 and

[4] A distribution is said to be asymmetric if its form does not present symmetry with respect to the central position. An elongated tail to the right characterises positive asymmetry.

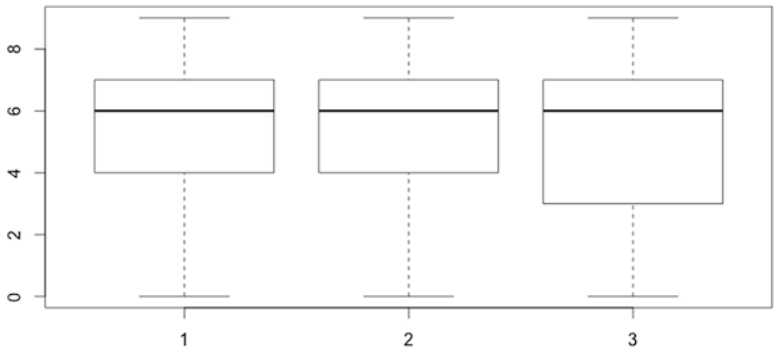


Fig. 5 Distribution of originality, significance, and internationalisation (Legend: 1 = originality, 2 = significance; 3 = internationalization)

Table 4 Correlation matrix of originality, significance, and internationalisation

	Originality	Significance	Internationalisation
Originality	1	0.92	0.87
Significance	0.92	1	0.89
Internationalisation	0.87	0.89	1

Table 5 Mean and Standard deviation of originality, significance, and internationalisation by CUN AREA in SSH

CUN AREA	Originality	Significance	Internationalisation
Area 10	6.53 (SD 0.09)	6.34 (SD 0.14)	6.34 (SD 0.11)
Area 11	6.40 (SD 0.14)	6.10 (SD 0.16)	5.82 (SD0.12)
Area 12	6.10 (SD 0.16)	5.99 (SD 0.28)	5.64 (SD 0.34)
Area 13	4.32 (SD 0.55)	3.81 (SD 0.49)	3.22 (SD 0.52)
Area 14	1.88 (SD 0.12)	1.45 (SD 0.18)	1.77 (SD 0.12)

12, the rating decreases to around 6 and less than 0.35 SD; in area 13 it is between 3 and 4 (with SD less than 0.5); in area 14 it is always less than 2, and the SD less than 0.20 (Table 5).

The average ratings are independent of the type of product, while there is a strong relationship with the CUN area.

Area 14 has evaluated all the products in a very severe manner, but the overall evaluation is on average 0.5, therefore, higher than that of area 13. So in all areas there is a strong correlation between overall score and the three parameters, except for area 14.

4 Do We See a Strategy in Selecting the Type of Publication? An Analysis of VQR Research Products and of Journal Articles

4.1 In Search of Patterns of Publication Types in VQR Data: A Multivariate Analysis

In this process, SSH areas submitted for the VQR 49,712 scientific products, distributed in area 10 (13,737 products from 76 disciplines), area 11 (9185 from 24), area 12 (11,565 from 21), area 13 (11,272 from 19), area 14 (3953 from 14). Each discipline has its own specificity not only for the content but also for the number of scholars. The smallest group has six scholars, the largest one several hundred. It is therefore clear that small groups, working on niche topics, find it very difficult to even find a reviewer who has the skills to be able to assess. Being able to find the criteria to standardise very different situations in an objective way is challenging.

Table 6 shows the number of products by type (scientific paper, monograph, curatorship, chapters in a book, article published in proceedings, other products such as patents, maps, software) selected for the evaluation (VQR 2004–2010). For all areas, except for area 13, articles, monographs, and chapters or contributions in a book are the most important products. In area 13, articles are the dominant product.

Table 7 shows the percentage distribution for SSH areas by products. We can see that in the various areas, except area 13 where the production is concentrated in

Table 6 VQR 2004–2010. Number of products submitted by type. CUN areas 10–14

CUN AREA	area 10	area 11	area 12	area 13	area 14
Article	3608	2193	3786	6943	1110
Monograph	3102	3021	2988	1451	1344
Edited book	241	167	86	70	115
Chap. Book	4547	3050	4172	2306	1300
Proceedings	1960	648	359	459	70
Other	279	106	174	43	14

Source: Our elaboration on DB ANVUR 2014 (97%)

Table 7 VQR 2004–2010. Percentage of products submitted by type. CUN areas 10–14

Area	area 10	area 11	area 12	area 13	area 14
Article	26.26	23.88	32.74	61.6	28.08
Monograph	22.58	32.89	25.84	12.87	34.00
Edited book	1.75	1.82	0.74	0.62	2.91
Chap. Book	33.10	33.21	36.07	20.46	32.89
Proceedings	14.27	7.05	3.10	4.07	1.77
Other	2.03	1.15	1.50	0.38	0.35

Source: Our elaboration on DB ANVUR 2014 (97%)

articles published in journals (61.6%), there is high similarity. In fact, the production is focused in a balanced manner on three main products: papers in journals, monographs or scientific treatises and chapters and/or contributions in a book. The curatorship, proceedings or "other" have a residual weight. The shares of the three main products are roughly similar across the disciplines.

Data show that in area 10 young scholars choose as a flagship product the monograph (31.7%), while other scholars prefer book chapters. A similar pattern is found in area 11. In area 12, young people mostly publish in journals, while scholars older than 40 spread their research through journals but also in book chapters. In area 13 the pattern is completely different: the article is for all age categories the most important way to disseminate research results. In area 14, half of young people focus their activities on scientific treatises or monographs (48.3%) and about one third on articles (27,1%). The group 41–55 distribute, fairly evenly, their production: about a third on monographs (32,8%), articles (31,0%) and book chapters (31.2%). The more mature show a slight propensity to publish in book chapters (37.3%), but they also consider monographs to be an important research product (32,3%).

As regards the collaboration with foreign co-authors, areas 10, 11, 12 and 14 have a low propensity to publish with non-Italian colleagues. International collaborations merely refer to edited books. Once again area 13 shows special features, since contributions by scholars who work abroad are abundant on all products: 17.2% for journal articles, 11.4% for edited books, 7.0% for proceedings, 5.2% for book chapters.

To display the associations between the product type in the different areas and the language used for the publication, we have applied the Multiple Correspondence Analysis (MCA) (Blasius and Greenacre 1994; Greenacre 2007).[5] We utilised the first two latent dimensions for the construction of the factorial map that explain 91.5% of the inertia (Fig. 6). The first dimension summarises 81.1% of the variability and represents the most relevant type of publication by area (label PUBB) and language (label LANG). The second axis is the residual type publications, synthesising 10.4% of the inertia. The analysis of these three variables simultaneously highlights the peculiarities of the different areas.

On the left side of the first axis is placed area 13, with the article journals (PUBB journal article) written in English (LANG ENG), while on the right side are the other areas. Areas 11, 12 and 14 are placed close to each other in the lower right quadrant in which monographs, edited books, and book chapters are located, as well as the Italian language. Area 10, despite being in the same area for the main publications (right side of the graph), is characterised by a group of residual other products (upper side of the graph) that are typical of Arts and Humanities (such as scientific translations, critical editions, exhibition catalogues and the like). In fact, area 10 has a particular position close to proceedings and other products in other languages. In

[5] For the implementation of the MCA, PCA and Clustering, we used the R software, library ca (Greenacre and Pardo 2006; Nenadic and Greenacre 2007), and FactoMineR (Husson et al. 2010).

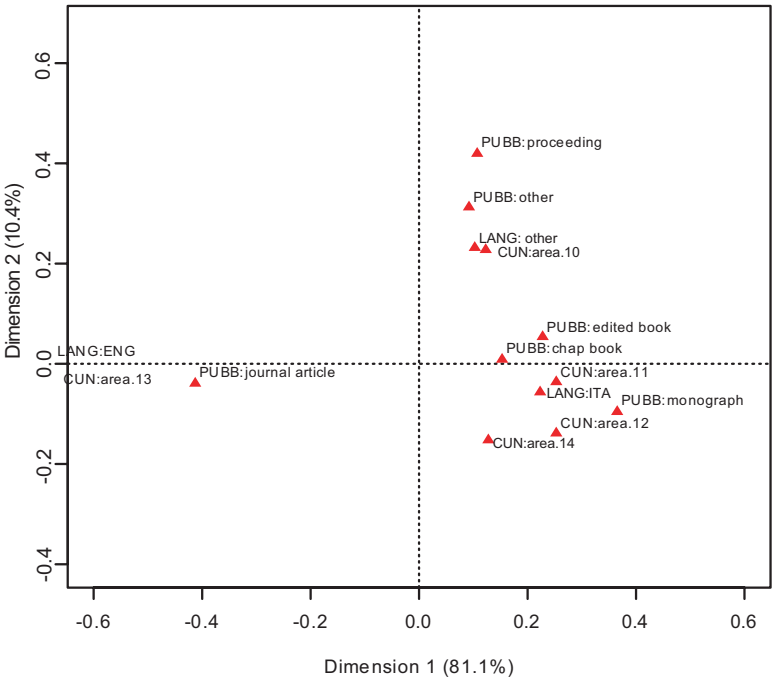


Fig. 6 Multiple Correspondence Analysis (MCA) of publications, language and scientific areas. Plane of the two axes

this area, scholars use many different languages to publish, such as Italian and English, but also Spanish, German, French, Russian, Arabic and many others.

We find a statistically significant relationship between the type of product and the language used (*p-value* < 0.001): English is the preferred language for a journal article, Italian for monographs, and other languages for proceedings. It is interesting to note that the age of the scholar has an impact on the choice of dissemination channels. There is also a statistically significant relationship with the age of scholars across areas (X^2 = 3858.9, df = 8, p-value <0.001). The CUN areas with a higher percentage of young people are area 12 (19.8%) and area 13 (20.1%), while the other areas (10, 11 and 14) have, respectively, 6%, 7% and 8% scholars aged less than 41 years and affiliated with universities with an academic position. In areas 10 and 11, on the contrary, there are more mature scholars: the age exceeds 55 years for 48% and 52% of academics, respectively. Area 14 has a large number of academics between 41 and 55 years old (48.5%).

It is interesting to ask whether different disciplines (at the disaggregated level of SSD, or *Settori Scientifico Disciplinari*, n = 154 in the SSH areas) follow the same publication patterns. We aggregate the data in order to detect latent profiles of each discipline, using Principal Component Analysis (PCA) (Jolliffe, 1986) and cluster analysis (Everitt et al. 2011). The results of PCA for 154 sectors show that there are

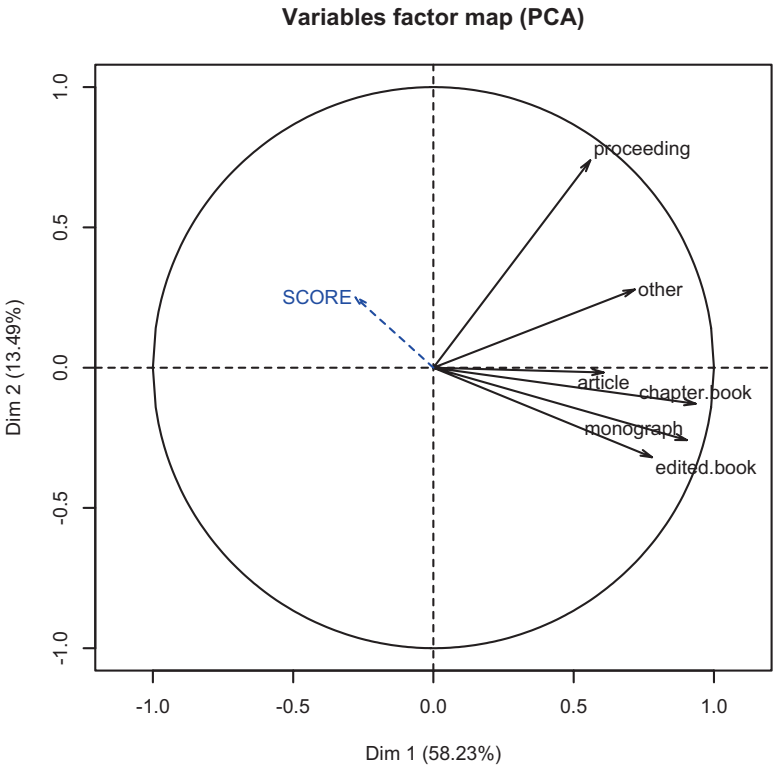


Fig. 7 Correlation circle. Illustrative variable

two latent dimensions: the factor of the publication products and the residual. The first factor explains 58.23% of variability and summarises the analysis for edited book, book chapter, monograph and article, while the second factor, explaining 13.49% of the variability, represents proceedings. The score, which we included as a supplementary variable, has a very low weight (Fig. 7).

We perform an agglomerative hierarchical clustering (Ward method) on the results from the two identified factors. The clustering analysis detects five groups, not corresponding to CUN areas (Fig. 8). Indeed, the CUN areas are located close to the centre of the orthogonal axis.

Figures 9 and 10 show that the cloud of points (representing single disciplines, or SSD) has the shape of a triangle. In correspondence with the vertex coincident with the first axis there is a greater concentration of points, therefore the SSDs belonging to it have similar characteristics: they are very small fields, mainly belonging to area 10. The products that characterise them are article, book chapter, and edited book. Cluster 2 prefers journal article and monograph and the SSDs belong to the areas 10, 12, 11 and 13. Cluster 3 selects journal articles and edited books, and here we find disciplines in areas 10, 12, 13 and 14. Cluster 4 chooses

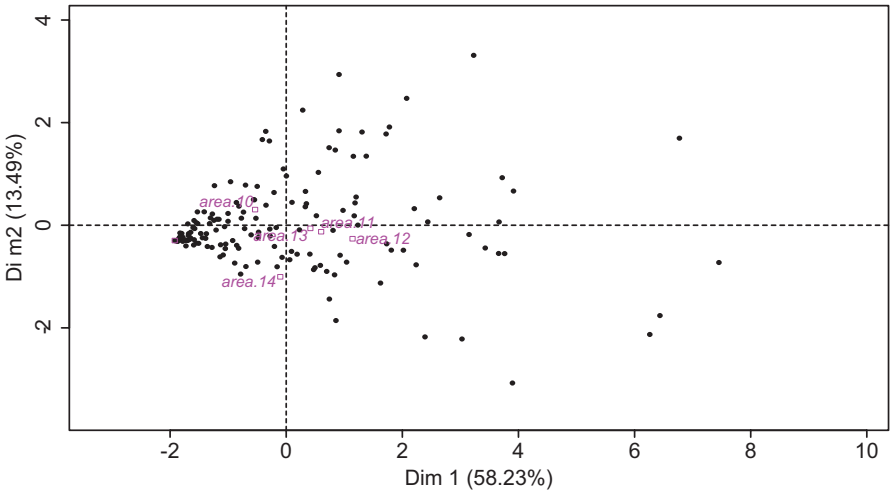


Fig. 8 Plane of the first two axes from the Principal Component Analysis. Illustrative variable: CUN area

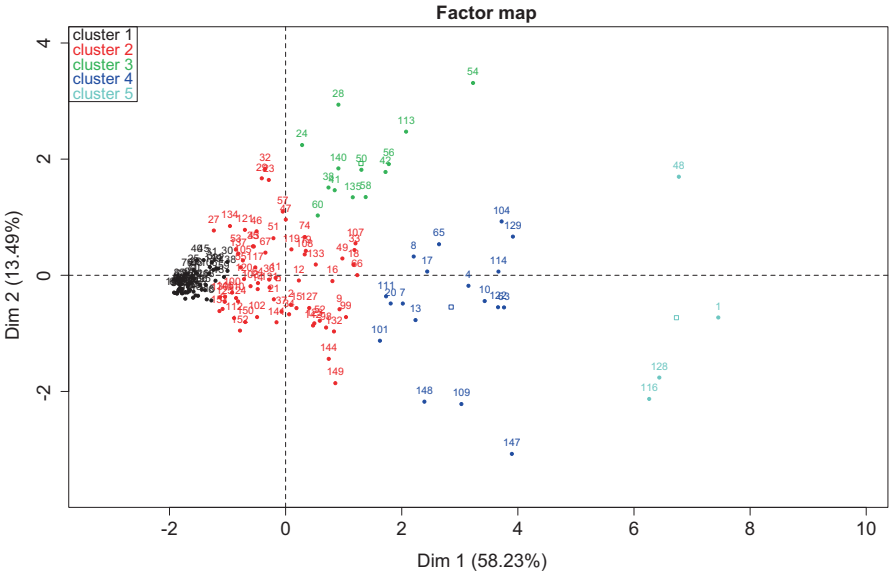


Fig. 9 Plane of the first two axes from the PCA – Individual points labelled by cluster

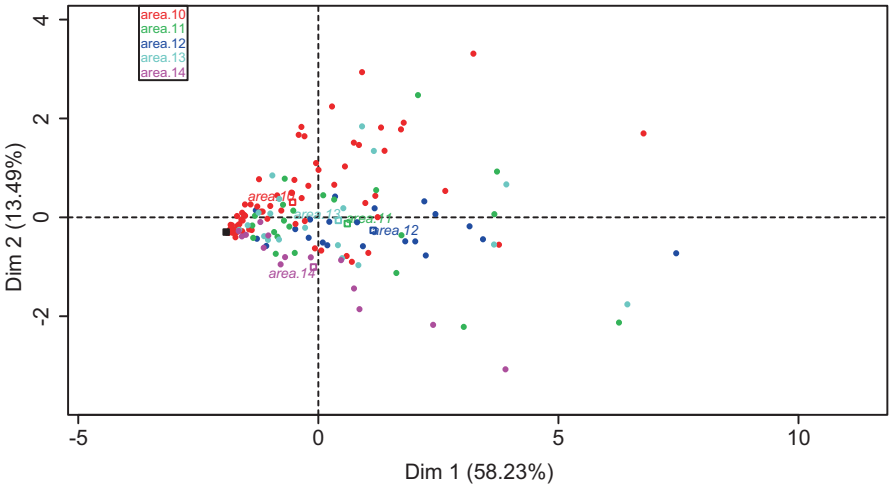


Fig. 10 Plane of the first two axes from the PCA – Individual points labelled by CUN area

articles and chapter books, and the SSDs belong to the areas 10, 12, 13 and 14. Finally, in cluster 5, the scholars indicate only journal articles: the disciplines here are part of areas 10, 11, 13 and 14.

It is quite clear that scientific production in the SSH is not dependent on the CUN area, but the boundaries are to be found elsewhere. For this reason, the search for quality standards is very complex and should be sought in objective criteria to measure the quality of research products. This does not only apply to journal articles, but first of all to books and book chapters. Multivariate statistical analysis can contribute to the study of epistemic differences across disciplines, as discussed in the chapter by Bonaccorsi in this volume. In this section we have used these techniques on data related to types of publications and languages. In the future it would be important to extend to other descriptors of research products.

4.2 In Search of Patterns in Journal Publication. An Analysis of the Self-Reported Production of Scholars in SSH

While in the VQR scholars were asked to select a small number of products (three for university staff, six for staff at PROs), it is clear that scholars produce much more. An interesting opportunity to expand the analysis is offered by the access to the loginmiur database managed by Cineca, made available by Anvur in the context of the research project at the origin of this chapter. Anvur made available a subset of data referring only to the number of articles by name of journal. No individual researcher data were accessible. This is due to current restrictions, until the notion of publicly available national repository (called “Anagrafe della Ricerca” in the

Table 8 Descriptive analysis of journal articles by CUN area. Period 2004–2013

CUN AREA	Mean	Sd	Cv
Area 10	7.62	6.42	84.25
Area 11	8.16	6.16	75.49
Area 12	13.27	12.05	90.81
Area 13	7.87	6.53	82.97
Area 14	7.23	4.31	59.61

Source: Based on DB CINECA

Italian legislation) will be completed and enforced in the legal and administrative framework.

The database cover all scholars affiliated with Italian universities (full professor, associate professor, researcher fixed term, researcher short term). It is important to remind that these data are self-reported. Each person affiliated with an Italian university receives from Cineca the access to a personal site through which are received all official communications about academic procedures from the ministry. Each person creates his/her own personal archive of publications. The categories are imposed by Cineca, but there is no ex-post control based on library procedures. This means that the interpretation of publication categories is left to the author's responsibility. Table 8 shows the average number of articles published in the decade 2004–2013.

In the decade under analysis, scholars in area 10 published 3531 articles in 1001 different journals. The journals with the largest number of publications are:

- L'Indice dei libri del mese (60 papers)
- L'analisi linguistica e letteraria (39)
- Bryn Mawr Classical Review (24)
- Eikasmos (23)
- Iride (21).

In terms of average number, scholars published 7.62 articles in the decade, less than one per year.

In the first quartile (Q1), i.e., the 25% least productive, scholars published 5 articles in 10 years; half of the scholars (median Q2) published 6 articles and the most productive (third quartile, or Q3) 8 articles.

In area 11 we find 720 different journals and 2879 articles, a slightly larger average number per journal (4 articles per journal vs 3 in area 10).

Journals with the highest number of publications are:

- Giornale Italiano di Psicologia (48 papers)
- Psicologia della Salute (25)
- Nuova Secondaria (25)
- Neuroimage (24)
- Archivio Storico Italiano (23)

The inter-quartile distribution is slightly better than the one in area 10: 5 for the least productive, 6 median value, 9 most productive quartile.

The concentration of journals is higher in area 12. In legal studies we find 495 journals, which have published 3301 articles in the decade, or approximately 7 articles each.

The top five journals are as follows:

- Giurisprudenza Italiana (83 papers)
- Foro Italiano (81)
- Giurisprudenza Costituzionale (51)
- Guida al diritto. Il Sole 24 Ore (45)
- Diritto pubblico comparato ed europeo (42).

In this area we find a much larger average value (13.27). The inter-quartile distribution is as follows: only 5 articles for the lower quartile, but 9 for the median and 17 for the most productive quartile.

In area 13 there are 720 journals with 2430 articles (3 on average). With an average of 7.87 articles in the decade, the least productive quartile published as few as 4 articles, the median 6 and the most productive 8 articles.

The journals with the largest number of publications are:

- Rivista Italiana di Economia, Demografia e Statistica (32 papers)
- Economia e Management (31)
- Economia e Politica Industriale (25)
- Rivista di Politica Economica (21)
- L'Industria (20).

Finally, in area 14, the smallest in SSH in Italy, we identify 221 journals with 810 articles, approximately 4 articles each. The inter-quartile distribution is 5, 6 and 8, respectively. The top five journals are as follows:

- Il Mulino (26 papers)
- Rassegna Italiana di Sociologia (23)
- La rivista delle politiche sociali (15)
- Salute e società (15)
- Filosofia Politica (14).

The following conclusions can be drawn from these data. First, it appears that the concentration of publication in journals is indeed very low. With an average number of articles equal to 3–4 per journal, and the top journals accounting only for a few decades of articles, it is clear that overall journal production is rather fragmented. This seems to be a structural characteristic of SSH, although with slightly different patterns, in particular in the legal area. Second, all top journals by number of articles are written in Italian language. This is another reminder of the specificity of SSH areas. Third, scholars also include in their self-reported data journals that are non-scientific, or are mixed-type at best. To make a few examples, *Il Mulino* is an established authority in cultural and political debates, but is not a scientific journal; *L'Indice dei libri del mese* is a journal entirely devoted to book reviews. This point calls attention to the need to carefully classify journals in terms of their scientific and academic value, as has been done by Anvur (see the chapter by Bonaccorsi,

Ferrara, Malgarini and Tindaro in this volume). At the same time it is an indication of the way in which scholars in SSH perceive their academic duty, which sometimes seem to include communication to a wider audience.

5 Conclusions

The analysis of VQR publications has been a useful exercise to understand the different strategies adopted by scholars in choosing where to publish their results.

We have discovered a large diversity of practices, which is influenced by the broad disciplinary area (CUN area), but also by the smaller disciplinary field (SSD) and by age. We find confirmation of the crucial role of monographs and book chapters but we also discover that, in area 10, the area of language, literary and art studies, proceedings and other research products play a key role. The Italian language is the most frequent mode of publishing monographs, book chapters and edited books. At the same time many research products, particularly in the Other product category, are often published in a large variety of non-Italian and non-English languages. Area 13, covering economics and business, as well as statistics and applied mathematics, is an outlier under many respects. Journal articles written in English are the key product.

Finally, we find evidence that the publication strategies of scholars do not completely overlap with the administrative definition of disciplines, as proposed by CUN. Perhaps the macro-CUN areas should be redrawn on the basis of the new characteristics the disciplines have assumed, as witnessed at a microlevel by their publication strategies. The scientific societies that guide and offer suggestions to scholars might be a bridge with the national Agency and with CUN in order to reopen the issue of disciplinary boundaries.

References

Abramo, G., & D'Angelo, C. A. (2014). How do you define and measure research productivity? *Scientometrics, 101*(2), 1129–1144.

Alves, A. D., Yanasse, H. H., & Soma, N. Y. (2016). An analysis of bibliometric indicators to JCR according to Benford's law. *Scientometrics, 107*, 1489–1499.

Blasius, J., & Greenacre, M. J. (1994). Computation of correspondence analysis. In *Correspondence analysis in the social sciences* (pp. 53–75). London: Academic Press.

Bonaccorsi, A., Secondi, L., Setteducati, E., & Ancaiani, A. (2014). Participation and commitment in third-party research funding: Evidence from Italian universities. *Journal of Technology Transfer, 39*, 169–198.

Cicero, T., Malgarini, M., Nappi, C. A., & Peracchi, F. (2013). *Bibliometric and peer review methods for research evaluation: A methodological appraisement (in Italian)*. MPRA Paper No. 50470. Available online at https://mpra.ub.uni-muenchen.de/50470/ (Munich Personal REPEC Archive).

Everitt, B. S., Landau, S., & Leese, M. (2011). *Cluster analysis* (5th ed.). Wiley.

Greenacre, M. (2007). *Correspondence analysis in practice* (2nd ed.). London: Chapman & Hall/CRC.
Greenacre, M. J., & Pardo, R. (2006). Subset correspondence analysis: Visualizing relationships among a selected set of response categories from a questionnaire survey. *Sociological Methods and Research, 35*, 193–218.
Hicks, D. (2012). Performance-based university research funding systems. *Research Policy, 41*, 251–261.
Husson, F., Le, S., & Pages, J. (2010). *Exploratory multivariate analysis by example using R*. Chapman and Hall.
Jolliffe, I. T. (1986). *Principal component analysis*. New York: Springer.
Nenadic, O., & Greenacre, M. (2007). Correspondence analysis in R, with two- and three-dimensional graphics: The ca package. *Journal of Statistical Software, 20*(3.) http://www.jstatsoft.org/v20/i03/.

Sitography

ARC. (2015). www.arc.gov.au/era-2015
ASCB. (2012). www.ascb.org/dora/
CNRS. (2016). https://sites.google.com/site/section37cnrs/Home/revues37
NAT. (2015). www.nature.com/news/bibliometrics-the-leiden-manifesto-for-research-metrics-1.17351
NSD. (2016). https://dbh.nsd.uib.no/publiseringskanaler/Forside
SIS. (2014). https://new.sis-statistica.org/wp-content/uploads/2014/12/statsoc-3_2014.pdf
WOS. (2016). https://wokinfo.com/essays/impact-factor/

Journal Ratings as Predictors of Article Quality in Arts, Humanities, and Social Sciences: An Analysis Based on the Italian Research Evaluation Exercise

Andrea Bonaccorsi, Antonio Ferrara, and Marco Malgarini

1 Introduction

There is widespread agreement that research assessment in Social sciences and Humanities (SSH) is complex for a variety of reasons. There are fundamental differences with respect to Science, Technology, Engineering and Mathematics (STEM) fields, especially in the academic publication's structure. In SSH books, book chapters and monographs represent a significant, sometimes dominant, share of scientific production, while journal articles are less central, and national languages are widely used (Finkenstaedt 1990; Hammarfelt 2012). The average number of publications per author, and of authors per publication, is smaller and the time window of citations much longer, so that citations are less immediately useful as indicators of the quality or impact of a publication and the bibliometric approach is thus of limited usefulness (Nederhof et al. 1989). A further complication stems from the number of research quality criteria which is larger in SSH than in other fields. There is less widespread agreement on these criteria (Hemlin 1996; Hemlin and Gustafsson 1996; Ochsner et al. 2012, 2013; Hug et al. 2013, 2014).

This chapter is a revised and expanded version of Bonaccorsi A., Cicero T., Ferrara A. and Malgarini M., *Journal ratings as predictors of articles quality in Arts, Humanities, and Social Sciences: an analysis based on the Italian Research Evaluation Exercise* [version 1; referees: 3 approved]. F1000Research 2015, 4:196 (doi: 10.12688/f1000research.6478.1).

A. Bonaccorsi (✉)
DESTEC, University of Pisa, Pisa, Italy

IRVAPP-FBK, Trento, Italy
e-mail: a.bonaccorsi@gmail.com

A. Ferrara • M. Malgarini
Agenzia Nazionale Valutazione del sistema Universitario e della Ricerca (ANVUR), Rome, Italy
e-mail: antonio.ferrara@anvur.it; marco.malgarini@anvur.it

© Springer International Publishing AG 2018

A. Bonaccorsi (ed.), *The Evaluation of Research in Social Sciences and Humanities*, https://doi.org/10.1007/978-3-319-68554-0_11

The state-of-the-art for SSH research assessment at an international level shows how several roads have been taken to face these challenges. There is widespread agreement that peer review remains the most critical evaluation methodology, and significant efforts are being made to render it more sophisticated, methodologically controlled, based on sound principles of evaluation methodology in social sciences, and free from unwanted biases, distortions, and unexpected side effects. Under this agenda, issues such as the notion of originality, unorthodox science, or interdisciplinarity are under examination (Guetzkow et al. 2004; Hammarfelt 2011).

Much effort is devoted to the classification and evaluation of non-indexed journals (in national languages), which are one of the primary vehicles for academic communication. However, the existing bibliometric databases are limited. The SSH case, which suffers from a scant coverage of relevant publications, partly due to a limited overlap between citing and cited documents (Frost 1979; Hammarfelt 2016), has been carefully discussed in the literature (Nederhof and Zwaan 1991; Nederhof and Noyons 1992; Archambault et al. 2006; Nederhof 2006; Hellqvist 2010). Whatever the specific metrics and the database adopted, the use of citations as the basis for SSH journal ranking has been subject to severe criticism (Christenson and Sigelman 1985; Campbell et al. 2006; Jarwal et al. 2009).

All of this should not lead to the conclusion that scientific production is divided into two irreducibly separated areas, one of which is, by constitution, subject only to qualitative judgement. The issue of the applicability of quantitative methods to the evaluation of research in SSH is open to debate. Classification of journals has been used in several countries for research assessment, sometimes sparking a heated debate.

2 The Debate on the Classification of Academic Journals

As of now, a relatively large body of literature exists on the issue of journal classification. Elsewhere (Ferrara and Bonaccorsi 2016) we discussed its appropriateness and the limits of citation-based and expert opinion-based journal rankings. Hereafter we will focus on criticisms concerning the impact of journal rankings on national languages, academic publication patterns, interdisciplinarity, paradigmatic pluralism and academic freedom. According to such critiques, the production of indicators might orient researchers towards opportunistic publication behaviour (Butler 2003a, b). The entry of new journals might be made too difficult and expensive (Lamp 2009); the rating of journals might lead to an under-investment in interdisciplinary fields (Rafols et al. 2012) and discourage from the undertaking of risky or unorthodox research. Journal classification may be considered the privilege of mainstream science (Rodriguez-Navarro 2009), so that some of the most interesting articles may not be published in top-rated journals (Starbuck 2005).

While these arguments have elements of truth, they do not intrinsically depend on the construction of indicators, as the latter only make the underlying dynamics of recognition more visible. The tension between normal and revolutionary science does not depend in itself on the academic journal system. There is always a trade-off

between pursuing mainstream science and taking a risk by looking for radically new discoveries (Carayol and Dalle 2007). While for individual authors and in the short term the argument might be valid, in the long run, and for large aggregates, the argument is untenable.

The concern that journal rankings might push researchers to switch to the English language, reducing the expressive role of national languages, which are important to all Humanities, is not supported by the literature. Bolton and Kuteeva (2012) carried out an extensive survey at Stockholm University about the use of English in research and teaching and concluded that while English has become standard in STEM fields, in SSH it was auxiliary to the national language.

Another concern is that the classification of journals might induce scholars to switch their publication habits from monographs to journal articles, often opportunistically, and contribute to the so-called "death of the monograph." This is an urgent and worrisome issue in fields like geography (Ward 2009) and literature (Thompson 2002) as monographs are at the core of research (Glanzel and Schoepflin 1999; Thompson 2002; Williams et al. 2009; Gimenez-Toledo and Roman-Roman 2009). Hammarfelt and De Rijcke (2015) have examined the impact of the evaluation system adopted by the Faculty of Arts at Uppsala University in Sweden. They found a significant increase in the share of peer-reviewed publications, without any distortion in the pattern of humanities publications or any "death of monographs." In their view, reaction to journal rating is local, context-dependent and mediated by disciplinary practices.

The interdisciplinarity impact is difficult to evaluate. Rafols et al. (2012) found that journal ranking in social sciences was detrimental to interdisciplinary research in economic areas. This is indeed a serious problem, worthy of consideration from any agency or body in charge of research assessment.

As for conformism and academic freedom, there is literature arguing that rankings service orthodoxy and the mainstream. This argument is mostly based on anecdotal evidence and disciplinary case studies (e.g. in pedagogical research, see the special issue of *Power and Education* on journal rankings). This observation has magnified the problem that in many cases departments and universities base some of their decisions on indicators of journal rankings, instead of assessments of individual articles (e.g. Hasselback et al. 2000; Lee 2006). For example, Gomez-Mejia and Balkin (1992) found that the number of a faculty member's publications, especially those in top-tier management journals, was a primary determinant of faculty pay (see also Park and Gordon 1996). Van Fleet et al. (2000) examined the use of journal lists by departments and warned against the risk that narrow and idiosyncratic lists might distort faculty research attitudes, particularly those of junior members, leading to detrimental conformism. Macdonald and Kam (2008) developed a radical critique of journal rating, pointing to their circularity and arbitrariness, and arguing that its introduction is a threat to non-orthodox and critical research traditions. Hogler and Gross (2009) advanced this direction and argued that journal rating, and indirectly the rating of business schools, is a device for the ideological manipulation of management education, a process that might be examined in the light of the Marxian theory of commodification.

However, some studies have assessed the overall impact of journal rating on academic communities at a national level. Most of these assessments come from Scandinavian countries where journal rating has been extensively introduced, with an impact on university funding. The so-called "Norwegian model," described by Schneider (2009) and Sivertsen (2016), has been especially influential. Developed between 2003 and 2004, it combined a point-based assessment with a performance-based funding system. It was used for funding allocation for the first time in 2006. The system is based on a two-level classification of all publications ("channels"), the collection of data on the whole scientific production, and the assignment of scores to different classes of products. It was also adopted in Denmark, Finland, and some Swedish universities and it triggered an increase in research both in output and the share of output channelled through level 2, or more prestigious, publications (Aagaard et al. 2015). The system's legitimacy is based on the involvement of scholars in the classification process. Challenges are identified in the areas of transparency, representation in individual committees, and the placement of channels among committees (Ahlgren et al. 2012). Ingwersen and Larsen (2014) carried out a detailed examination of the point-based system's impact in Denmark. This system was introduced after Norway's initial experience finding that research article publication was positively affected by the introduction of performance indicators. The latter has, so far, not resulted in an increase in duplicate or redundant publications or a decline in citation impact. Overall, these studies do not lend support to the pessimistic view that journal rating would damage the independence, freedom and established traditions in the practice of research and style of SSH publishing.

3 The Italian Experience

This paper reports on a large experiment in the classification of journals in SSH carried out in Italy between 2012 and 2014 for the National Scientific Habilitation (*Abilitazione Scientifica Nazionale*: ASN). The exercise was based upon a mandatory legal provision to rate *all* journals, to calculate the overall candidates' academic production as part of the national procedure to become an associate professor or full professor. This exercise asked the National Agency for the Evaluation of Universities and Research Institutes (ANVUR) to evaluate all journals in which *at least one* Italian scholar published *at least one* paper in 2002–2012. This figure was more than 60,000 titles.

While the rating of journals was followed in several national contexts, it is only in the Italian exercise that there is the opportunity to carry out a controlled experiment to test the robustness of journal classification. In fact, two independent evaluations were carried out on the same set of journals. Firstly, a panel of experts classified all journals as academic and non-academic (i.e. popular, professional, technical, cultural and political). Then a subset of academic journals was rated as 'A-class.' This exercise was based on the reputation, esteem, diffusion and impact

of journals – a qualitative, expert-based, reputational basis. Elsewhere we have detailed the entire process (Ferrara and Bonaccorsi 2016).

Individual articles published in those journals were rated by a large number of individual referees as part of a nation-wide research assessment exercise (*Valutazione della Qualità della Ricerca*, VQR 2004–10). At least for the considered disciplines, this exercise relied entirely on peer review, carried out by a large number of referees *external* to the panels in charge of the assessment process (Ancaiani et al. 2015). Experts were expected to grade each article on its own merits – which they did, sometimes assigning low scores to articles published in prestigious journals. While referees could have known about the ratings mentioned above (published while the VQR was still ongoing), there is no evidence that any article received higher scores only because it appeared in an outstanding journal. In those fields where journals were indexed and ranked (not simply rated), expert judgements were stricter than expected and some top class papers were assigned a lower score than that allocated to the same class by the journal ranking (Bertocchi et al. 2015). In addition, the Report from the expert panel in Law explicitly clarified that the indication of rating of journals had no systematic influence on the assessment of individual articles, as discussed by Peruginelli and Faro in this volume. This was further evidence that referees did not succumb to any 'halo effect' emanating from the standing of the journal in which the articles were published.

This does not negate that the quality of the peer review process can sometimes be an issue. This problem is not entirely new but has taken a different order of magnitude recently, due to the acceleration of scientific production and the peer-review burden placed upon researchers. Several studies on the peer review process have drawn attention to a variety of biases (Chubin and Hackett 1990; Cicchetti 1991; Daniel 1993; Campanario 1998a, b). Studies by Mahoney (1977) and Travis and Collins (1991) suggested that the familiarity of referees with a subject domain and the affiliation to schools of thought might affect their decisions due to cognitive and institutional particularism. The latter investigated the extent to which experts were biased by their cognitive cronyism, or allegiance to a specific view of scientific practice, irrespective of how professionally they carried out the peer review. This issue has been extensively discussed, but the empirically supported conclusion is that particularism does not dominate in scientific judgement (Cole et al. 1981; Cole 1992). More recently there is the issue of distortions in the peer-review process induced by the need to control costs. Gläser and Laudel (2005) identified several problems in the peer review process carried out at a distance ("remote evaluation" or Taylorised assessment), often used to minimise costly interaction among experts.

However, there is no reason to think that the VQR peer review process has been vulnerable to such problems. The involvement of many external experts, each entrusted with a reasonable number of 'research outcomes' submitted for evaluation, minimised any concerns that an excessive workload would impinge on the ability to read the publications carefully and assess them in a balanced way.

One may argue that the two evaluations were not independent since a similar exercise had preceded the rating of journals for ASN purposes. This had been implemented by the expert panels that were supervising the VQR process (see Baccini

2016). However, this last exercise had a much more limited scope and identified only a relatively small number of top journals, starting from pre-determined lists which were provided by scientific societies. The primary goal of journal rating within the VQR was to identify those journals from which it was expected that Italian researchers would submit their *best* articles. It was a top-down exercise. On the contrary, the ASN exercise started from the total number of entries in the self-administered website of scholars, which included more than 60,000 entries. It was a bottom-up exercise and led to a much larger number of classifications. The smaller number of journals in the VQR exercise was arranged into up to three tiers, while the ASN journal ratings (devised by a smaller number of experts, who were *not* members of the VQR panels) were derived from parsing a much larger set of publication venues to exclude those which were non-academic. Academic journals were arranged in only two tiers – A-class versus the rest.

The top tiers of the two lists overlap to some extent. This mostly results from reputational factors – at least a large part of the affected academic communities share the opinion that some academic journals (because of their history, intellectual origin, editorial policy, editorial board reputation and scientific committee) possess higher qualities than others. Once we account for the influence of these reputational aspects, it is fair to conclude that the two exercises of journal ratings discussed here were different enough in their purposes and methods as to be considered as substantially, if not entirely, independent. As noted above, the paper by Bertocchi et al. (2015) shows that, in economic disciplines, scores given by referees in VQR to individual articles (our dependent measure in the model below) did not automatically follow the rating of the journals in which these articles were published. A similar remark can be found in the case of legal studies. An essential element to exclude that referees followed the VQR journal rating in evaluating articles automatically so that our model's variables were not independent.

It is possible to carry out a controlled experiment, extending to all SSH, except economics and business, using the analysis initiated by Ferrara and Bonaccorsi (2016) for journals of anthropology, education sciences, geography, history, library science, palaeography, and philosophy. In the following, we first introduce the database used for the analysis and test the influence of the journal classification on the article score. The paper will be concluded with a discussion of the results obtained.

4 Methodology

The paper is based on a dataset including data on all the journal articles submitted for evaluation by Italian scholars in the disciplinary areas of architecture, arts and humanities, history and philosophy, law and sociology and political science. Submissions for evaluation took place within the framework of VQR 2004–10. Italy's national research assessment exercise involved all professors and researchers affiliated to the Italian Universities and Public Research Organisations (PROs) as of November 2011. According to adopted rules, SSH research evaluation was entirely

Table 1 Description of dataset

Area of assessment	Acronym	Full professor	Associate professor	Researcher	Other	N° of articles	N° of articles in Class A journals
Architecture	Area08	280	278	353	7	918	360
Antiquities, philology, literary studies, art history	Area10	1040	1191	1322	26	3579	1954
History, philosophy, pedagogy, and psychology	Area11	713	680	726	8	2127	1086
Law	Area12	1488	983	1337	30	3838	2637
Political and social sciences	Area14	338	409	442	9	1198	664
Total		3859	3541	4180	80	11,660	6701

based on peer review, with the exception of economics and psychology (see the chapters by Bonaccorsi in this volume). Research quality was assessed against *relevance* criteria, intended as contribution to the advancement of the state of art in the field, adequacy, efficacy, timeliness and duration of impacts; *originality and innovation*, intended as a contribution to the creation of new knowledge in the field; *internationalisation*, intended as its position in the international research. Five Groups of Evaluation Experts (GEV in the Italian acronym) carried out the evaluation; one for each SSH area (Architecture; Arts and Humanities; History and Philosophy; Law; Sociology and political sciences). Reviewers were instructed by the GEV to evaluate articles only on the basis or their merit, regardless of the journal in which they were published and of the publication language. Each article had a possible rating of Excellent (A), Good (B), Fair (C) or Limited (D); to each class corresponded to a score ranging from 1 (for articles A-rated) to zero (for articles deemed as limited). Negative scores were assigned if the article was deemed as non-academic (−1) or for plagiarism or fraud (−2, see Ancaiani et al. 2015 for details). In the human and social sciences fields, a substantial fraction of articles – namely, 6701 out of 11,660 (Table 1) – appeared in journals deemed as 'A-class' according to the ASN procedure which was intended to select the best researchers for the ranks of associate and full professors.

According to the relevant Ministerial Decree (No. 76/2012), those journals, were 'internationally recognised as excellent because of the rigor of their peer review procedures and because of their diffusion among, esteem by, and impact on, the scholarly community of a field, as indicated by their presence in the major national and international databases' (our translation). Most of the remaining articles appeared in journals considered as 'academic' for ASN purposes, while a minority were published in journals that remained 'uncategorised'. The dataset's main feature

Table 2 Preliminary analysis of the association between the evaluation of research product and the journal evaluation

		Evaluation of Journal			
		A	Not A	Not academic	Total
Evaluation of research product	A	1344	573	20	1937
	B	3184	1743	92	5019
	C	1322	1096	80	2498
	D	837	1176	150	2163
	Non-academic and other	14	21	8	43
	Total	6701	4609	350	11,660

Pearson $\chi^2 = 630.9$*; p-value = 0.000*

is that it allows a comparison between the evaluations of journals and individual articles.

A preliminary analysis shows that there is a relationship between the evaluation of individual articles and journals where the article is published (Table 2). The non-parametric statistic for categorical data (Pearson χ^2) is statistically significant at 1%,[1] showing that the two distributions are dependent and the two ratings are mutually related. In the following, we will examine this relationship and the controls of several author-level and article-level variables, more thoroughly.

5 The Influence of Journal Classification on the Article Score

We assume that the probability of an article i, published in the journal j, receiving a score equal to $x \in \{-2; 1\}$ is influenced by the class assigned to the journal and the controls for several article characteristics:

$$P\left(Score_{i,j} = x\right) = F\left(Journal\,class_{i,j}, j, Paper\,characteristics_{i,j}\right) \quad (1)$$

Among the controls, we considered the publication language (Italian or not) and the age (distinguishing among three age classes; less than 40 years, between 41 and 55 years and more than 55 years); scientific activity sector (Scientific Areas 8, 10, 11, 12, 14); academic status (full professor; associate professor; researcher; other); and the researcher's gender. We add two binary variables controlling the existence of international co-author(s) and for the referees' nationality (allowing for the possibility of international referees). Finally, we added a variable for the size of the author's scientific area. It uses an ordered probit model, which is an extension of the standard binary probit model, used when the dependent variable has ranked and

[1] All the statistical analyses have been performed using the software STATA ver. 13 (http://www.stata.com/stata13/)

multiple discrete variables, alternatively considering the whole sample or each scientific area. In the first case, we also control the possible area-specific effects.

To avoid the "dummy trap", we normalised those articles written in Italian with no international co-author, evaluated by an Italian reviewer, presented by a female researcher in sociology and political science, aged less than 40. This means that the statistical significance, sign and magnitude of estimated parameters are interpreted as control group differentials. The total available observations amount to 11,660 varying from a minimum of 918 in Architecture to a maximum of 3838 in Law (Table 3).

The main result was that at the aggregate level and in each scientific area the article score was higher as the journal ranking improved. The probability of receiving a high score grew if the article was published in a high-ranking journal according to ASN's experts' evaluation. When assessing the control variables, we confirmed most of the results which already emerged in a previous paper using the same data (Cicero et al. 2016): article scores are higher for papers not written in Italian, with international co-authors, published by an under-40, male, full or associate professor. We found that at an aggregate level, and in most areas, an international reviewer and a lower number of professors in the specific scientific sector (SSD, *Settore Scientifico Disciplinare*) results in a higher article score. A possible interpretation of the first result is that the expert groups responsible for the evaluation assign international reviewers to more international papers which have a greater probability of receiving a high score – given that the level of internationalisation was one of the VQR evaluation criteria (see Ancaiani et al. 2015). The negative relationship between area size and article score result already emerged in Ferrara and Bonaccorsi (2016) and is now extended to all SSH. A possible interpretation is that small fields may be favoured by a "proximity bias" among authors and reviewers, thus resulting, *ceteris paribus*, in higher article scores.

As a final check, once controlling for the same variables considered in model 1, we concentrated on the probability of receiving an excellent score and related it to the fact that the article is published in a top, A-Class journal:

$$P\left(Score_{i,j} = "E"\right) = F\left(Journal\,class_{i,j} = "\mathrm{A}", Paper\,characteristics_{i,j}\right) \qquad (2)$$

In (2), F is the logistic function, and the model is estimated as a logit, a class of models allowing the prediction of a binary response based on the specified predictors. A desirable feature of the logit model is that the regression coefficients may easily be transformed into an odds ratio, expressing the change in the odds of the occurrence under scrutiny (in our case, the odds for a paper of receiving an 'Excellent' evaluation) due to a small change of a given predictor. In our case, we were particularly interested in the odds associated with the classification of a journal as a top, Class A journal. Estimation results for both the aggregate sample and each scientific area are presented in Table 4.

According to logit estimations, the probability of receiving an excellent evaluation is positively affected by the journal in which the paper is published. Publishing

Table 3 Ordered probit model (Dependent variable: article score)

Variables	Total	Architecture	Arts&Hum.	Hist.& Phil.	Law	Sociology & Pol. Sci.
Journal rating	0.417***	0.542***	0.400***	0.379***	0.503***	0.323***
Architecture	0.134***					
Arts and Humanities	0.720***					
Hist. & Philosophy	0.471***					
Law	0.259***					
Italian language	−0.372***	−0.518***	−0.148***	−0.623***	−0.281***	−0.704***
41–55 years	−0.151***	−0.256	−0.208***	0.0492	−0.265***	−0.265**
More than 55 years	−0.582***	−0.662***	−0.726***	−0.394***	−0.563***	−0.572***
Associate professor	0.318***	0.206**	0.308***	0.265***	0.439***	0.234***
Full professor	0.818***	0.788***	0.690***	0.660***	1.096***	0.679***
Other personnel	−0.277**	−0.157	−0.329	0.359	−0.415**	−0.742*
Male	0.0777***	0.184**	0.0882**	0.0257	0.0542	−0.00491
International coauthors	0.301***	0.505***	0.122	0.237	0.902***	0.205
International reviewer	0.153***	0.363***	0.185***	0.0243	0.225***	0.0722
Number of Professors in the scientific sector (SSD)	−0.00131***	−0.00297***	−0.00248***	−0.00146**	−0.000921***	−0.00390***
Constant cut1	−2.854***	−2.514***	−2.081***	−2.990***	−1.867***	−2.953***
Constant cut2	−1.718***	0.363	−0.454***	−2.106***	−1.777***	−0.490***
Constant cut3	−1.695***	1.076***	0.200*	−0.357**	0.378***	0.402**
Constant cut4	0.302***	2.387***	1.574***	0.318**	1.162***	1.539***
Constant cut5	1.026***			1.651***	2.710***	
Constant cut6	2.400***					
Observations	11,660	918	3579	2127	3838	1198
Pseudo R-squared	0.0814	0.0919	0.0543	0.0720	0.0926	0.0832

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 4 Logit model (Odds ratio)

Variables	Total	Architecture	Arts, & Hum.	Hist. & Phil.	Law	Sociology & Pol. Sci.
Top journal classification	1.952***	2.513***	1.834***	2.424***	1.990***	1.311
Architecture	1.210					
Arts and Humanities	3.042***					
History and Philosophy	2.031***					
Law	1.084					
Italian language	0.488***	0.311***	0.681***	0.333***	0.529***	0.243***
41–55 years	0.878	0.408**	0.697**	1.144	0.807	0.671
More than 55 years	0.411***	0.248***	0.303***	0.506**	0.572***	0.252***
Associate professor	1.793***	1.283	1.815***	1.825***	2.629***	1.620*
Full professor	4.650***	3.263***	3.831***	4.023***	9.909***	4.877***
Other personnel	1.660	–	1.360	3.057	2.470	1.293
Male	1.247***	1.664**	1.263***	1.155	1.077	1.028
International coauthors	1.611***	2.357**	1.118	1.558	5.149***	1.511
International reviewer	1.352***	1.566**	1.393***	1.178	1.560***	1.490**
Full prof. in the SSD	0.998**	0.992**	0.996**	0.998	1.000	0.990***
Constant	0.065***	0.201***	0.258***	0.137***	0.0332***	0.236***
Observations	11,660	911	3579	2127	3838	1198
Pseudo R-squared	0.116	0.140	0.0738	0.122	0.129	0.143

*** $p < 0.01$, ** $p < 0.05$, *$p < 0.1$

in a Class A journal almost doubled the probability of receiving an excellent evaluation. In each scientific area, the odds of receiving an excellent evaluation were more than doubled by the publication in a Class A journal in architecture and history and philosophy. The effect is somewhat lower, but still highly significant, in law, and arts and humanities, while disappearing in sociology and political sciences. Logit estimation broadly confirmed the results already emerging from the ordered probit model. The odds of receiving an excellent evaluation were increased by publishing in a foreign language, with an international co-author (albeit only in law and architecture) and when the submitting author was 40 years old or younger, an associate or full professor and a male. Gender is significant at the aggregate level and in architecture and humanities, but not in the remaining areas. Having an international

reviewer and publishing in an SSD with a lower number of full professors helped in obtaining an excellent evaluation.

6 Conclusions

Using a large dataset of journal articles published in SSH, the paper proves that independent classifications of journals may be considered as good predictors of the score assigned to individual articles. More specifically, we found that, after controlling some articles' characteristics, the probability of receiving a better score grew with the quality profile of the journal in which the article was published. The probability of receiving an excellent score almost doubles when the paper was published in a top, A-Class journal. The findings held both at the aggregate level and for each specific sub-areas that were considered in the analysis. While peer review must remain the main evaluation methodology, our results indicate that expert-based journal classification may be considered a useful supporting tool in a large evaluation exercise, since it may provide reviewers with valuable information which is apt to support expert evaluation.

References

Aagaard, K., Bloch, C., & Schneider, J. W. (2015). Impacts of performance-based research funding systems: The case of the Norwegian publication indicator. *Research Evaluation, 24*(2), 106–117.

Ahlgren, P., Colliander, C., & Persson, O. (2012). Field normalized citation rates, field normalized journal impact and Norwegian weights for allocation of university research funds. *Scientometrics, 92*(3), 767–780.

Ancaiani, A., Anfossi, A., Barbara, A., Benedetto, S., Blasi, B., Carletti, V., Cicero, T., Ciolfi, A., Costa, F., Colizza, G., Costantini, M., di Cristina, F., Ferrara, A., Lacatena, R. M., Malgarini, M., Mazzotta, I., Nappi, C. A., Romagnosi, S., & Sileoni, S. (2015). Evaluating scientific research in Italy: The 2004-2010 research evaluation exercise. *Research Evaluation, 24*(3), 242–255.

Archambault, E., Vignola-Gagnè, E., Cotè, G., Larivière, V., & Gingras, Y. (2006). Benchmarking scientific output in the social sciences and humanities: The limits of existing databases. *Scientometrics, 68*(3), 329–342.

Baccini A. (2016). Comments on Bonaccorsi A, Cicero T, Ferrara A. and Malgarini M., *Journal ratings as predictors of articles quality in Arts, Humanities and Social Sciences: an analysis based on the Italian Research Evaluation Exercise*. https://f1000research.com/articles/4-196/v1.

Bertocchi, G., Gambardella, A., Jappelli, T., Nappi, C. A., & Peracchi, F. (2015). Bibliometric evaluation vs. informed peer review: Evidence from Italy. *Research Policy, 44*, 451–466.

Bolton, K., & Kuteeva, M. (2012). English as an academic language at a Swedish university: Parallel language use and the 'threat of English'. *Journal of Multilingual and Multicultural Development, 33*(5), 429–447.

Butler, L. (2003a). Explaining Australia's increased share of ISI publications – The effects of a funding formula based on publication counts. *Research Policy, 32*, 143–155.

Butler, L. (2003b). Modifying publication practices in response to funding formulas. *Research Evaluation, 12*, 39–46.

Campanario, J.-M. (1998a). Peer review for journals as it stands today – Part 1. *Science Communication, 19*(3), 181–211.

Campanario, J.-M. (1998b). Peer review for journals as it stands today – Part 2. *Science Communication, 19*(4), 277–306.

Campbell, K., Goodacre, A., & Little, G. (2006). Ranking of United Kingdom law journals: An analysis of the research assessment exercise 2001 submissions and results. *Journal of Law and Society, 33*, 335–363.

Carayol, N., & Dalle, J. M. (2007). Sequential problem choice and the reward system in Open Science. *Structural Change and Economic Dynamics, 17*, 167–191.

Christenson, J. A., & Sigelman, L. (1985). Accrediting knowledge: Journal stature and citation impact in social science. *Social Science Quarterly, 66*(4), 964–975.

Chubin, D. E., & Hackett, E. J. (1990). *Peerless science: Peer review and U.S. science policy*. Albany: State University of New York Press.

Cicchetti, D. V. (1991). The reliability of peer review for manuscript and grant submissions: A cross-disciplinary investigation. *Behavioural and Brain Sciences, 14*, 119–186.

Cicero, T., Malgarini, M., & Benedetto, S. (2016). Determinants of research quality in Italian universities: Evidence from the 2004 to 2010 evaluation exercise. *Research Evaluation, 25*(3), 257–263.

Cole, S. (1992). *Making science. Between nature and society*. Cambridge, MA: Harvard University Press.

Cole, S., Cole, J. R., & Simon, G. (1981). Chance and consensus in peer review. *Science, 214*, 881–886.

Daniel, H.D. (1993). *Guardians of science. Fairness and reliability of peer review*. Weinheim: VCH.

Ferrara, A., & Bonaccorsi, A. (2016). How robust is journal rating in humanities and social sciences? Evidence from a large-scale multi-method exercise. *Research Evaluation, 25*(3), 279–291.

Finkenstaedt, T. (1990). Measuring research performance in the humanities. *Scientometrics, 19*(5/6), 409–417.

Frost, C. O. (1979). The use of citations in literary research: A preliminary classification of citation functions. *Library Quarterly, 49*(4), 399–414.

Gimenez-Toledo, E., & Roman-Roman, A. (2009). Assessment of humanities and social sciences monographs through their publishers: A review and a study towards a model of evaluation. *Research Evaluation, 18*, 201–213.

Glanzel, W., & Schoepflin, U. (1999). A bibliometric study of reference literature in the sciences and the social sciences. *Information Processing and Management, 35*, 31–44.

Gläser, J., & Laudel, G. (2005). Advantages and dangers of 'remote' peer evaluation. *Research Evaluation, 11*(2), 141–154.

Gomez-Mejia, L., & Balkin, D. (1992). Determinants of faculty pay: An agency theory perspective. *Academy of Management Journal, 35*(5), 921–955.

Guetzkow, J., Lamont, M., & Mallard, G. (2004). What is originality in the humanities and the social sciences? *American Sociological Review, 69*(2), 190–212.

Hammarfelt, B. (2011). Interdisciplinarity and the intellectual base of literature studies: A citation analysis of highly cited monographs. *Scientometrics, 86*(3), 705–725.

Hammarfelt, B. (2012). Harvesting footnotes in a rural field: Citation patterns in Swedish literary studies. *Journal of Documentation, 68*(4), 536–558.

Hammarfelt, B. (2016). Beyond coverage: Toward a bibliometrics for the humanities. In M. Ochsner et al. (Eds.), *Research assessment in the humanities* (pp. 115–132). London: Springer Open 2016.

Hammarfelt, B., & De Rijcke, S. (2015). Accountability in context: Effects of research evaluation systems on publication practices, disciplinary norms, and individual working routines in the Faculty of Arts at Uppsala University. *Research Evaluation, 24*(1), 63–77.

Hasselback, J. R., Reinstein, A., & Schwan, E. S. (2000). Benchmarks for evaluating the performance of accounting faculty. *Journal of Accounting Education, 18*, 79–97.

Hellqvist, B. (2010). Referencing in the humanities and its implications for citation analysis. *Journal of the American Society for Information Science and Technology, 61*(2), 310–318.

Hemlin, S. (1996). Social studies of the humanities: A case study of research conditions and performance in ancient history and classical archaeology and English. *Research Evaluation, 6*(1), 53–61.

Hemlin, S., & Gustafsson, M. (1996). Research production in the arts and humanities. A questionnaire study of factors influencing research performance. *Scientometrics, 37*(3), 417–432.

Hogler, R., & Gross, M. A. (2009). Journal rankings and academic research: Two discourses about the quality of faculty work. *Management Communication Quarterly, 23*(1), 107–126.

Hug, S. E., Ochsner, M., & Daniel, H.-D. (2013). Criteria for assessing research quality in the humanities: A Delphi study among scholars of English literature, German literature and art history. *Research Evaluation, 22*(5), 369–383.

Hug, S. E., Ochsner, M., & Daniel, H.-D. (2014). A framework to explore and develop criteria for assessing research quality in the humanities. *International Journal of Education Law and Policy, 10*(1), 55–64.

Ingwersen, P., & Larsen, B. (2014). Influence of a performance indicator on Danish research production and citation impact 2000–12. *Scientometrics, 101*(2), 1325–1344.

Jarwal, S. D., Brion, A. M., & King, M. L. (2009). Measuring research quality using the journal impact factor, citations and 'ranked journals': Blunt instruments or inspired metrics? *Journal of Higher Education Policy and Management, 31*(4), 289–300.

Lamp, J. W. (2009). At the sharp end: Journal ranking and the dreams of academics. *Online Information Review, 33*(4), 827–830.

Lee, F. S. (2006). The ranking game, class, and scholarship in American mainstream economics. *Journal of Economics, 3*, 1–41.

Macdonald, S., & Kam, J. (2008). Quality journals and gamesmanship in management studies. *Management Research News 31(8),*595-606.

Mahoney, M. J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research, 1*(1977), 161–175.

Nederhof, A. J. (2006). Bibliometric monitoring of research performance in the social sciences and the humanities: A review. *Scientometrics, 66*(1), 81–100.

Nederhof, A. J., & Noyons, E. C. M. (1992). International comparison of departments' research performance in the humanities. *Journal of the American Society for Information Science, 43*(3), 249–256.

Nederhof, A. J., & Zwaan, R. A. (1991). Quality judgements of journals as indicators of research performance in the humanities and the social and behavioural sciences. *Journal of the American Society for Information Science, 42*(5), 332–340.

Nederhof, A. J., Zwaan, R. A., De Bruin, R. E., & Dekker, P. (1989). Assessing the usefulness of bibliometric indicators for the humanities and the social sciences – A comparative study. *Scientometrics, 15*(5/6), 423–435.

Ochsner, M., Hug, S. E., & Daniel, H.-D. (2012). Indicators for research quality in the humanities: Opportunities and limitations. *Bibliometrie – Praxis und Forschung, 1*, 4.

Ochsner, M., Hug, S. E., & Daniel, H.-D. (2013). Four types of research in the humanities: Setting the stage for research quality criteria in the humanities. *Research Evaluation, 22*(2), 79–92.

Park, S., & Gordon, M. (1996). Publication records and tenure decisions in the field of strategic management. *Strategic Management Journal, 17*(2), 109–128.

Rafols, I., Leydesdorff, L., O'Hare, A., Nightingale, P., & Stirling, A. (2012). How journal rankings can suppress interdisciplinarity. The case of innovation studies in business and management. *Research Policy, 41*(7), 1262–1282.

Rodriguez-Navarro, A. (2009). Sound research, unimportant discoveries: Research, universities, and formal evaluation of research in Spain. *Journal of the American Society for Information Science, 60*(9), 1845–1858.

Schneider, J. W. (2009). An outline of the bibliometric indicator used for performance-based funding of research institutions in Norway. *European Political Science, 8*(3), 364–378.

Sivertsen, G. (2016). Publication-based funding: The Norwegian model. In M. Ochsner et al. (eds.), *Research Assessment in the Humanities*, Springer Open 2016, London, 79–90.

Starbuck, W. H. (2005). How much better are the most prestigious journals? The statistics of academic publication. *Organization Science, 16*, 180–200.

Thompson, J. W. (2002). The death of the scholarly monograph in the humanities? Citation patterns in literary scholarship. *Libri, 52*, 121–136.

Travis, G. D. L., & Collins, H. M. (1991). New light on old boys: Cognitive and institutional particularism in the peer review system. *Science, Technology & Human Values., 16*(3), 322–341.

Van Fleet, D., McWilliams, A., & Siegel, D. (2000). A theoretical and empirical analysis of journal rankings: The case of formal lists. *Journal of Management, 26*(5), 839–861.

Ward, K. (2009). The future of research monographs: An international set of perspectives. *Progress in Human Geography., 33*(1), 101–126.

Williams, P., et al. (2009). The role and future of the monograph in arts and humanities research. *ASLIB Proceedings, 61*, 67–82.

Part IV
Exploring New Indicators for the Multilevel Assessment of Research

Google Scholar as a Citation Database for Non-bibliometric Areas: The EVA Project Results

Alfio Ferrara, Stefano Montanelli, and Stefano Verzillo

1 Introduction

The use of bibliometric approaches for scholar and research assessment is widely enforced in STEM disciplines and it has been recently investigated also for social sciences and humanities where the use of quantitative methods is not a consolidated practice (Garfield 1980; Linmans 2010; Nederhof 2006). A number of commercial databases exists, like for example NCBI-PubMed, ISI-WoS, and Elsevier-Scopus, and they are recognized as reliable data sources for supporting calculation of citation indexes (Ferrara and Salini 2012). These databases provide "trustworthy" citation calculation, in that authors, titles, and venues of indexed publications are checked and verified. On the other side, issues about limitation and partial coverage of these databases with respect to the overall scientist production are also recognized (Archambault et al. 2006). For social sciences and humanities, open issues are even more challenging due to the fact that reference citation databases are still missing for both reliability and coverage.

In this chapter, we present the EVA (Extraction, Validation, and Analysis) project and related results about the use of Google Scholar as web database for calculation of citation indexes in non-bibliometric scientific areas, such as social sciences and humanities. The core research issue of EVA is to investigate the quality of publication records that can be retrieved from Google Scholar. In particular, the focus of this chapter is on the problem of properly disambiguating author names in retrieved records, with the aim at assigning scholars to the set of publications they actually

A. Ferrara (✉) • S. Montanelli
Department of Computer Science, Università degli Studi di Milano, Milan, Italy
e-mail: alfio.ferrara@unimi.it; stefano.montanelli@unimi.it

S. Verzillo
Department of Economics, Management and Quantitative Methods,
Università degli Studi di Milano, Milan, Italy
e-mail: stefano.verzillo@unimi.it

© Springer International Publishing AG 2018

A. Bonaccorsi (ed.), *The Evaluation of Research in Social Sciences and Humanities*, https://doi.org/10.1007/978-3-319-68554-0_12

authored. To this end, we present the *author disambiguation techniques* developed in the EVA project based on two different strategies called *similarity-oriented* and *specificity-oriented* characterized by the use of latent semantic indexing and text normalization techniques, respectively. In EVA, similarity and specificity strategies can be exploited as alternative options as well as in a combined way for enabling a flexible configuration of the author disambiguation process. The results of the EVA project are presented on a case-study about the publication records retrieved from Google Scholar for a dataset of Italian academic researchers belonging to non-bibliometric scientific areas. The goal of the EVA case-study is twice. First, we provide evaluation results about the effectiveness of the EVA techniques for author disambiguation. Second, we provide a descriptive analysis of the obtained results. As a further contribution of EVA about the coverage of Google Scholar, a comparative evaluation of the case-study results against the Elsevier-Scopus database is also provided.

The chapter is organized as follows. In Sect. 2, we provide the literature review. In Sect. 3, we present the EVA approach and related phase-organization to author disambiguation and bibliometric data analysis. The author-based disambiguation techniques developed in EVA are illustrated in Sect. 4. In Sects. 5 and 6, we discuss the results obtained by applying the EVA approach to an Italian case-study and the results obtained by evaluating EVA against the Elsevier-Scopus commercial database, respectively. Concluding remarks are finally provided in Sect. 7.

2 Literature Review

In the recent years, the idea to exploit Google Scholar as a citation database has been proposed as a possible alternative to commercial solutions like for example ISI-WoS and Elsevier-Scopus (Aguillo 2012; Archambault et al. 2006; Falagas et al. 2008; Kousha and Thelwall 2007). In particular, in the related work, the focus is on coverage aspects. On the one side, it is recognized that Google Scholar outperforms commercial databases on the number of indexed publications (also in non-English language), as well as on the number of considered scientific areas, especially in the field of humanities and social sciences. On the other side, drawbacks about accuracy and quality of the retrieved results are widely recognized as well. In most of the existing approaches where Google Scholar is considered for citation extraction, the discussion is on bibliometrics aspects with focus on index selection for addressing a given analysis/evaluation problem (Ferrara and Salini 2012). As a matter of fact, issues about retrieval techniques and quality of extracted metadata are only marginally discussed. This is the case of systems and tools for large-scale bibliographic-data analysis, such as for example Academic analytics, Global Research Benchmarking (GRB), InCites – Thomson Reuters, Scival – Elsevier, Evaluation Support System – Research Value2, in which the supported procedure for data acquisition and validation are only marginally described (Biolcati-Rinaldi et al. 2012).

Other related works are about author disambiguation over a set of publication records, with the aim at evaluating the authorship relations between scholars and publications. The goal is to distinguish those relations that are correct (and need to be confirmed) from those relations that are incorrect mainly due to homonyms (and need to be discarded). Some interesting work in this field are (D'Angelo et al. 2011; Ferreira et al. 2010; Han et al. 2004, 2005a, b; On and Lee 2007; Smalheiser and Torvik 2009; Tang et al. 2012; Torvik et al. 2005; Yang et al. 2008). In the literature, a popular disambiguation method exploits co-authorship relations and it is based on the idea that a co-author of a scholar *sn* usually appears in multiple publications of *sn*. Thus, when a set of homonyms can be authors of a publication *p*, disambiguation is enforced by considering the other authors of *p* and by selecting the homonym-scholar that has the largest overlap between her/his sets of co-authors in past publications and the authors of the publication *p*. Disambiguation approaches based on co-authors are ineffective in those scientific areas where single-author publications are a frequent practice, such as in humanities and social sciences. As an alternative, the use of disambiguation solutions based on keyword and linguistic analysis techniques has been also proposed (Han et al. 2005b; Tang et al. 2012). However, we stress that this kind of solutions are poorly effective on scholars that are used to publish in different languages. Disambiguation can be also characterized by the execution of a preliminary clustering step with the aim at generating groups of publications with similar authors. In (Han et al. 2005b; Torvik et al. 2005), the use of an unsupervised clustering procedure is proposed. As an alternative solution, in (Ferreira et al. 2010; Tang et al. 2012), a learning approach based on a gold dataset of disambiguated publication is enforced. In these related work, the main focus is on how to perform disambiguation from a technical point of view, but the specification of a comprehensive framework is mostly missing where all the relevant aspects from dataset acquisition to disambiguation configuration and result release are adequately addressed. In particular, what is really missing is the capability to dynamically select the appropriate disambiguation strategy in which different techniques are effectively combined on the basis of the considered disambiguation case-study.

Original Contribution of the Chapter With respect to the literature, the EVA contribution is the specification of a single, comprehensive approach for author-based disambiguation. In EVA, two different disambiguation strategies have been developed to be invoked alone or in combination according to a dynamic setup mechanism that can be configured according to the dataset to disambiguate. The EVA approach has been conceived to work with publication datasets acquired from a web repository with possible unclean and duplicate metadata such as Google Scholar. In this respect, an experimental evaluation is provided to assess the effectiveness of the overall EVA approach by comparing the results of the EVA techniques applied to a dataset retrieved from Google Scholar against the results extracted from a commercial citation database.

3 The EVA Approach

The EVA approach has been conceived to enforce data analysis functionalities on a dataset of Google Scholar publications related to a selection of scholars/researchers. The approach receives in input (i) the scholars *SD* to consider, and (ii) the publications *GS* available on Google Scholar for the researchers belonging to *SD*. The publications in *GS* are retrieved by searching on Google Scholar the full name of each researcher in *SD*. Given a scholar name $s(n) \in SD$, $GS_s \subseteq GS$ represents the set of *GS* publications retrieved for *s(n)* from Google Scholar. As a matter of fact, it is possible that "mismatching" publications, namely publications not authored by the scholar *s(n)*, are included in GS_s. For this reason, the EVA approach is characterized by the use of the following disambiguation stages (see Fig. 1).

Syntax-Based Disambiguation This step works on syntax-level ambiguities of scholar names and it is concerned with the retrieval mechanism of Google Scholar. When a "full-name query" is submitted, the set of publications returned by Google Scholar is selected according to an approximate, surname-based matching procedure. As an example, it is possible that publications with author S. Ferrara and A. Ferrara are both included in the publications returned by Google Scholar for the query Alfio Ferrara. Given a scholar name *s(n)*, the EVA step of syntax-based disambiguation has the goal to detect and remove from GS_s the retrieved publications that are mismatching on author names, even when (first and/or second) names are shortened to the initial letter. Conventional cleaning techniques based on regular expressions are employed to this end.

Author-Based Disambiguation This step works on identity-level ambiguities. Given a scholar name *s(n)*, the EVA step of author-based disambiguation has the goal to detect and remove from GS_s the retrieved publications that are authored by a homonym of *s(n)*. To this end, two different author-disambiguation strategies called

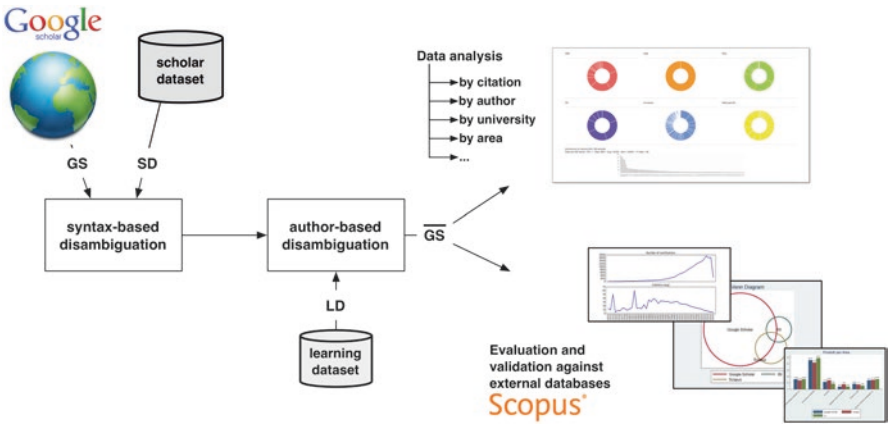


Fig. 1 The EVA approach

Table 1 The EVA case-study on Italian university researchers

Non-bibliometric area	Number of scholars
Civil engineering and architecture	1856
Historical artistic sciences	4942
History, philosophy, education and psychology	3392
Law	4706
Economics and statistics	4749
Political science	1697
Total	21,342

similarity-oriented and *specificity-oriented* have been specifically developed in EVA. These strategies represent a distinguishing feature of the EVA approach and a detailed description is provided in Sect. 4.

Syntax- and author-based disambiguation steps produce the set of publications $\overline{GS}$ as a result. In EVA, the set $\overline{GS}$ is exploited for (i) analyzing the publications authored by the scholars in *SD* (Sect. 5), and (ii) evaluating the EVA results against the Scopus citation database (Sect. 6).

Running Example The EVA project is focused on a case-study about the Italian university researchers belonging to non-bibliometric scientific areas. It is important to note that the Italian scenario is characterized by two main peculiar features. First, in Italy, the distinction between bibliometric and non-bibliometric scientific areas is defined by the Italian Ministry of University and Research (MIUR). Second, each Italian university researcher is associated with a reference scientific area. As a result, in the EVA case-study, the dataset of considered scholars *SD* has been retrieved from the MIUR by selecting the names of scholars associated with non-bibliometric scientific areas. A summary table of the scholar dataset used in the EVA case-study is shown in Table 1. For the above scholar dataset, a set *GS* containing 887,514 publications has been retrieved from Google Scholar and they have been submitted to author-based disambiguation.

As a running example throughout the chapter, we consider the following publications retrieved from Google Scholar for an Italian researcher associated with the Historical Artistic Sciences area.[1]

p_1: Renaissance Siena: Art for a City
p_2: Ludic maps and capitalist spectacle in Rio de Janeiro
p_3: Guidelines for the Management of Atrial Fibrillation

[1] The complete name of the area is Antiquities, Philological-Literacy and Historical Artistic Sciences. It has been shortened only for readability. The name of the Italian researcher is irrelevant for the clarity of the example and it is omitted.

The considered researcher has homonyms in different scientific areas, such as for example Political or Medical science. As a result, the publications retrieved from Google Scholar contain various mismatching records and the use of author disambiguation techniques is actually required. In particular, for the considered publications, only p_1 is authored by the considered Italian researcher, while p_2 and p_3 are authored by homonyms. In the following, we apply the author-based disambiguation techniques of EVA and we discuss the obtained results on the considered publications p_1, p_2, and p_3.

4 Author-Based Disambiguation Techniques

Consider a publication $p \in GS_S$ retrieved from Google Scholar in reply to a full-name query $s(n) \in SD$.[2] In EVA, the goal of author-based disambiguation techniques is to enforce an automatic decision-making mechanism for determining whether to confirm or to discard the proposed authorship relation $s(n) \longleftrightarrow p$. In other words, we aim at detecting and removing incorrect authorship relations due to homonyms on scholar names. To this end, the author-based disambiguation techniques of EVA are based on the following two requirements:

– *Scholars are associated with a reference scientific areas of research*. In EVA, for a scholar $s(n) \in SD$, we call *s(sa)* the scientific area in which *s(n)* places most of her/his publications. The specification of *s(sa)* can be exploited in different ways. As an option, the scholar $s(n) \in SD$ can manually specify her/his reference scientific area *s(sa)* within a set of available alternatives. As another option, the area *s(sa)* can be assigned to a scholar by a trusted authority (e.g., a research and education ministry, a research centre administration).
– *Scientific areas of research are associated with a dictionary of featuring keywords*. In EVA, a set of supported scientific areas $SA = \{sa_1, \ldots, sa_k\}$ is predefined and each area $sa \in SA$ is represented as a dictionary of featuring keywords.

In EVA, author disambiguation is based on the idea that an authorship relation $s(n) \longleftrightarrow p$ is confirmed when the title of the publication p is "coherent" with the terminology (i.e., the dictionary of featuring keywords) of the scientific area *s(sa)* of the scholar *s(n)*. To this end, two different disambiguation strategies called *similarity-oriented* and *specificity-oriented* have been developed in EVA. Moreover, a two-phase process articulated in *preparation* and *execution* has been specified as described in the following (see Fig. 2).

[2] In this chapter, we call publication p the bibliographic record returned by Google Scholar in response to a query.

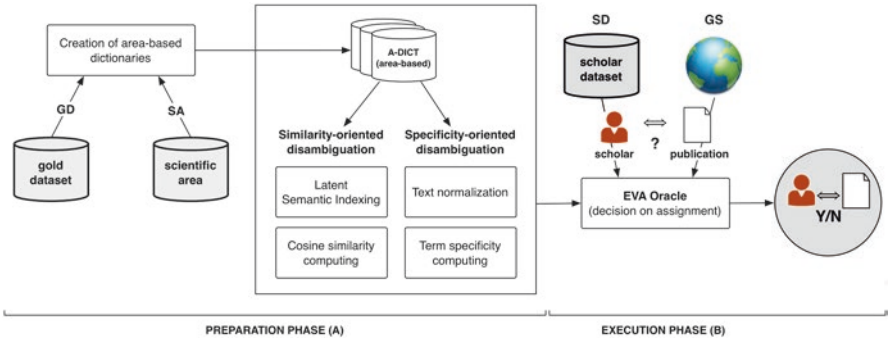


Fig. 2 The author-based disambiguation techniques of EVA

4.1 The Preparation Phase

This is a batch, preparatory phase for enabling to invoke similarity- and specificity-oriented disambiguation. Both strategies require the preliminary *creation of area-based dictionaries* in which a scientific area $sa \in SA$ is associated with a set of featuring keywords. These keywords are extracted from a *gold dataset* containing a set of publications that are expert-assigned and validated with respect to the scientific areas SA. A publication $p \in LD$ is defined as $p = < p_t, p_a >$, where p_t is the publication title and p_a is the scientific area that has been expert-assigned. Given a set of scientific areas $SA = \{sa_1, \ldots, sa_k\}$, the creation of area-based dictionaries generates a set $ADICT = \{AD_1, \ldots, AD_k\}$, where $AD_j \in ADICT$ contains the featuring keywords associated with the scientific area sa_j. The area-based dictionaries ADICT are exploited by both similarity-oriented and specificity-oriented disambiguation strategies.

Similarity-Oriented Disambiguation This strategy relies on *latent semantic indexing* (LSI) techniques, that are based on the idea to represent a potentially large set of documents over a relatively small number of considered dimensions (e.g., 400 dimensions in our case-study) (Dumais 2004). Dimensions are also called *topics* and they have the goal to make explicit the latent variables that can be inferred by observing the term distribution and co-occurrence over the considered documents. A document is created for each scientific area $sa_j \in SA$ and it corresponds to the set of featuring keywords $AD_j \in ADICT$. The set AD_j contains the titles of the publications belonging to the scientific area sa_j in the gold dataset (i.e., $AD_j = \{p_t | < p_t, p_a > \in LD \bigwedge p_a = sa_j\}$). In similarity-oriented disambiguation, the execution of LSI techniques over area-based documents produces a set of document-vectors $\overrightarrow{DOCLSI}$ as a result, where a document-vector $\overrightarrow{AD_j} \in \overrightarrow{DOCLSI}$ provides the representation of the document AD_j (and thus of the scientific area sa_j) with respect to the LSI dimensions.

Consider an authorship relation $s(n) \longleftrightarrow p$ to confirm or to discard. In similarly-oriented disambiguation, the idea is that the more the title of the publication p is similar to the document-vector $\overrightarrow{AD_j} \in \overline{DOCLSI}$ associated with the scientific area $s(sa_j)$ of the scholar $s(n)$, the higher is the likelihood that the authorship relation is correct and can be confirmed. To this end, the set of document-vectors *DOCLSI* produced as result by LSI is exploited as input for *cosine similarity computing*. Given the publication $p \in GS_s$, we call $\overrightarrow{p_t}$ the vector representation of p with respect to the document-vectors *DOCLSI*. The cosine similarity between the publication $p \in GS_s$ and the scientific area $s(sa_j)$ is calculated as follows:

$$sim\left(p,s\left(sa_j\right)\right) = \frac{\overrightarrow{p_t} \cdot \overrightarrow{AD_j}}{\overrightarrow{p_t} \quad \overrightarrow{AD_j}}$$

where $\overrightarrow{p_t} \cdot \overrightarrow{AD_j}$ is the scalar product between the two vectors $\overrightarrow{p_t}$, $\overrightarrow{AD_j}$, and $\overrightarrow{p_t}$, $\overrightarrow{AD_j}$ represent the Euclidean norm of the vectors $\overrightarrow{p_t}$, $\overrightarrow{AD_j}$, respectively.

Specificity-Oriented Disambiguation This strategy relies on *text normalization* techniques for transforming an input text into a set of string tokens (or simply tokens in the remaining of the chapter) that represent the relevant lexical elements of the input text. Normalization is enforced by applying a sequence of *natural language processing* (NLP) functions, such as for example tokenization, upper-case and elision removal, and stemming that are employed in the Italian case-study.[3] Consider the set AD_j containing the titles of the publications belonging to the scientific area sa_j in the gold dataset. In specificity-oriented disambiguation, the execution of normalization techniques over area-based documents produces a set of documents $\overline{DOCNORM}$ as a result, where a document $\overrightarrow{AD_j} \in \overline{DOCNORM}$ contains the tokens extracted from the publication titles of AD_j through the execution of normalization techniques.

Consider an authorship relation $s(n) \longleftrightarrow p$ to confirm or to discard. In specificity-oriented disambiguation, the idea is that the more the terms/tokens in the title of the publication p are "specific" of the scientific area $s(sa_j)$ (i.e., the tokens appear frequently in $\overrightarrow{AD_j}$ and rarely in other areas), the higher is the likelihood that the authorship relation is correct and can be confirmed. To this end, the set of documents $\overline{DOCNORM}$ produced as result by normalization is exploited as input for *term specificity computing*. Given the publication $p \in GS_s$, we call $\overline{p_t}$ the tokens extracted from the title of p. The specificity between the publication $p \in GS_s$ and the scientific area $s(sa_j)$ is calculated as follows:

[3] A detailed description of normalization techniques and related NLP functions is provided in (Manning et al. 2008).

$$spec\left(p,s\left(sa_j\right)\right)=\left(\sum_{i=1}^{i=r}\frac{f\left(t_i,\overline{AD_j}\right)-f\left(t_i,\overline{DOCNORM}\right)}{\sqrt{f\left(t_i,\overline{DOCNORM}\right)}}\right)^3$$

where $t_i \in \overline{p_t}$ is a token belonging to the normalized title of the publication $p \in GS_s$, r is the number of tokens in $\overline{p_t}$, $f\left(t_i,\overline{AD_j}\right)$ is the number of occurrences of the token t_i in $\overline{s\left(sa_j\right)}$, and $f\left(t_i,\overline{DOCNORM}\right)$ is the number of occurrences of t_i in all the documents of $\overline{DOCNORM}$. Negative values of $spec(p, s(sa_j))$ of means that the tokens of $\overline{p_t}$ rarely appear in $\overline{s\left(sa_j\right)}$, thus the title of the publication p is not coherent with the scientific area $s(sa_j)$. Conversely, positive values of $spec(p, s(sa_j))$ means that the tokens of $\overline{p_t}$ frequently appear in $s\left(sa_j\right)$, thus the publication p is coherent with the scientific area $s(sa_j)$.

4.2 The Execution Phase

This is an executive phase where disambiguation is invoked. This phase is based on the so-called *EVA Oracle* which is in charge of disambiguation resolution. Given a scholar name $s(n) \in SD$ and a publication $p \in GS_s$, the EVA Oracle exploits similarity- and specificity-oriented strategies to decide whether to confirm or to discard the authorship relation $s(n) \longleftrightarrow p$. Three disambiguation modalities are enforced by the EVA Oracle for taking the decision:

- *Modality-by-similarity* (*modsim*). This modality exploits the similarity-oriented disambiguation strategy. According to *modsim*, the authorship relation $s(n) \longleftrightarrow p$ is confirmed when the result of cosine similarity $sim(p, s(sa_j))$ is higher than a similarity threshold th_{sim} denoting the minimum degree of similarity required for authorship-relation confirmation. In the EVA case-study, the similarity threshold is experientially set to $th_{sim} = 0.8$.
- *Modality-by-specificity* (*modspec*). This modality exploits the specificity-oriented disambiguation strategy. According to *modspec*, the authorship relation $s(n) \longleftrightarrow p$ is confirmed when the result of the specificity function $spec(p, s(sa_j))$ is higher than a specificity threshold th_{spec} denoting the minimum degree of specificity required for authorship-relation confirmation. In the EVA case-study, the specificity threshold is experientially set to $th_{spec} = 0$.
- *Modality-by-similarity-specificity* (*modsimspec*). This modality exploits both the similarity- and the specificity-oriented disambiguation strategies. According to *modsimspec*, the authorship relation $s(n) \longleftrightarrow p$ is confirmed when both the criteria of *modsim* and *modspec* are satisfied. The relation is discarded otherwise.

4.2.1 Running Example

Consider the scholar *s(n)* associated with the scientific area s(sa) = Historical Artistic Sciences and the publications p_1, p_2, and p_3 presented in Sect. 3. According to similarity- and specificity-oriented strategies, the results of cosine similarity computing (*sim*) and term specificity computing (*spec*) are the following:

$sim(p_1,s(sa)) = 0.855$; $spec(p_1,s(sa)) = 0.0237$
$sim(p_2,s(sa)) = 0.564$; $spec(p_2,s(sa)) = 0.0125$
$sim(p_3,s(sa)) = 0.485$; $spec(p_3,s(sa)) = -0.0011$

In the EVA case-study, the modality-by-similarity-specificity is exploited for author-based disambiguation. This choice is the result of an extensive experimentation and it has been selected for enabling to obtain the best performance in terms of effectiveness. In this respect, only the authorship relation with the publication p_1 is confirmed (i.e., $sim(p_1,s(sa)) > th_{sim}$ and $spec(p_1,s(sa)) > th_{spec}$), while the relations with p_2 and p_3 are discarded.

4.3 Assessment of the EVA Disambiguation Techniques

The disambiguation techniques exploited in EVA require to properly set up a set of parameters, that are the number of dimensions for latent semantic indexing, the similarity threshold for the similarity disambiguation modality, and the specificity threshold for the specificity disambiguation modality. In order both of experimentally tuning the parameters driving disambiguation and of assessing the EVA disambiguation techniques, we build a dataset of about 350,000 publications extracted from the curriculum vitae of Italian scholars working in all the scientific areas, either bibliometric or non-bibliometric. Since those publications are taken from the curricula, the authorship relation between each publication and the corresponding scholar is correct and, as a consequence, also the relation holding between a publication and the scientific area of the author is correct.

The dataset has been then split in a *training* and a *testing* set. The training set has been used for tuning the EVA parameters, while the testing set has been used for evaluating the quality of the EVA disambiguation process.

In particular, we measured the quality of disambiguation by computing the *precision* and *recall* of the disambiguation process. The precision is the fraction of authorship relations that are confirmed by EVA that are actually correct. The recall is the fraction of the authorship relations of an author that are actually confirmed by EVA. In order to set up the parameters, we executed several tests with different combinations of parameter values over the training set of authorship relations and we measured precision and recall. As a result, we empirically set up to 400 the number of dimensions of latent semantic indexing, to 0.8 the similarity threshold,

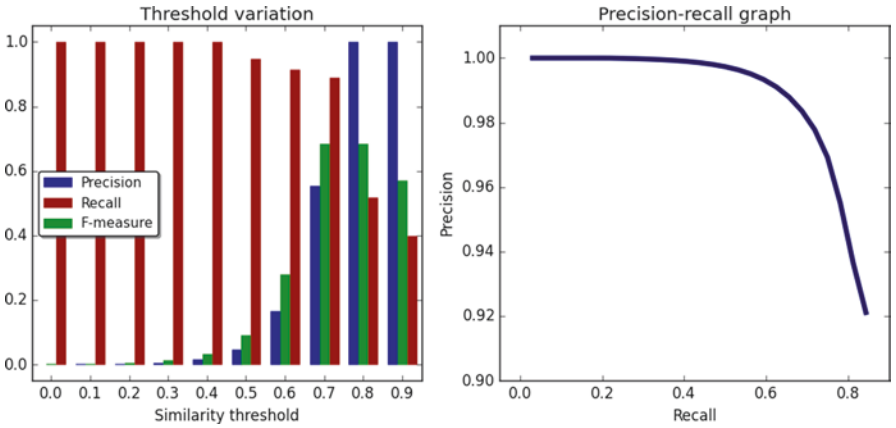


Fig. 3 Precision and recall of the EVA disambiguation techniques

and to 0 the specificity threshold. In particular, we observed that parameter that mainly affects the quality of disambiguation is the similarity threshold. We report the values of precision, recall and F-measure (i.e., the harmonic mean of precision and recall) collected at different levels of similarity in Fig. 3. We note that the balance of precision and recall, represented by the F-measure, reaches the highest values with similarity threshold values of 0.7 and 0.8. We choose 0.8 because with that value of similarity threshold, precision is maximal. This choice is motivated by the fact that we preferred to maximize the correctness of the authorship relation (i.e., precision) rather than the coverage of all the author publications. However, using the testing set we correlated the levels of precision and recall achieved by the EVA disambiguation techniques by means of the Precision-Recall graph (see Fig. 3). The graph report the behavior of the disambiguation precision at different levels of recall. In particular, we note how the EVA disambiguation techniques are capable of correctly validate more than 90% of authorship relations (0.9 precision) by covering correctly more than 80% of author publications (0.8 recall).

5 Evidence from Italian Academia

The approach of adopting Google Scholar to evaluate non-bibliometric disciplines proposed in this chapter may also contribute to the description of the overall Italian research activity in these fields in the last decades and to testing indirectly the validity of a bibliometric-style approach to analysis in fields in which the international scientific journals are not the main vector to disseminate research findings. The main focus of this section is to provide a synthetic description of recent research trends in non-bibliometric disciplines in Italian academia over the last few years.

Table 2 Descriptive statistics

Discipline	Stats	Publication (Sum)	Citation (Avg)	Citation (Sum)	H Index	Academic Seniority
Civil engineering and architecture	Mean	6.69	4.8	18.64	1.25	14.17
	Std	7.71	24.56	65.05	1.58	16.89
	N. Obs.	1353	1353	1353	1353	1286
Antiquities, philological-literacy and historical artistic sciences	Mean	8.41	4.83	36.55	1.59	17.8
	Std	12.79	20.35	332.6	2.14	18.32
	N. Obs.	3836	3836	3836	3836	3.742
History, philosophy, education and psychology	Mean	10.72	5.75	63.19	2.23	18.1
	Std	13.19	10.83	310.41	3.13	20.23
	N. Obs.	2756	2756	2756	2756	2756
Law	Mean	7.73	4.46	22.29	1.49	14.76
	Std	11.61	13.92	81.33	1.7	19.64
	N. Obs.	3071	3071	3071	3071	2991
Economics and statistics	Mean	22.31	13.88	281.97	5.26	16.51
	Std	25.82	31.34	996.39	5.17	17.02
	N. Obs.	4224	4224	4224	4224	4192
Political science	Mean	12.6	8.73	99.92	2.9	16.99
	Std	15.4	18.33	345.01	3.3	17.34
	N. Obs.	1427	1427	1427	1427	1404
Total	Mean	12.41	7.54	104.5	2.69	16.6
	Std	17.92	22.18	562.02	3.67	18.44
	N. Obs.	16,667	16,667	16,667	16,667	16,306

We collapsed the publication records obtained from GS at the researcher level to describe the current status of the research in non-bibliometric disciplines published by academics who were active in the Italian higher education system on 31 December 2014. The descriptive statistics (means, standard deviations and numbers of observations) are reported in Table 2 for the selected disciplines. A first look at the data provides a general idea of great heterogeneity across disciplines, with, on the one hand, the average researcher in economics and statistics publishing 22 research papers or products (e.g. conference proceedings, book chapters or journal articles) indexed by Google Scholar with on average 14 citations each and having an H-Index of 5 after 16 years of research activity and, on the other hand, the representative researcher in law publishing fewer than 8 papers with 4 citations (on average) and having an H-Index of 1:5 after 14 years of activity. However, considering the fact that the data that we used are essentially count data resulting from the collapse of repeated events at a certain point in time (e.g. at the lower level the number of citations of an article while at the aggregate level the cumulative or mean citation counts of individual publications, etc.) analysed by groups of researchers (e.g. researchers within a specific discipline), examining the simple descriptive statistics of a selected set of indexes could clearly be misleading (Bornmann et al. 2008). In fact, the assumption of normally distributed data usually required by common statistical

tests is clearly violated (count data are better represented by Poisson or negative binomial distribution) and a simple means comparison may lead to a distorted picture when comparing different research subsets of individuals. For these reasons simple box plot diagrams (Fig. 3) may be a more convenient way of visually summarizing the differences across discipline distributions. Indeed, outliers usually provide important information on very productive (inactive) academics or highly (lowly) cited scholars, such as academic stars or inactive individuals, even if they cannot depict the entire output of a research group.

Figure 3 shows the box plots of the H-Index and of the total number of EVA publications by discipline. The smallest observations have zero citations and the H-Index for all the disciplines in both cases as well as the lower quartiles (the maximum values of the intervals containing the less productive and less cited 25% of all academics) are close to zero except for economics and statistics and political sciences. In addition, the medians (the maximum value of the interval containing 50% of all observations) and upper quartiles (containing 75% of all observations) are quite different for economics and statistics and political sciences if compared with the other non-bibliometric disciplines conveying the impression of a research attitude that is clearly situated in the middle between hard sciences and arts and humanities (Checchi et al. 2014). Finally, any observed value outside the ends of the whiskers is considered unusual or an academic outlier.

When the academic discipline is our main unit of analysis, some measures of concentration (e.g. Lorenz curves and Gini indexes) may also be calculated to distinguish between research disciplines exhibiting some sort of "collective strength" and groups with more "individual strength" (Daniel and Fisch 1990; Burrell 2006). To describe these pattern of research better, a set of Lorenz curves representing the cumulative number of papers published against the cumulative number of citations of researchers belonging to each of the selected disciplines is provided in Fig. 4. In this way, the Lorenz curves capture the degree of concentration with the implicit assumption that each individual researcher in a specific discipline contributes to an equal share of the total number of citations. If each researcher had an equal value in the shares of the total citations of the discipline or of the H-Index, the Lorenz curve would result as a straight line (the diagonal) reflecting the pattern of perfect equality.

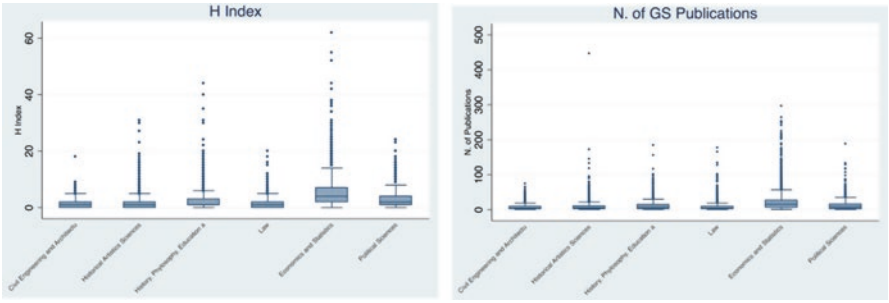


Fig. 4 Boxplots of the H-Index and number of EVA products

Otherwise, if the observed Lorenz curve deviates from the diagonal line, the researchers do not contribute equally to the total number of citations or the H-Index. Figure 4 clearly show that, in the non-bibliometric disciplines of Italian academia, this is not the case.

For example, for civil engineering and architecture and for economics and statistics, the 20% most productive scientists account for about 35% of the total citations in their fields, while for the other disciplines the inequality is a little higher, with 20% of the most productive academics receiving around 45% of the EVA citations in their respective field. The research impact, as measured by the total citations, is more concentrated in a relatively small group of researchers in arts and humanities with respect to economics and statistics or engineering. Overall it seems that in non-bibliometric disciplines "collective strength" is the common pattern instead of situations of "individual strength", with large fractions of citations relating to very small fractions of researchers.

6 External Validity Assessment of the EVA Project

A comparative evaluation of the EVA project data using established bibliometric data sources is also useful for assessing the external validity of the project and discussing the empirical evidence regarding the coverage and reliability of the bibliometric indexes computed using the EVA data set. The most reliable benchmark data source should be the ANPREPS database "National Archive of Professors and Researchers scientific publications" containing the entire academic production of Italian professors as prescribed by the Law 1 = 2009 (art. 3 bis-2), but unfortunately this database has never been realized in Italy. Hence, we focus on international publications only as collected by the two commercial databases ISI-Thomson and Scopus-Elsevier (Fig. 5).

We collected publications of the 16,667 Italian academics active at the end of 2014 in non-bibliometric fields as they were defined by MIUR. The substantive difference between these resources is mainly related to the difference in the extent of coverage among research disciplines. On the one hand, a better representation of publications in arts and humanities is guaranteed by Scopus-Elsevier, while, on the other hand, ISI-Thomson is preferable for more scientific sub-disciplines, such as statistics or psychometrics. For these reasons each database is peculiar and it is necessary to take into account its own characteristics when it is used as reference point. In general, the comparisons with EVA provided in this section must therefore be considered as relative comparisons between benchmark databases instead of absolute comparisons with respect to a true reference point.

Table 3 shows the relative composition of the three data sets by discipline. A similar pattern can be identified between EVA, Scopus and ISI publications collected for Italian academics. About 45% of the overall number of published papers is authored by researchers in the economics and statistics fields, around 15% by authors in history, philosophy, education and psychology, 10% in law and 15% in

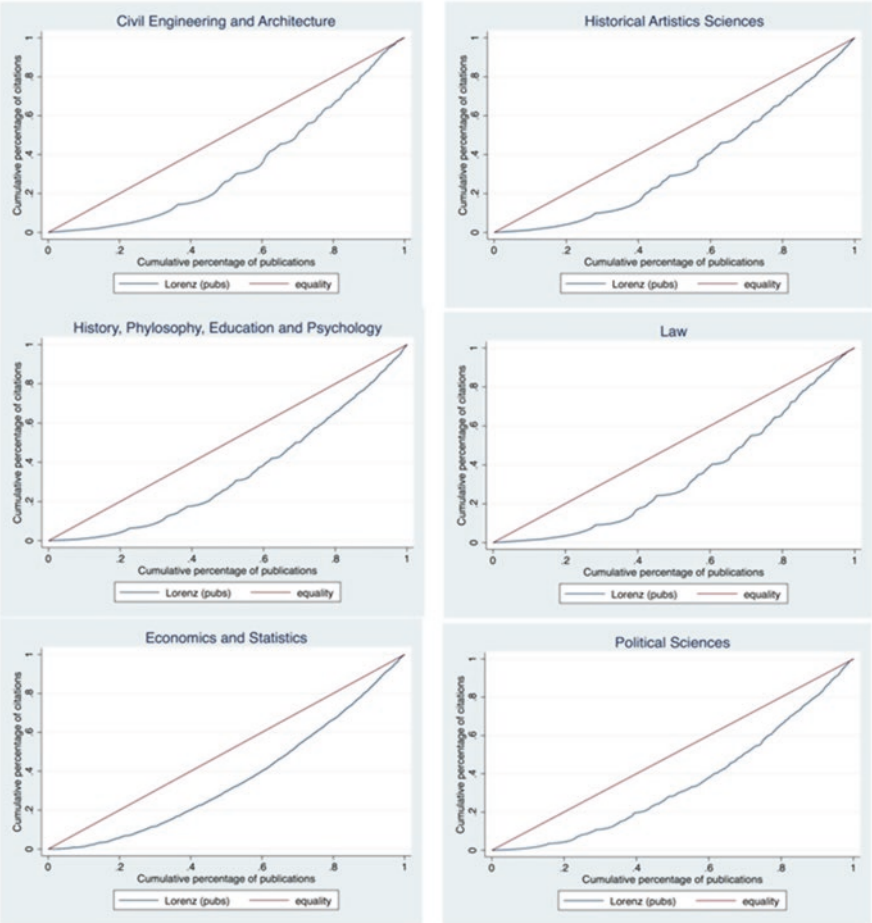


Fig. 5 Lorenz Curves by Discipline

Table 3 Relative frequencies of papers by Source

Discipline	EVA	Scopus	ISI
Antiquities, philological-literacy and historical arts sciences	15.68	13.87	15.98
Economics and statistics	45.72	41.85	49.07
Law	11.56	14.09	8.87
Civil engineering and architecture	4.32	7.92	4.42
Political science	8.62	7.51	5.69
History, philosophy, education and psychology	14.1	14.76	15.97
Total	100.0	100.0	100.0

antiquities, philological-literacy and historical artistic sciences. ISI-Thomson seems to be more biased towards disciplines such as economics and statistics (49% vs 45% of EVA and 41% of Scopus) and history, philosophy, education and psychology (16% versus 14%), while it is less representative of the arts and humanities (e.g. 8% in law versus 11% of EVA and 14% of Scopus). In addition, Scopus is more oriented towards arts and humanities, with a larger fraction of papers in law (14% versus 11% of EVA and 9% of ISI) and in civil engineering and architecture (8% versus 4% of the others). To this end, EVA can be considered as the most balanced source, with a good degree of coverage of all the non-bibliometric disciplines.

In the previous section, we described the main features of the research produced by Italian academics in non-bibliometric disciplines that emerged when adopting the EVA approach. However, the external validation of EVA requires us to compare each paper collected by following the suggested approach with the whole sets of papers in the ISI and Scopus repositories to evaluate their degree of completeness. To this end, some matching algorithms were employed to check whether an EVA record was also collected by the other sources or not. Indeed, differences in titles' syntaxes or in authors' names between the three sources may alter the results of standard matching procedures. Table 4 shows that a conflict arises when the same bibliographic record published in 2014 by "Amendola A." and co-authors is collected with two similar "but not equal" strings for the title in EVA (Google Scholar) and Scopus. In this case, exact matching would fail to recognize the same paper and a less precise matching criterion is required.

In contrast, Table 5 suggests that calibrating the matching algorithm of titles' strings is a considerable challenge given the fact that less stringent criteria may be misleading as well. Table 5 below reports four papers, two collected by EVA and two by Scopus, with the same title in three cases out of four and different publication years (2012, 1997 and 2006), for which it is very difficult to calibrate the matching algorithm (whereby stringent and relaxed criteria produce very different results).

Table 4 Example of matching conflict between different data sources

Id paper	Title	Year	Author	Source
34,889	CFE network: The annals of computational and financial econometrics	2014	Amendola, A. Et al.	EVA
178	CFE network: The annals of computational and financial econometrics: 2nd issue	2014	Amendola, A. Et al.	Scopus

Table 5 Example of matching conflict within the same data source

Id Paper	Title	Year	Source
28,635	15 comparative law and economics	2012	EVA
28,667	Comparative law and economics	1997	EVA
51,885	Comparative law and economics	2006	Scopus
51,886	Comparative law and economics	2012	Scopus

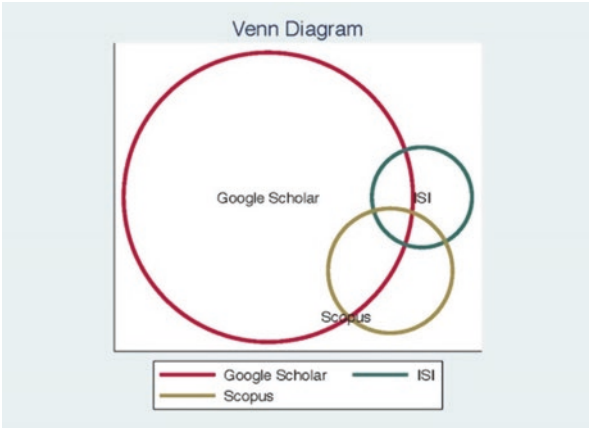


Fig. 6 Venn diagram

Table 6 Intersection of data sources by field

Discipline	EVA in Scopus	EVA in ISI	EVA in ISI and Scopus
Civil engineering and architecture	5.6	1.26	0.87
Antiquities, philological-literacy and historical arts sciences	1.69	0.76	0.32
History, philosophy, education and psychology	4.86	2.24	0.87
Law	1.06	0.36	0.15
Economics e statistics	15.8	7.89	5.21
Political sciences	4.33	1.56	0.75
Total	8.91	4.27	2.67

For the external validity assessment exercise of EVA, we adopted a conservative approach with a highly selective heuristic matching algorithm that guarantees a high degree of reliability of the corresponding associations of records between different data sources. The selected procedure, for example, discarded all the possible associations reported in Table 5 requiring a higher level of correspondence between the title strings. The association of two records in different data sources requires two conditions: a maximum one-year lag between their publication years and an edit distance lower than 2.[4]

A Venn diagram of the matched papers following the described procedure between the three sources is proposed in Fig. 6. It represents the universe of all publications (before 2012) collected by the three data sets and their set representation derived by the adoption of the matching algorithm. It shows that the fraction of papers shared by EVA and Scopus is conditioning by the number of papers in EVA larger than the one shared with ISI (9% versus 4%), but both are subsets of limited size with a very small

[4]The edit distance between two titles' strings (S_1 and S_2) is the minimum number of operations (inclusion, substitution or deletion) on single characters needed to transform S_1 into S_2.

intersection (2:6% of the EVA's size). Surprisingly, the intersection between ISI and Scopus accounts for around 30% of ISI and 15% of Scopus products.

If we consider the intersections of EVA with both ISI and Scopus separately, we observe some peculiar differences. Table 6 shows the existence of large heterogeneity across the disciplines, with a relatively larger fraction of EVA publications shared with Scopus (15%), ISI (8%) and both (5%) for individuals in the field of economics and statistics.

In addition, Fig. 7 confirms for the EVA data set the validity of one of the most common empirical laws in bibliometric disciplines: Lotka's law. It describes the frequency of publication by authors, showing that a large fraction of papers is authored by a small number of researchers. The distribution of the number of authors against the number of contributions made by the authors is highly asymmetric, with a higher concentration of articles among a few authors (great producers), while the remaining articles are distributed among several authors. The empirical estimate of Lotka's law between Scopus and EVA is similar, while it is steeper in ISI, identifying a substantially robust picture of the non-bibliometric research produced in Italy in the observed period. Moreover, Fig. 8 shows that the citation distributions are quite similar, because they do not particularly differ among the sources analysed.

Finally, we analyse the citation counts of each paper as they emerged from different data sources. First, we notice a strongly positive set of correlations between the citation counts, which are all higher than 0.85 and statistically significant (Tables 7 and 8).

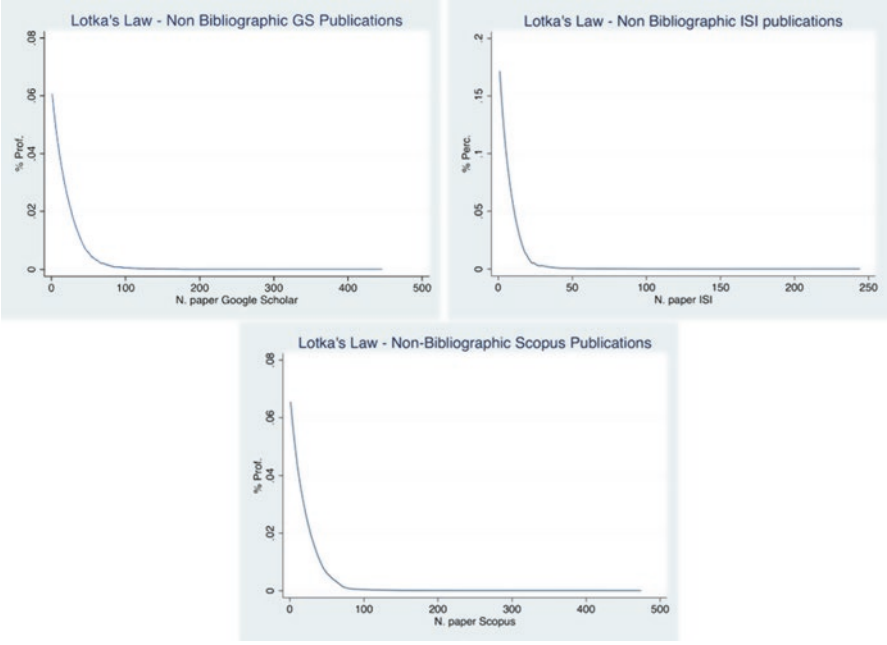


Fig. 7 Lotka Law by bibliometric source

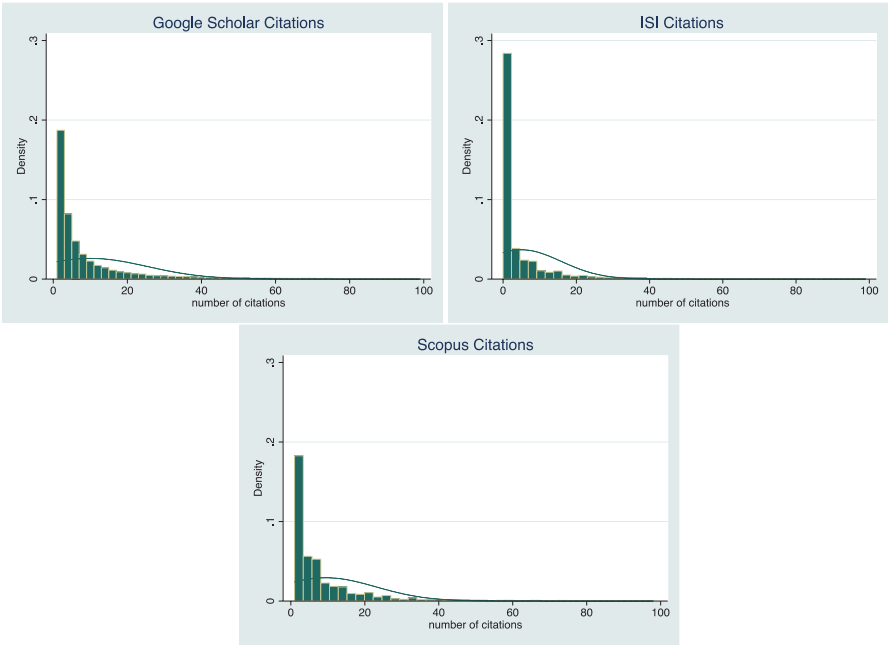


Fig. 8 Distribution of citations by bibliometric source

Table 7 Correlation matrix of citations

Correlation matrix	EVA	Scopus	ISI
EVA	1		
Scopus	0.8749*	1	
ISI	0.8510*	0.9171*	1

In general, it is very interesting to notice that the greater correlation between paper citations is associated with the publications that are simultaneously indexed in both ISI and Scopus (0.91), which account for a small fraction of the EVA products, but both refer to commercial resources. However, the correlations between citations received by papers collected simultaneously in Scopus and EVA as well as in ISI and EVA are both larger than 0.85. In particular, this applies to the economics and statistics disciplines. Finally, the slopes of the three lines in Fig. 9 evidence that the ratio between the citations received in Scholar and in ISI is lower (1/3) than the ratio of the citations collected from Scholar and Scopus (1/2).

Table 8 Correlation matrix of citations by field

Correlation Matrix	EVA	Scopus	ISI
Civil engineering and architecture			
EVA	1		
Scopus	0.8986*	1	
ISI	0.7139*	0.7280*	1
Antiquities, philological-literacy and historical arts sciences			
EVA	1		
Scopus	0.8533*	1	
ISI	0.9090*	0.7680*	1
History, philosophy, education and psychology			
EVA	1		
Scopus	0.9490*	1	
ISI	0.8526*	0.8312*	1
Law			
EVA	1		
Scopus	0.5713*	1	
ISI	0.7257*	0.6294	1
Economics and statistics			
EVA	1		
Scopus	0.8790*	1	
ISI	0.8603*	0.9206*	1
Political sciences			
EVA	1		
Scopus	0.9128*	1	
ISI	0.7207*	0.7747*	1

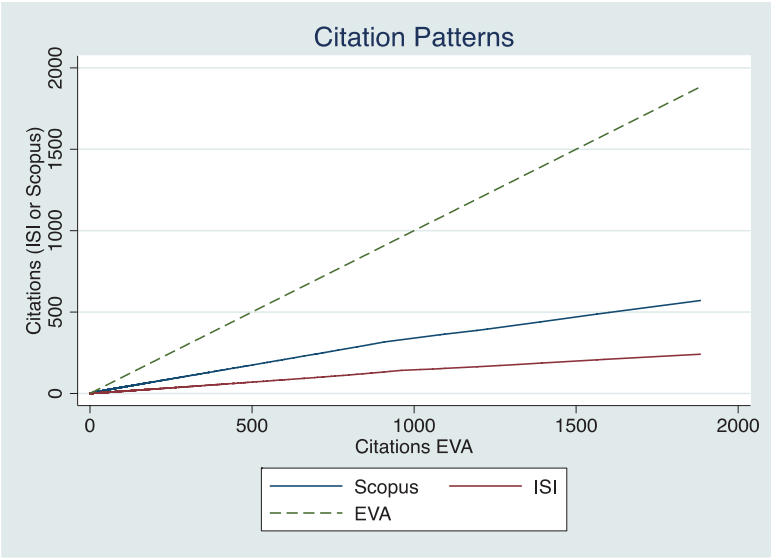


Fig. 9 Lotka Law by Bibliometric Source

7 Concluding Remarks

The main research question motivating the EVA project concerns the use of Google Scholar as a reference bibliographic database for those research areas, mainly humanities and social sciences, where the use of citation-based bibliometric indexes is not a common practice. The use of bibliometrics in general and the use of citations in particular, either for research assessment or just as a mean for achieving a better insight of the disciplines, is a main issue of discussion among social science and humanities researchers. Several authors raised epistemological objections motivated by the nature of the scientific communications in these areas and the kind of publications practices, such as the low proportion of journal articles, the literature post-publication citation rate and the local relevance of social sciences and humanities knowledge (Archambault and Larivière 2010; Prins et al. 2016). However, EVA focused the attention on a second set of objections, based on the idea that the coverage of these areas by the databases commonly used for hard sciences is largely inadequate to represent the scientific production of researchers operating in social sciences and humanities. The EVA results show that Google Scholar can be an alternative for assessing the scientific research in non-bibliometric areas, but only by accurately using suitable techniques for data analysis and quality verification. A first remarkable limit in using Google Scholar is due to the Google terms of service, which clearly state that Scholar is a service for searching data and it is not intended nor usable as a database for downloading data. This means that analysis must be limited to search results provided by Google Scholar online. As a consequence, Google Scholar is mainly usable for analysis of research at the individual level or when dealing with small groups of researchers such as Departments of small institutions. Larger collections of products can be analyzed as well, but this requires time for accessing records online and collecting complete analysis results. Moreover, the data access limitations require also to get as much information as possible from the search answer page provided by Scholar rather than by the complete publication record. A consequence of this limitation is that the information concerning the publication venue and type is very hard to achieve and it is not reliable is several cases. Despite these limits, however, it is indisputable that Google Scholar is an invaluable source of information for what concerns citations. Our results show that less than 10% of publications retrieved from Scholar are available on Scopus and less than 5% of ISI, with some remarkable differences among scientific areas. Proportionally, the number of citations from Scholar is also higher due to the fact that the number of citing publications indexed by the Scholar database is definitively larger than Scopus and ISI. From the statistical point of view, we observe that on large groups of publications there is a good level of correlation between citation distribution in the three databases. However, there are differences at the individual level. As a consequence, the picture we take of the scientific production of individuals and small groups from Google Scholar is completely different from Scopus or ISI in many cases and provides a more realistic insight of the scientific production of researchers, especially for humanities, law and some fields of political sciences.

Finally, we believe that in dealing with bibliometric analysis the crucial issue, even more decisive than having correct measures or indexes, is the quality of data. The main issue about data quality with Google Scholar is to correctly disambiguate the authorship of publication records. When searching for author names on Google Scholar, only the 23% of records retrieved can be correctly attributed to the author. Less than 10% of these can be easily disambiguated by relying on the author name format, but the majorities are due to real ambiguities due to homonymy, which makes the disambiguation process difficult. When working with tools for research assessment based on Google Scholar, this should be the main issue about the trustworthiness of the results. Solutions proposed in literature are mainly based on co-authorship relations which are rare and often untrustworthy when dealing with social sciences and humanities, where it is a common practice to publish work with no more than one or two authors. In EVA, we proposed a new solution based on the analysis of terminology, which takes into account both latent semantics and the specificity of terms with respect to scientific areas. The EVA techniques have been proved to be effective in disambiguating authorship in different areas independently from the publication language. Moreover the EVA system can be easily tuned for the purpose of achieving the required level of precision and recall. Our future work will be devoted to further improve disambiguation. The idea is to start from known publications of a given author in order to model language specificity at the level of single authors and not only of scientific areas. Such an improvement will be used to achieve disambiguation also for that authors working in different areas but on interdisciplinary fields.

References

Aguillo, I. F. (2012). Is Google scholar useful for bibliometrics? A webometric analysis. *Scientometrics, 91*(2), 343–351.

Archambault, É., & Larivière, V. (2010). The limits of bibliometrics for the analysis of the social sciences and humanities literature. *World Social Science Report*, 251–254.

Archambault, É., Vignola-Gagne, E., Côtè, G., Larivire, V., & Gingrasb, Y. (2006). Benchmarking scientific output in the social sciences and humanities: The limits of existing databases. *Scientometrics, 68*(3), 329–342.

Biolcati-Rinaldi, F., Ferrara, A., Pinotti, L., & Salini, S. (2012). Lesson learned by Unimival researchers during the comparative bibliometric analysis project (ABC). in *Rassegna Italiana di Valutazione*, v. XVI, n. 52, pp. 81-100, ISSN 1826-0713.

Bornmann, L., Mutz, R., Neuhaus, C., & Daniel, H. D. (2008). Citation counts for research evaluation: Standards of good practice for analyzing bibliometric data and presenting and interpreting results. Ethics in science and environmental politics, 8(1), 93–102.

Burrell, L. Q. (2006). Measuring concentration within and co-concentration between informetric distributions: An empirical study. *Scientometrics, 68*(3), 441–456. https://doi.org/10.1007/s11192-006-0122-0.

Checchi, D., De Fraja, G., & Verzillo, S. (2014). Publish or Perish: An Analysis of the Academic Job Market in Italy. Tech. Rep. Discussion Paper 10084, CEPR Discussion Paper.

D'Angelo, C. A., Giuffrida, C., & Abramo, G. (2011). A heuristic approach to author name disambiguation in bibliometrics databases for large-scale research assessments. *Journal of the American Society for Information Science and Technology, 62*(2), 257–269.

Daniel, H. D., & Fisch, R. (1990). Research performance evaluation in the german university sector. *Scientometrics, 19*(5), 349–361. https://doi.org/10.1007/BF02020698.

Dumais, S. T. (2004). Latent semantic analysis. *Annual Review of Information Science and Technology, 38*(1).

Falagas, M. E., Pitsouni, E. I., Malietzis, G. A., & Pappas, G. (2008). Comparison of PubMed, Scopus, web of science, and Google scholar: Strengths and weaknesses. *The FASEB Journal, 22*(2), 338–342.

Ferrara, A., & Salini, S. (2012). Ten challenges in modeling bibliographic data for Bibliometric analysis. *Scientometrics, 93*(3), 765–785.

Ferreira, A. A., Veloso, A., Gonçalves, M. A., & Laender, A. H. (2010). Effective self-training author name disambiguation in scholarly digital libraries. In *Proceedings of the 10th annual ACM joint conference on digital libraries* (pp. 39–48). Aarhus.

Garfield, E. (1980). Is information retrieval in the arts and humanities inherently different from that in science? The effect that ISI©'S citation index for the arts and humanities is expected to have on future scholarship. *The Library Quarterly*, 40–57.

Han, H., Giles, L., Zha, H., Li, C., & Tsioutsiouliklis, K. (2004). Two supervised learning approaches for name disambiguation in author citations. In *Proceedings of the 4th joint ACM/IEEE conference on digital libraries (JCDL 2004)* (pp. 296–305). Tucson.

Han, H., Xu, W., Zha, H., & Giles, L. (2005a). A hierarchical naive Bayes mixture model for name disambiguation in author citations. In *Proceedings of the ACM symposium on applied computing* (pp. 1065–1069). Santa Fe.

Han, H., Zha, H., & Giles, L. (2005b). Name disambiguation in author citations using a Kway spectral clustering method. In *Proceedings of the 5th joint ACM/IEEE conference on digital libraries (JCDL 2005)* (pp. 334–343). Denver.

Kousha, K., & Thelwall, M. (2007). Google scholar citations and Google web/URL citations: A multi-discipline exploratory analysis. *Journal of the American Society for Information Science and Technology, 58*(7), 1055–1065.

Linmans, A. J. (2010). Why with bibliometrics the humanities does not need to be the weakest link. *Scientometrics, 83*(2), 337–354.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval* (Vol. 1). Cambridge: Cambridge University Press.

Nederhof, A. J. (2006). Bibliometric monitoring of research performance in the social sciences and the humanities: A review. *Scientometrics, 66*(1), 81–100.

On, B. W., & Lee, D. (2007). Scalable name disambiguation using multi-level graph partition. In *Proceedings of the SIAM International conference on data mining* (pp. 575–580). Minneapolis, Minnesota.

Prins, A. A., Costas, R., van Leeuwen, T. N., & Wouters, P. F. (2016). Using google scholar in research evaluation of humanities and social science programs: A comparison with web of science data. *Research Evaluation, 25*(3), 264–270.

Smalheiser, N. R., & Torvik, V. I. (2009). Author name disambiguation. *Annual Review of Information Science and Technology, 43*(1), 1–43.

Tang, J., Fong, A. C. M., Wang, B., & Zhang, J. (2012). A unified probabilistic framework for name disambiguation in digital library. *IEEE Transactions on Knowledge and Data Engineering, 24*(6), 975–987.

Torvik, V. I., Weeber, M., Swanson, D. R., & Smalheiser, N. R. (2005). A probabilistic similarity metric for Medline records: A model for author name disambiguation. *Journal of the American Society for Information Science and Technology, 56*(2), 140–158.

Yang, K. H., Peng, H. T., Jiang, J. Y., Lee, H. M., & Ho, J. M. (2008). Author name disambiguation for citations using topic and web correlation. In *Proceedings of the 12th European conference on digital libraries* (pp. 185–196). Aarhus, Denmark.

Assessing the Reliability and Validity of Google Scholar Indicators. The Case of Social Sciences in Italy

Ferruccio Biolcati-Rinaldi, Francesco Molteni, and Silvia Salini

1 Introduction

One of the main debates in bibliometrics concerns its applicability to the fields of social sciences and humanities (SSH). The main problem stems from the structure of the scientific literature that characterises SSH and that reflects its peculiarities (Nederhof 2006). If we can assume that articles published in international journals are substantially comprehensive of the natural and life sciences literature, for SSH it is necessary to consider other communication channels of scientific literature (Hicks 2004).

Social sciences literature is not limited to what is published in international journals (Moed 2005, 42 and 122–131; Hicks 2004, 475–477), and it should be supplemented with the literature published through other scientific channels. One of these channels is certainly represented by books, which not only constitute a scientific research outcome more used in SSH than in other fields, but which also have more impact (in terms of citations) than articles. In addition, the relatively small values of the correlation coefficients between the citations from books and from articles support the hypothesis that these are somewhat two distinct forms of literature, although interdependent and with some overlaps (Hicks 2004, 481–482; Clemens et al. 1995).

Another important communication channel for SSH is represented by national journals. This is an obvious consequence of the fact that the subjects of interest are often nationally (if not locally) framed as well as the fact that the stakeholders

F. Biolcati-Rinaldi (✉) • F. Molteni
Dipartimento di Scienze sociali e politiche, Università Statale di Milano, Milan, Italy
e-mail: ferruccio.biolcati@unimi.it; francesco.molteni@unimi.it

S. Salini
Dipartimento di Economia, management e metodi quantitativi, Università Statale di Milano, Milan, Italy
e-mail: silvia.salini@unimi.it

© Springer International Publishing AG 2018

A. Bonaccorsi (ed.), *The Evaluation of Research in Social Sciences and Humanities*, https://doi.org/10.1007/978-3-319-68554-0_13

(politics, media, public opinion) with whom researchers relate are national and local. While international journals are often included in Web of Science (WoS) or Scopus, national journals seldom are. It follows that national literature is not adequately represented in commercial bibliographies, with the exception of the US and UK. However, it should be considered that things are rapidly changing. The internationalisation of social sciences is increasing due to powerful forces such as globalisation processes, digital innovation, European funding mechanisms that require the establishment of international partnerships, and national assessment exercises that emphasise international publications (Hicks 2004, 484–489).

In recent years, both Web of Science and Scopus have opened themselves to these other channels of disseminating scientific literature, although in different ways. Nonetheless, the highest expectations for the resolution of these problems are placed by many researchers on Google Scholar. In particular, it is assumed that Scholar is a more extensive bibliography, able to cover most of the literature of SSH. It is also assumed that the data quality is still good enough to ensure the validity and reliability of bibliometric indicators derived from this source. This chapter aims to verify this hypothesis in the case of social sciences in Italy, combining the methodological practices for the assessment of indicators' reliability and validity with the appropriate data processing and data analysis procedures.

The chapter is divided into six sections, including the introduction and conclusions. The following section details the research design, introducing the description of the process of scientific production, the definitions of coverage, the reliability and validity and the field of application of the chapter, i.e., a subset of the social and political sciences in Italy. Then the data sources analysed in the chapter (Google Scholar, Scopus, Web of Science) are discussed in the third section. The fourth section will focus on the definition of the dataset used for the analysis and on the problem of Scholar coverage, while the fifth section presents the results of estimates of reliability and validity of the bibliometric indicators drawn from Scholar.

2 Research Design

The aim of this section is to define the main coordinates behind the analysis, namely the various dimensions of the scientific production and the related bibliometric indicators, the desirable properties of such indicators and a description of the field of application of the analysis.

Our working hypothesis relies on three different dimensions of scientific production: output, recognition and relevance (of the kind and venue of the publication).

By *output* we mean the ability of the researcher to transform his or her scientific work into the main research product, that is, publications. Using this kind of definition, we aim to distinguish between the quantitative aspects of research work (the number of publications) and the qualitative ones (their scientific value). These represent different dimensions of scientific production, and the relationship between the two needs to be empirically assessed (Martin and Irvine 1983, 65–66; Research

Evaluation and Policy Project 2005, 12). Already in the sixties, Cole and Cole (1967) pointed out this distinction, developing a typology of scientists based on the intersection between quantity and quality of scientific production. Starting from a sample of 120 university professors of physics in the United States, they identified four different typologies; two of them were identified by the convergence of the two dimensions ("the prolific": high quantity and quality; "the silent": low quantity and quality) and the other two by the divergence ("the mass producer": high quantity and low quality; "the perfectionist": low quantity and high quality). Indicators of scientific output are, for example, the number of publications in a given period in the different data sources, the number of books, chapters, articles etc.

With the second dimension, we adopt the interpretation of citations as a measure of scientific *recognition*. Relying on the normative structure of science described by Merton (1973), recognition represents the main incentive and reward mechanism for researchers. The main goal for scientific researchers is indeed to be recognised as competent members of a scientific community and possibly to have a central role in this community. There are several ways to achieve this scientific recognition – eponymy, awards, being part of an editorial committee, panel of referees and so on – but the primary means is surely being quoted in a scientific publication. Citations are also frequently used as impact indicators for scientific publications (Martin and Irvine 1983, 67–72), but here we intend to adopt the perspective of scientific recognition. This approach, being more general, permits consideration of the heterogeneity behind the reasons for citations, whereas the approach relying on impact implicitly presumes a rational approach that considers citations as being used strictly for scientific argumentation.[1] Examples of indicators of scientific recognition are the number of citations per publication and the author's H-index.

The third dimension (*relevance*) relies on the importance given by the scientific community to the various kinds of publications (books, articles, conference proceedings, maps, patents etc.) or to the various venues (journals, publishers, book series etc.). Regarding the latter, the underlying assumption is that as the relevance of the publication venue increases, the selection criteria become more compelling; it follows that the acceptance rate can somehow indicate the quality of publication. At the same time, publication in a relevant venue increases the odds of being read and quoted. With some simplification, we can state that the main objective of this kind of exercise is to provide SSH with an indicator similar to the *Impact Factor*, built upon peer assessment instead of journal citations. The publication of an article in a journal classified as Class A, according to the list published by ANVUR under the National Scientific Habilitation, is an example of an indicator of relevance.

For each of these three dimensions, it is possible to identify a series of indicators. These indicators should hold some desirable properties (reliability and validity), which we will try to assess. In addition, we will try to estimate the coverage of the scientific production provided by the different bibliometric datasets.

[1] For an overview of the debate, see Moed (2005, 193–208) and the Research Evaluation and Policy Project (2005, 12–14).

The analysis of the coverage, reliability and validity of bibliometric indicators will follow a two-level structure: a base level concerning the publications and an aggregate level concerning the authors. The base level about publications will permit an evaluation of the degree of coverage of the sources, using cross-comparison checks. The aggregate level will permit the assessment of the reliability and validity of the indicators.

According to Corbetta (2003, 81)

> Reliability has to do with the 'reproducibility' of the result, and marks the degree to which a given procedure for transforming a concept into a variable produces the same results in tests repeated with the same empirical tools (stability) or equivalent ones (equivalence).

Following this definition, bibliometric indicators, which are basically procedures for the translation of the different dimensions of scientific production into variables, can be considered reliable if they provide the same results with equivalent tools: in this case, with different data sources. The notion of reliability should not be confused with validity, which is defined as follows.

> Validity, on the other hand, refers to the degree to which a given procedure for transforming a concept into a variable actually operationalizes the concept that it is intended to (Corbetta 2003, 81).

Also in this case, there are different procedures with which to evaluate the validity of indicators. Herein we propose a validation criterion that tests "the correspondence between the indicator and an external criterion that, for some reason, is deemed to be correlated with the concept" (Corbetta 2003, 83). This test can be applied to all the dimensions linked to the quality of scientific production (recognition and relevance) using Research Quality Evaluation (*Valutazione della Qualità della Ricerca* – VQR 2004–2010) individual scores or National Scientific Habilitation (*Abilitazione Scientifica Nazionale* – ASN 2012) results as criteria.

The analysis presented here will focus on a selection of social and political sciences falling into the so-called Area 14 of the scientific sectors classification elaborated by MIUR (the Italian Ministry of Education, University and Research) upon the advice of the National University Council (CUN) (labelled *CUN areas* for this reason in the academic language). More specifically, Area 14 includes the following scientific disciplines:

- SPS/01 Political philosophy
- SPS/02 History of political thought
- SPS/03 History of political institutions
- SPS/04 Political science
- SPS/05 American history and institutions
- SPS/06 History of international relations
- SPS/07 General sociology
- SPS/08 Sociology of culture and communication
- SPS/09 Economic sociology and sociology of work and organisations
- SPS/10 Urban and environmental sociology
- SPS/11 Political sociology

- SPS/12 Sociology of law, deviance and social change
- SPS/13 African history and institutions
- SPS/14 Asian history and institutions

In administrative language, the disciplines are labelled *Settore concorsuale*, meaning that the recruitment process ("*concorso*", or competition) takes place within the disciplinary boundaries established in the classification. Each of the SCs may include one or more sub-disciplines, labelled *Settore scientifico disciplinare* (SSD). In the rest of the chapter we will deal with the SC aggregation level.

As will be noticed, Area 14 is a heterogeneous set of disciplines, with some disciplines close to the humanities. The VQR final report states that the share of articles among all publications subject to evaluation in the VQR ranges between 22.9% for SPS/06 History of international relations and 37.6% for SPS/04 Political science. The heterogeneity is even greater if we look at the share of publications in English, which ranges from 5.2% for SPS/03 History of political institutions to 38.6% for SPS/04 Political Science (ANVUR 2013).

3 Google Scholar and Other Bibliometric Data Sources

The data source upon which this research focuses is Google Scholar, the application for scientific research that is part of the well-known web search engine (http://scholar.google.com/). Using the *PageRank* algorithm (Franceschet 2011), Google Scholar indexes various kinds of research products (articles, working papers, reports, books, theses) found in the websites of publishers and publishing houses, academic and professional associations, universities, research institutes. For each research product, the algorithm also measures the number of times it is cited in other research products. Scholar is much more extensive than other data sources, but this advantage implies issues of data quality. The data quality significantly increases in the case of personal profiles; in this case, Scholar permits the removal of all those spurious publications or the unification of different records referring to the same publication. According to some authors, the main underlying problem is that "all the information about contributions selection, inclusion criteria, timing of updates and ways of indexing is essentially classified" (Baccini 2010, 75). Scholar can be consulted directly, or some interfaces can be used; among many of these (Scholarometer, Quadsearch, Scirus and others), the best known is Publish or Perish, developed by Anne-Wil Harzing (http://www.harzing.com/pop.htm). This interface permits the generation of many bibliometric indicators and the cleaning of the data obtained from the Scholar dataset. This is an important feature, considering the problems mentioned above.

Besides Scholar, this research focuses on other two data sources. The first one (Scopus) was released in 2004 by Elsevier (www.info.scopus.com) and contains a commercial bibliographical repertoire of abstracts and citations. Compared to Web of Science, Scopus is more extensive and includes more social sciences and

humanities journals and national journals. According to 2014 data, Scopus indexes more than 21,000 journals (compared to the 12,000 of WoS) and 50,000 books (similar to WoS). Scopus and Web of Science share similar problems regarding journal admission (see below); despite being more extensive, the coverage problems for various fields in Scopus are lessened but not solved. Starting from Scopus data, some indicators similar to the Impact Factor are produced; examples include the *SCImago journal rank* (SJR) (www. scimagojr.com) and the *Source normalized impact per paper* (SNIP) (www. journalindicators.com).

Web of Science, managed by the Institute for Scientific Information (ISI) and supplied by Thomson Reuters (http://wokinfo.com), is organised into three different sections, the Science Citation Index Expanded, the Social Sciences Citation Index and the Arts & Humanities Citation Index. It aims at a selective coverage of influential journals as well as books and conference proceedings. As already discussed, WoS coverage for the social sciences and humanities is scarce when compared to that for the natural sciences or medicine (Moed 2005, 42). The admission criteria are demanding, especially for traditional national journals (regularity of publication, acceptance of international editorial conventions, number of citations per journal, English language, procedure of selection). The Journal Citation Reports, linked to WoS, develops some indicators of relevance for scientific journals, including the well-known Impact factor (IF).

4 The Construction of the Dataset and the Coverage of Scholar

4.1 The Population Under Investigation

The data analysed here refer to university professors who fall within the so-called Area 14 (Political and Social Sciences) and who were in service on 31 December 2014. By university professors we mean full professors, associate professors and assistant professors (researchers) with open-ended or temporary contracts. The list has been extracted through the website of the Ministry of Education, University and Research (http://cercauniversita.cineca.it/). Some information was added by drawing on the research data of *The Italian university system: a regional analysis*, a project developed for the Regional Observatory of Education and Training Systems of Regione Puglia with the scientific coordination of Daniele Checchi.

Table 1 shows the distribution of the main characteristics of the population: gender and age group, academic degree and discipline (*Settore concorsuale*), size of the university and area where the university is located. For many of these variables, the information is available even at a more disaggregated level. For the analyses performed here, the level of aggregation has been chosen in order to ensure a sufficient number of cases for group analysis.

Table 1 Frequency distribution of population characteristics (percentages and absolute values)

	%	(N)
Gender		
Male	60.9	(1034)
Female	39.1	(663)
Age group		
1940–1949	13.1	(218)
1950–1959	23.4	(391)
1960–1969	33.7	(562)
1970–1983	29.8	(498)
Academic degree		
Full professor	22.2	(376)
Associate professor	29.9	(507)
Assistant professor (fixed term contract)	38.5	(655)
Assistant professor (temporary contract)	9.4	(159)
Discipline		
14/A1 – Political philosophy	6.2	(105)
14/A2 – Political science	12.9	(218)
14/B1 – History of political thought and institutions	11.1	(189)
14/B2 – History of international relations, non-European history and institutions	8.8	(150)
14/C1 – General sociology	23.7	(401)
14/C2 – Sociology of culture and communication	18.3	(311)
14/C3 – Political sociology, sociology of law, deviance and social change	6.9	(117)
14/D1 – Economic sociology, urban and environmental sociology	12.1	(206)
University size		
Large	55.1	(922)
Medium	25.5	(427)
Small	16.0	(267)
Very small	3.4	(57)
University area		
North-West	23.8	(404)
North-East	19.2	(325)
Centre	24.3	(412)
South	19.7	(335)
Islands	10.4	(177)
Online	2.6	(44)

4.2 The Scholar Dataset

The Scholar dataset was developed within the project *EVA – Extraction, validation and analysis of Google Scholar data in non-bibliometric scientific sectors,* developed by Alfio Ferrara, Stefano Montanelli and Stefano Verzillo, and described in a chapter in this volume.

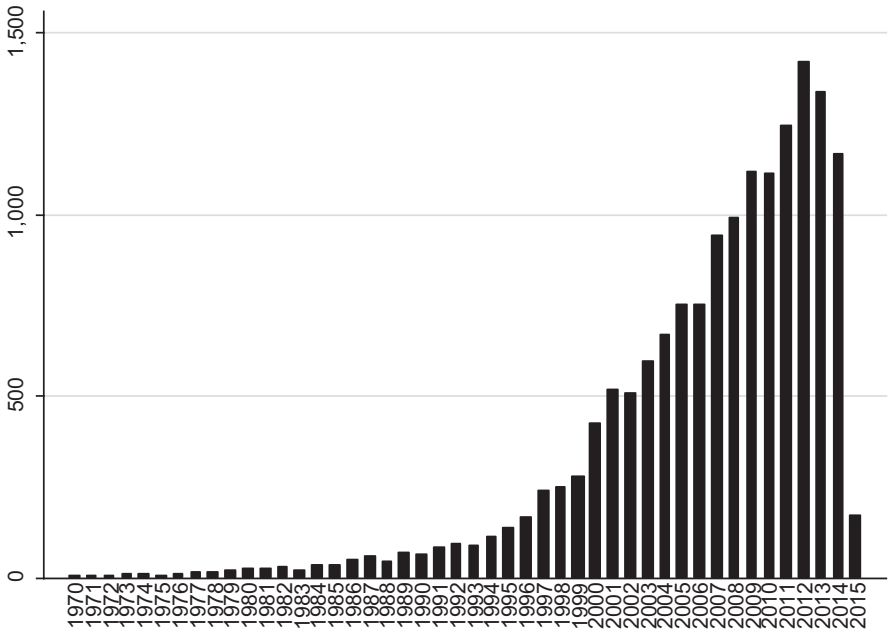


Fig. 1 Publication year of contributions in Scholar (absolute values)

Our work started from a dataset of 17,307 contributions; 1641 of these cases (about 10%) were dropped due to the lack of information on the publication date or because the publication date proved patently implausible (prior to 1970), given the age structure of the population considered here. In this chapter, analyses are carried out mostly on contributions and not on publications, that is, the same publication recurs as many times as there are co-authors who belong to the target population described above.

Figure 1 describes the distribution of the publication years of the contributions. In interpreting this graph, as for the following graphs related to Scopus and Web of Science, keep in mind that the trend of scientific production, as estimated by the three bibliometric datasets, mixes different phenomena. On the one hand, it reflects the coverage capacity of the scientific production, which grows over time, and, on the other hand, the trend of scientific production. In turn, the latter suffers from two contrasting trends, the increase in individual productivity (perhaps following the introduction of evaluation) and the subsequent phases of expansion and contraction of academic staff (from 2000 to 2008, the number of professors rose from 1287 to 1801 and then subsequently dropped to 1679 in 2015). The difficulty of separating the two phenomena obviously has implications for any longitudinal comparison.

Figure 1 shows how the trend of scientific production has increased exponentially since the mid-nineties, so much so that in 2012 it reached a maximum of more

than 1400 contributions. In 2013 and 2014 there is a decrease in the number of publications, but again this has nothing to do with the trend of scientific production but with the delay in some publications being indexed in Scholar.

It is useful to underline how Scholar, compared with Scopus and WoS, is particularly poor in information. For example, neither the type of contribution (monograph, chapter, article etc.) nor the publication language or the name of the journal is immediately available. Information about the name of the journal would allow linking of the articles to the many bibliometric indicators available for journals. In some cases, this information can be retrieved but with some margin of error. In any case, the type of analysis that can be performed on the basis of the Scholar dataset is limited compared to the possibilities offered by the other datasets.

4.3 The Scopus Dataset

While with Scholar data were retrieved in an automatic way using an algorithm, for Scopus (and WoS) data were retrieved in a controlled manner by an expert. This choice implies higher costs in identifying the contributions, which would consume much more human time and much less computer time, but this procedure presumably yields more accurate results. It should be emphasised that the difference between automatic and controlled retrieval in terms of the accuracy of the results must be seen in relative terms; even controlled retrieval is prone to errors (double surnames, composite names, abbreviated forms, problems in the identification of the author's membership). These errors derive from a variety of technical barriers of the selection procedure and from the knowledge of a large and complex field which, even if entrusted to an expert, cannot be complete.

All publication years available in Scopus were considered here. The data collection was carried out between June and July 2015 and was completed for a subset of professors in November of that year. The procedure resulted in the identification of 4126 contributions. Figure 2 shows the trend of publication years in Scopus; this trend is similar to that of Scholar (exponential since the mid-nineties), even if the level is much lower. Here the peak is in 2013 with nearly 600 publications. Again, what is possible to appreciate is more the development of the capacity of the repertoire to include the scientific production of Italian political and social scientists rather than the trend of production.

4.4 The Web of Science (WoS) Dataset

Similarly to what was done for Scopus, in WoS the identification of the contributions of Italian scholars was not carried out automatically by an algorithm but in a controlled manner by an expert.

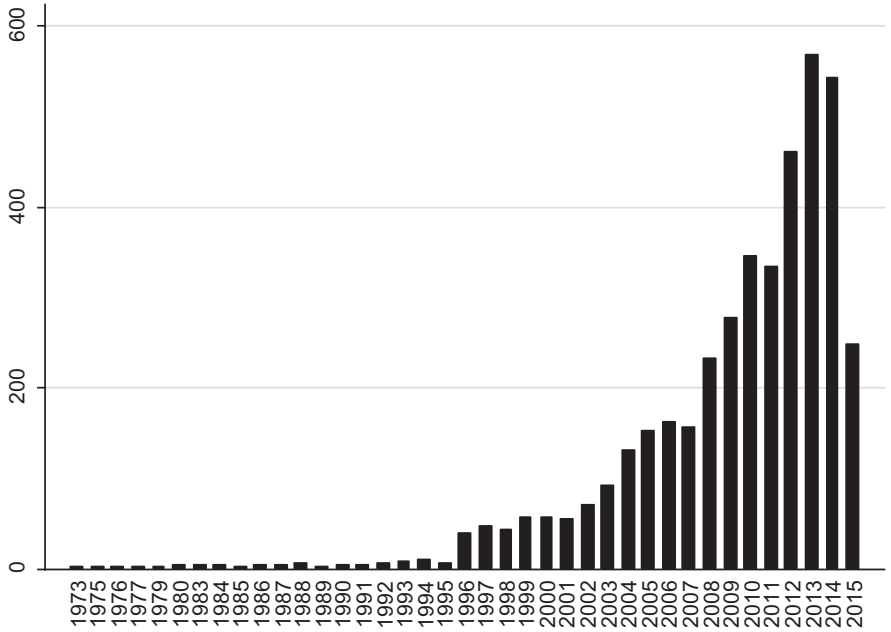


Fig. 2 Publication year of contributions in Scopus (absolute values)

We considered all available years and all bibliographic repertoire indexes of the Web of Science Core Collection: Science Citation Index Expanded (SCI-EXPANDED) from1985 to the present; the Social Sciences Citation Index (SSCI) from 1985 to the present; the Arts & Humanities Citation Index (A&HCI) from 1985 to the present; the Conference Proceedings Citation Index – Science (CPCI-S) from 1990 to the present; the Conference Proceedings Citation Index – Social Science & Humanities (CPCI-SSH) from 1990 to the present; the Emerging Sources Citation Index (ESCI) from 2015 to the present) together with the Book Citation Index (from 2005 to the present). Data collection from the Web of Science Core Collection was carried out in December 2015 and from the Book Citation Index in February 2016.

The WoS dataset consists of 2941 contributions. Figure 3 shows the publication year trend, which is similar to the one already seen for Scholar and Scopus but with a period of stability between 2009 and 2012. The level here is even lower; the maximum is reached in 2014 with little more than 300 contributions.

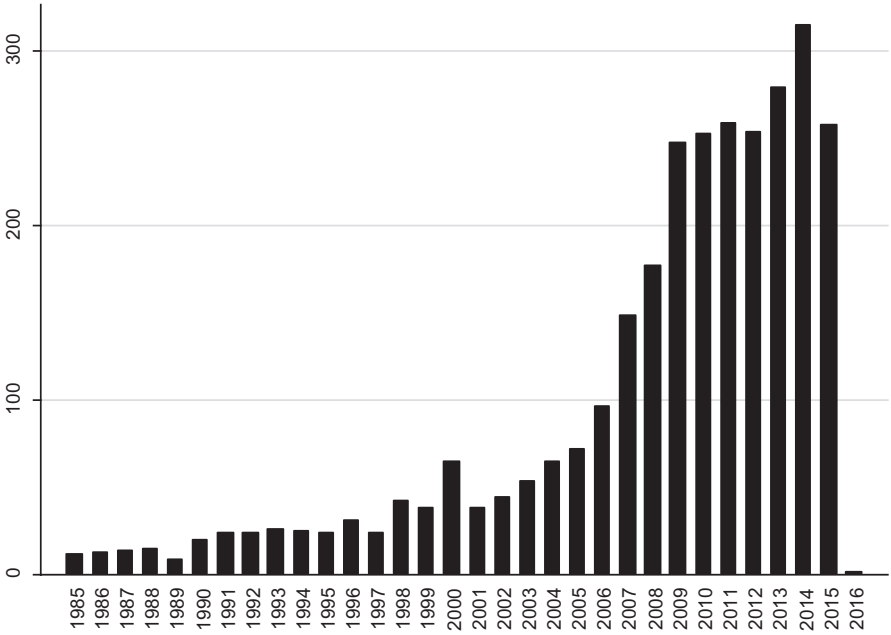


Fig. 3 Publication year of contributions in WoS (absolute values)

4.5 The SUA-RD 2011–2013 Dataset and the Coverage Algorithm

In addition to the citation databases, the SUA-RD dataset was considered. The dataset considers the publication activities of Area 14 inserted within the SUA-RD procedure in the 2011–2013 period and it was made available by ANVUR.[2] The SUA-RD data is used to analyse the Scholar coverage.

The four datasets considered – Scholar, Scopus, WoS and SUA-RD 2011–2013 – were collected by different sources. In order to be able to perform a comparison between the publications they included, it was necessary to process some matching algorithms. This need was due to the possibility – not so remote – that the same record was indexed differently in different databases (i.e. with slight differences in the title or even with different years of reference). The heuristic configuration adopted in the following project was particularly selective, to search with a high degree of reliability, even at the cost of discarding valid correspondences. In detail, for two records from two different databases to be considered as a match, the following conditions must be jointly verified:

[2] We would like to thank Andrea Bonaccorsi and Marco Malgarini, who allowed us to use SUA-RD 2011–2013 data.

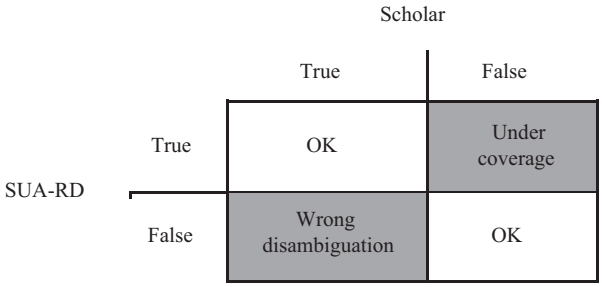


Fig. 4 Trade-off between correct disambiguation and proper coverage

- The publication year of the two must not differ by more than a year.
- The edit distance (minimum number of operations required to make two strings equal) between the two titles must be less than 2.

Before proceeding, it is necessary to make a clarification. As just mentioned, the data used for the following analysis were not the raw output of Scholar but the data disambiguated within the EVA project. It should be emphasised that there is a trade-off between correct disambiguation and proper coverage (see Fig. 4). If you want to increase correct disambiguation, this inevitably increases the number of products that really are in SUA but that are not identified in Scholar – the under coverage. If you want to increase the coverage, you must be willing to tolerate a higher disambiguation error, that is, increase the number of instances that are not in SUA but that match with a record in Scholar.

The parameters and thresholds have to be properly calibrated to favour one or the other objective. The following analyses, then, are conditioned by the choices that were made in the EVA project, and changes in these calibration parameters may have led to slightly different results.

Figure 5 shows the consistency of the SUA-RD publications in the period 2011–2013 and compares this with the raw output of Scholar, with the cleaned output of Scholar disambiguated by author (EVA1), with the cleaned output of Scholar disambiguated by author and sector (EVA2) and with Scopus and WoS. What is striking is certainly the different order of magnitude of the publications in the SUA-RD, not only compared to WoS and Scopus but also compared to EVA. SUA-RD collected 5728 contributions in 2011, 6112 in 2012 and 5593 in 2013; EVA2 never exceeds 1500 contributions, which is still fewer than those of Scopus and WoS. Even if the observation window is very narrow, it should be noted that for the SUA-RD there are no signs of the exponential curve that characterises the citation databases. This confirms that a similar trend does not capture the dynamics of scientific production but rather the growing ability of citation databases to cover such production.

SUA-RD is used as a benchmark. Therefore, it is important to evaluate how it covers Scopus and WoS in order to understand the appropriateness of this database.

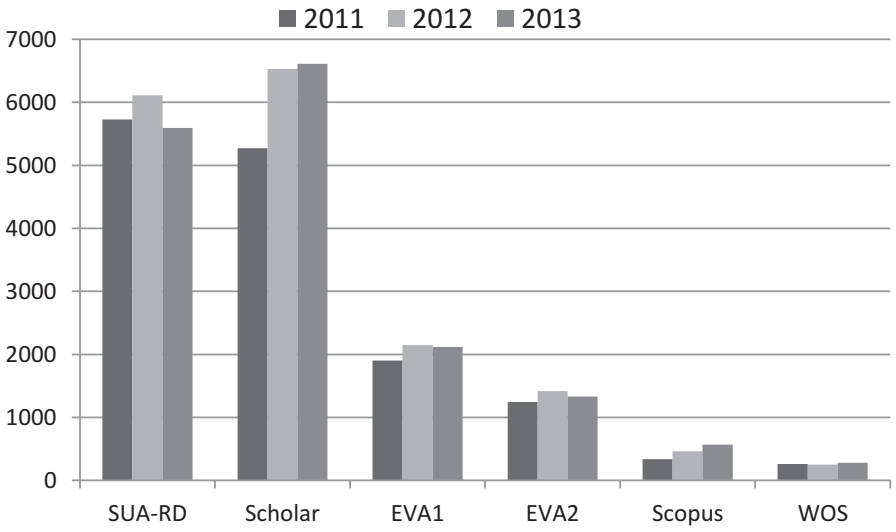


Fig. 5 Year of publication of contributions in SUA-RD, Scholar, EVA, Scopus and WoS (absolute values)

Only 61.6% (total 1356) of Scopus records are covered by SUA-RD, and only 60.7% (total 780) of WoS records are covered by SUA-RD. These percentages respectively become 64.7% and 74.9% if only the type of article is considered, and they are obviously lower, at 46.1% and 26.2%, respectively, if other types of record are considered which are different from article, chapter and book chapter.

This result highlights the fact that there are inconsistencies in how the data are stored in citation databases and in the SUA-RD: sometimes the titles are in Italian and sometimes in English; the years do not coincide; in one of the two sources, a subtitle is present; there are duplicates.

Note that this result refers to the period 2011–2013. In recent years, the institutional archives of universities have standardised and implemented mechanisms able to connect the publication to the corresponding records in the citation databases when they are present. If the exercise were to be repeated now, the result would surely be very different. Moreover, in the period considered only a few institutional archives of universities were Open Access. Recently, all institutional research archives have become Open Access. This has introduced transparency, reproducibility, accountability and visibility. In future exercises, the inconsistencies will probably be drastically reduced even if the problem is not completely solved.

Another issue is the types of publications represented by the databases. The distribution of product types, as seen from Fig. 6, is very different in the three databases. Scopus and WoS primarily include articles, while the modal type for SUA-RD is instead the book chapter. Books comprise more than 15.2% of the production of scholars in Area 14, according to SUA-RD, but less than 3.2% in Scopus and WoS. Many books could be in Italian, and in fact, the percentages of book chapters

Fig. 6 Distribution of type of publication by database

Table 2 Coverage of SUA-RD by three sources: Scholar, Scopus and WoS

	Scholar		Scopus		WoS	
SUA	N	%	N	%	N	%
No	14,925	85.2	16,772	95.3	17,149	97.1
Yes	2588	14.8	836	4.7	513	2.9
Total	17,513	100	17,608	100	17,662	100

are also very different, that is, 38.5% versus 16.6% for Scopus and 21.2% for WoS. It is important to note also that the Other category includes records that are very different: conference proceedings, translations, reviews, forewords and afterwords in SUA-RD and discussions and rejoints in Scopus and WoS.

After discussing the reliability of the database SUA-RD used as a benchmark, we shall now proceed with the analysis of the coverage of the three data sources Scholar, Scopus and WoS with respect to this benchmark database. As highlighted in Table 2, the Scholar coverage is much higher than that of Scopus and WoS, although the 14.8% coverage does not lead to the conclusion that Scholar is currently able to cover the scientific production of the scholars of Area 14.

We will now assess whether the coverage of Scholar differs by type of publication. We also consider the ASN classification for the publication, where present. The coverage may also vary according to the characteristics of the authors or universities the authors belong to. The same analyses were also carried out for Scopus and WoS, but they are not reported, since the percentage of coverage differs little.

A first difference, as seen in Table 3, can be observed in the type of publication. Articles and books are covered by Scholar with respective percentages of 27.7% and 17.6%.

Table 3 Coverage of SUA-RD in Scholar by the type of publication

Type		No	Yes	Total
Article	N	3540	1356	4896
	%	72.3	27.7	
Book	N	2187	468	2655
	%	82.4	17.6	
Book chapter	N	6151	566	6717
	%	91.6	8.4	
Other	N	3099	194	3203
	%	93.9	6.1	
Total	N	14,887	2584	17,471
	%	85.2	14.8	

The coverage changes according to the level of the publications. In particular, 33% of the publications considered as Class A according to the ASN classification are covered by Scholar. Following the same ASN classification, 27% of scientific journals considered are covered by Scholar. Even the presence of a foreign co-author appears to increase the level of coverage: 27% of publications with a foreign co-author are covered by Scholar.

Regarding the characteristics of the authors, a different coverage does not seem to emerge, depending on the author's qualifications, while the disciplinary sector the publication belongs to seems to affect the SUA-RD coverage by Scholar. The sectors which are more covered are A2 (Political Science), C2 (Sociology of cultural and communication) and D1 (Economic sociology, urban and environmental sociology), with a coverage of 19%, 19% and 18%, respectively. The sectors which are less covered are A1 (Political Philosophy) at 10% and B1 (History of political thought and institutions) at 6%.

We now consider the characteristics of the university to which the authors belong. No difference is shown as a result of the size of the university. But the coverage is much higher for universities of the North-West (23%) and the North-East (18%) than for those in the Centre (10%) or the South and the Islands (8% for both).

As part of a further analysis, we wondered whether the ranking of universities or sectors greatly varied, considering Scholar instead of SUA-RD.

The following scatter plot shows the ranks of the universities (Fig. 7) and the sectors (Fig. 8) according to their rank on SUA-RD (horizontal axis) and their rank on Scholar (vertical axis). The plot shows that the sorting does not change substantially, even if, in the case of the universities, a large number change their ranking by 10 positions and five universities even change their ranking by 20 positions. The Spearman correlation indices are 0.77 for the universities and 0.89 for the sectors.

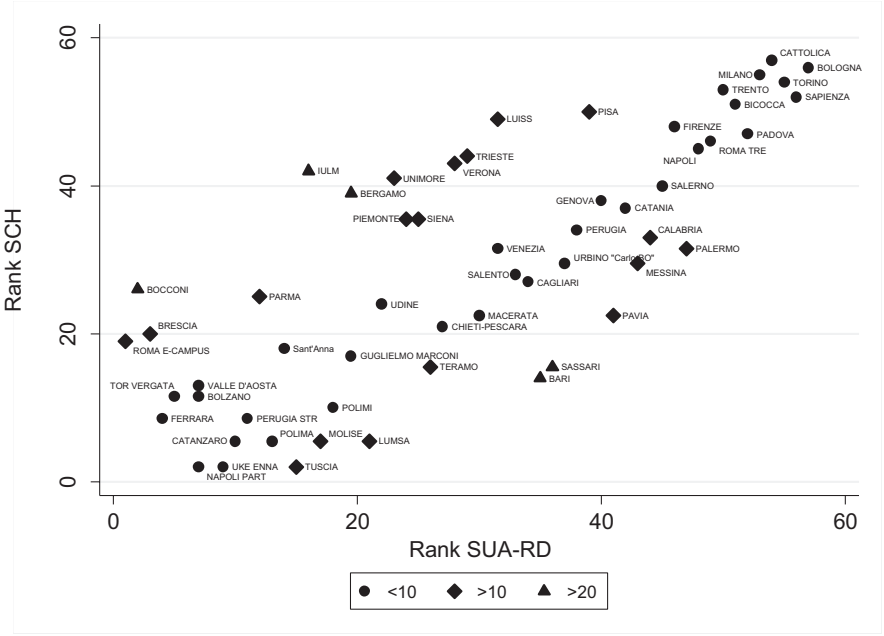


Fig. 7 Ranks of universities for SUA-RD and Scholar

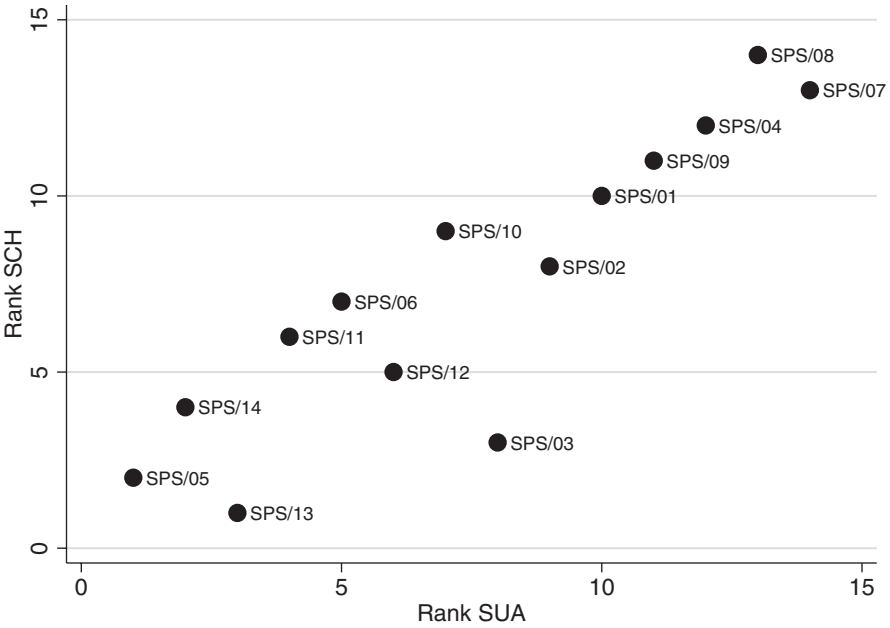


Fig. 8 Ranks of sectors for SUA-RD and Scholar

5 Reliability and Validity of Bibliometric Indicators

5.1 Descriptive Statistics of Output and Recognition Indicators

Starting from the large amount of data gathered, we decided to focus on three kinds of indicators.

- Number of contributions in Scholar, Scopus and WoS.
- Average number of citations per contribution in Scholar, Scopus and WoS.
- H-index in Scholar, Scopus and WoS.

The first indicator refers to scientific output, whereas the other two refer to scientific recognition. The average number of citations indicator was preferred to a simple arithmetical sum in order to normalise these values for the scientific output of the authors. Another option could have been to normalise for the activity period (computed as the difference between the publication years of the first and last contributions), but the number of contributions better grasped the volume of scientific output. Also, a preliminary analysis of distribution asymmetry confirmed the choice to normalise according to the number of contributions.

We also decided not to take into account other indicators related to the kind of publication or linked to the journal they were published in (Class A, IF, SCImago, SNIP etc.) because of the difficulty for Scholar in accurately identifying the kind of publication and the journal for the scientific articles. For further clarification, see the chapter by Alfio Ferrara, Stefano Montanelli and Stefano Verzillo in this volume.

All the indicators refer to the period 1998–2014, since all three data sources report reliable numbers for each of the years considered.

Table 4 presents the features of the statistical distribution for the nine indicators considered here. The results are not surprising. Regarding the number of contributions and the average citations per contribution, Scopus and WoS show similar

Table 4 Descriptive statistics of bibliometric indicators

	Median	Mean	Max	Std. dev.	Variation coefficient	Asymmetry	Kurtosis	(N)
Number of contributions								
Scholar	4	8.2	147	11.7	1.4	3.7	26.3	(1697)
Scopus	1	2.1	41	3.9	1.8	3.3	19.0	(1697)
WoS	0	1.4	53	3.1	2.1	5.5	60.8	(1697)
Number of citations per contribution								
Scholar	2.0	5.5	216.6	13.5	2.4	8.5	102.5	(1372)
Scopus	0.6	2.5	72.6	5.8	2.2	5.9	52.4	(924)
WoS	1.0	2.5	87.0	6.0	2.3	7.0	74.9	(692)
H-index								
Scholar	2	2.6	22	2.9	1.1	2.3	10.1	(1372)
Scopus	1	1.1	12	1.5	1.3	2.4	11.3	(924)
WoS	1	1.0	10	1.2	1.2	2.0	9.3	(692)

Table 5 Correlation coefficients for the indicators of output and recognition in Scholar, Scopus and WoS (Spearman correlations for the original versions) and Cronbach's Alpha

	Contributions	Citations per contributions	H-Index
Scholar-Scopus	0.48	0.64	0.65
Scholar-WoS	0.42	0.59	0.57
Scopus-WoS	0.65	0.74	0.72
Cronbach's Alpha	0.77	0.76	0.86
(N)	(1697)	(534)	(534)

values and they are very different from Scholar; the median of the distribution is significantly higher for Scholar, not to mention the range (here it coincides with the maximum value of the distribution) and the standard deviation. The situation changes slightly if we consider the variability in relative terms by means of the coefficient of variation, which shows similar distributions. The values for asymmetry and kurtosis are very high; to deal with them, we will use different techniques and different versions of the original variables through the entire analysis. Worthy of note is the median value for the number of citations in WoS (0), meaning that at least half of the authors from Area 14 have no publications at all in WoS.

The values for asymmetry and kurtosis are indeed lower for the H-index for all three data sources; this peculiarity makes the H-index very desirable from a statistical point of view if compared to the other citation indicators. The H-index is defined by its creator in this way: "A scientist has index h if h of his or her N_p papers have at least h citations each and the other $(N_p - h)$ papers have $\leq h$ citations each" (Hirsch 2005, 16,569). For instance, an H-index equal to 10 means that a researcher has 10 publications which are quoted at least 10 times. It is evident that this indicator aims to consider both scientific output and recognition.

5.2 Reliability of Output and Recognition Indicators

As mentioned above, the reliability of the indicators of scientific output and recognition is assessed here by looking at the reproducibility of the results and is measured on the basis of a correlational analysis. Table 5 presents the main results of this analysis. First of all, all correlation coefficients are rather high, which means a general reliability for the three indicators in the three data sources considered. Secondly, for all three indicators, the correlation coefficients for Scopus and WoS are higher than the correlation coefficients for Scholar. Regarding this latter, its correlation with Scopus is always higher than the correlation with WoS.

This correlation structure seems to capture the specificities of the three data sources. If we hypothesise a continuum ranging from international journal articles to different types of publications even at the national level, we find a clear order of the three data sources on this continuum (WoS, Scopus, Scholar), with Scopus being much closer to WoS than to Scholar.

Moreover, it should be noted that the correlation coefficients between the citation indicators are higher than the correlations between the indicators counting the number of contributions. This might mean that the structure of scientific recognition is more coherent if compared to the structure of scientific output. Put another way, the various specificities of the three data sources are more evident for the indicators concerning scientific output, whereas the dimension of scientific recognition is captured in a more stable way from all three sources.

To take into account the distribution asymmetry for the indicators of contributions and average citations per contribution, in addition to the rank correlations (Spearman) for the original variables, the standard correlation coefficients (Pearson) were also calculated on alternative versions of such indicators. In the truncated version, a threshold value – equal to three interquartile differences over the third quartile – was defined, and the cases with higher values were restricted to this threshold. The version using deciles recoded the cases according to the related decile, whereas the dichotomous version distinguished the values equal to "0" from all the other values coded as "1". The square root was computed for the average number of citations per contribution, as suggested by the analysis based on the "ladder of powers" to reduce the distributions' asymmetry (Hamilton 2012, 129–132). All the correlation coefficients between these different versions showed stable values, with the exception of the dichotomous one, which presented lower correlations, probably due to the loss of information. All these procedures were not necessary for the H-index, which presented a more symmetrical distribution. Table 5 shows the Spearman correlations similarly to what was done for the other indicators.

As further proof of the general reliability of the indicators considered, Table 5 shows the values of the Cronbach's Alpha computed for each indicator for the three data sources. In every case, the values are higher than the usual threshold of acceptability (0.7).

5.3 Validity of Indicators of Scientific Recognition: VQR 2004–2010

In this section, we will deal with the issue of the validity of indicators of scientific recognition. As previously stated, we will apply the procedure of criterion validation which evaluates the indicator's validity on the basis of associations with an external criterion linked to the measured property. This evaluation is based on the assumption that the citation variables are indicators of scientific recognition, a property which is presumed to be linked with the quality of the scientific product.

This analysis comprises two different parts. In this section, we will use the average scores obtained in the Research Quality Evaluation (*Valutazione della Qualità della Ricerca* – VQR 2004–2010) as external criteria, whereas in the next section we will use the scores from the National Scientific Habilitation (*Abilitazione Scientifica Nazionale* – ASN 2012). Both these procedures are managed by ANVUR

and are based on *peer review* processes aiming to ascertain the scientific quality of the research. It should be noted that in the first case only the scientific production is considered, whereas in the second other aspects of scientific work – research projects, teaching activities, research activity in international institutes, awards etc. – are also taken into account.

For what concerns the VQR 2004–2010, the researchers were asked to choose three scientific products. The variable considered here is the average score for these three products. One of the following scores was assigned to each product: "1" for "Excellent", "0.8" for "Good", "0.5" for "Acceptable" and "0" for "Limited". In the case of missing products, the score was "-0.5", in cases of products that were impossible to evaluate, it was "-1" and it was "-2" for verified situations of scam or plagiarism. Concerning Area 14, results from the VQR show that 36.0% of the products scored "Excellent-Good", 30.6% scored "Acceptable" and 27.3% "Limited". The remaining 4.0% consisted of penalised products, including missing products, that is, products not delivered (167 products) as well as explicitly penalised products, including non-eligible products, plagiarism, self-plagiarism or scams (13 products). All this information was taken from the final report of the Research Quality Evaluation (VQR 2004–2010) (ANVUR 2013).

In this case, we considered the H-Index and the number of citations per product as indicators of scientific recognition. For the latter, given its asymmetry problems, we used the rank correlation coefficient (Spearman's Rho), whereas for the H-Index we used the usual Pearson's correlation coefficient. For both indicators, we considered two different versions, the original one covering the whole observation window (1998–2014) and a second version covering only the period considered by the VQR (2004–2010).[3]

All the correlation coefficients are reported in Table 6. Starting from the H-Index computed for the extended time span (1998–2014), the table shows that all the correlation coefficients are statistically significant, ranging from 0.33 (Scholar-VQR) to 0.29 (Scopus-VQR) and 0.26 (WoS-VQR). Using the VQR score as the external criterion, the Scholar H-Index seems to be the most valid, followed by Scopus and WoS. If we use the 2004–2010 time span – the same period considered by the VQR – the correlation coefficients as well as their order are basically the same. Even if we had computed the Spearman's Rho, either the values would remain the same or the differences would increase for the 2004–2010 observation window.

If we look at the correlation coefficients between the VQR scores and the other indicator of scientific recognition – the average number of citations per contribution – we see lower values. These coefficients still remain statistically significant. In addition, the range stating a higher validity for Scholar (0.27) and a lower for

[3] Given that the VQR scores are highly confidential, these analyses were conducted directly by ANVUR. In this regard, we thank Marco Malgarini, Tindaro Cicero and Marco De Santis Puzzonia from ANVUR for their precious collaboration. We also need to report that given some difficulties in combining the CAVIB dataset with that of the VQR 2004–2010, the analysis was conducted for 1394 over 1697 cases (82% of the total sample).

Table 6 Correlation coefficients between the average score on VQR 2004–2010 and the original versions of the indicators of scientific recognition for Scholar, Scopus and WoS for the two different time spans

	H-Index/VQR			Citations per contribution/VQR		
	Pearson Correlation			*Spearman Correlation*		
1998–2014	*r*	*p*	*(N)*	*rho*	*p*	*(N)*
Scholar	0.33	0.000	(1127)	0.27	0.000	(1127)
Scopus	0.29	0.000	(750)	0.23	0.000	(750)
WoS	0.26	0.000	(559)	0.21	0.000	(559)
2004–2010	*r*	*p*	*(N)*	*rho*	*p*	*(N)*
Scholar	0.33	0.000	(970)	0.27	0.000	(970)
Scopus	0.30	0.000	(463)	0.26	0.000	(463)
WoS	0.25	0.000	(374)	0.19	0.000	(374)

Scopus (0.23) and WoS (0.21) is confirmed.[4] The results are highly consistent even if we look at the 2004–2010 time span.

5.4 Validity of Indicators of Scientific Recognition: ASN 2012

The analysis of the indicators of scientific recognition based on the ASN 2012 uses a different version of the same criterion validation procedure. In the case of VQR 2004–2010, we used a *concomitant validation* process based on the assumption that the quality of the publications evaluated with a peer review process in the VQR 2004–2010 and the value of scientific recognition measured by the citation indicators were two contemporary features. With this being a symmetrical association the use of correlation coefficients is justified.

In the case of ASN 2012, the situation is different. Here, the underlying question is whether the use of citation indicators can predict the outcome of the habilitation procedure. This kind of procedure is called *predictive validation*. In this case, the association is asymmetrical. We assumed that the citation indicators impacted the habilitation, and therefore, we ran a logistic regression with the result of the habilitation procedure (yes/no) as dependent variable and the continuous bibliometric indicators as independent variables.

Although ASN considered other aspects of scientific work beyond the scientific quality of publication (research projects, teaching or research activities in international institutes, awards etc.), the results according to the two criteria (VQR and ASN) are essentially coherent.

Table 7 shows the results for the logistic regression model for the two citation indicators (average citations per contribution and H-Index) and for the three

[4] If instead of the Spearman correlation coefficients we had calculated the Pearson's coefficient, the results would have been even lower.

Table 7 Logistic regressions for habilitation with different normalised bibliometric indicators

	OR	Std. Err.	z	P > |z|	Pseudo R^2	Count R^2	Adjusted Count R^2	(N)
H-index (normalised)								
Scholar	1.58	0.119	6.13	0.000	0.05	0.60	0.19	(696)
Scopus	1.60	0.156	4.83	0.000	0.04	0.57	−0.02	(509)
WoS	1.39	0.143	3.23	0.001	0.02	0.64	0.00	(397)
Average citations per contribution (normalised square root)								
Scholar	1.34	0.112	3.57	0.000	0.01	0.55	0.10	(696)
Scopus	1.22	0.088	2.88	0.004	0.01	0.58	0.00	(509)
WoS	1.22	0.111	2.20	0.028	0.01	0.64	0.00	(397)
ASN indicator (normalised)								
1	0.98	0.071	−0.39	0.695	0.00	0.48	−0.04	(696)
2	2.09	0.235	6.61	0.000	0.04	0.61	0.21	(509)
3	1.93	0.163	7.84	0.000	0.07	0.64	0.28	(397)

H-Indexes, average citations per contribution in Scholar, Scopus and WoS and ASN 2012 indicators

bibliometric indicators computed within ASN. These three indicators, hereby used as *benchmarks*, are the following:

1. Number of books
2. Number of journal articles or book chapters
3. Number of articles published in "Class A" journals

In order to allow a comparison between the various effects (odds ratio and predicted probabilities), we also used the base 10 normalised version of the indicators. This means that all the variables were scaled to make the range of variation uniform (from 0 to 10) (Corbetta et al. 2001, 80–82). For the abovementioned reasons, the normalisation of the variables concerning the average number of citations per contribution was computed starting from the square root of the original variables.

Table 7 shows some features of the nine models. Given the predictive validation perspective, the more relevant results regard the fit of the models rather than the size of the effects. The table displays three different indexes of model fit. The Pseudo (McFadden) R^2 is based on the maximum likelihood function, whereas the two different versions of Count R^2 define the model fit as a predictive power, which corresponds exactly to what we want to evaluate. The Count R^2 corresponds to the proportion of cases rightly predicted by the model, i.e. to the proportion of qualified researchers and non-qualified researchers rightly predicted by the model. However, this index does not consider the fact that even without the model it is possible to predict the dependent variable – the habilitation – using its mean, the proportion of habilitated researchers: if the value of the proportion of habilitated researchers exceeds 0.50 we can predict a successful habilitation for all the candidates, if the proportion is lower than 0.50 we can predict failure for all the candidates. Conversely, the Adjusted Count R^2 also considers the so-called Model 0, based on the mean of the dependent variable, and returns the proportion of rightly predicted cases among

those not explicitly predicted by Model 0 (Menard 2002, 27–36). Put another way, beyond the cases rightly predicted by Model 0 (based on the mean of the dependent variable), how many cases can Model 1 – the one with the bibliometric indicator – predict?

The results shown in Table 7 are rather clear. The Scholar H-Index permits the prediction of 19% of cases, more than Model 0. Regarding the average citations per contribution indicator in Scholar, the Adjusted Count R^2 is instead 0.10. These values are lower than the two best performing indicators from ASN (the number of articles published in "Class A" journals and the number of published articles or book chapters), but they are decisively higher than the bibliometric indicators from Scopus and WoS, which are unable to increase the predictive ability of Model 0.

6 Conclusions

The set of results which we have discussed suggests to keep distinct the bibliometric level from the evaluative one. It must be stressed that bibliometrics and evaluation are not the same thing. Even if the connection between evaluation and bibliometrics is perceived by many scholars as very close, the two disciplines are relatively autonomous. Bibliometrics emerged from science studies and can be seen as a *tool* of research evaluation.

At the bibliometric level, the main results are as follows. First, the correlational analysis of the different indicators confirms the reliability of bibliometric indicators. The specificities of the different repertoires are also confirmed, with WoS and Scopus correlating highly with each other and Scholar correlating more with WoS than with Scopus. To assess the validity of the scientific recognition indicators, we used criterion validation. In particular, we used two criteria: the individual scores of VQR 2004–2010 and the results of ASN 2012. Both criteria confirm the validity of the scientific recognition indicators based on all three data sources, although Scholar is systematically more valid than the others.

Combining the two sets of results, a sort of scale (Scholar/Scopus/WoS) emerges, with the reliability increasing from Scholar to WoS while the validity instead increases from WoS to Scholar. Such a pattern is likely to be associated with the scientific perimeter covered by the different bibliometric datasets. While WoS and Scopus preferably cover articles in international journals, Scholar is more open to other types of publications and to national literatures. These different scientific perimeters become an advantage in terms of reliability for WoS and Scopus, which are more similar to each other, and an advantage in terms of validity for Scholar, which better discriminates among those researchers who do not publish articles in international journals.

Then there are some side results. The scientific recognition indicators are more reliable than output indicators. Furthermore, according to all the results – and not only in the validation analysis – the H-index seems to work better than the average

number of citations for contribution. The bibliometric longitudinal comparisons are biased by a coverage capacity of bibliometric datasets that is growing over time.

Finally, we must not forget to mention a relevant weakness of Scholar that might go unnoticed. The difficulties in identifying for each contribution the type of publication (book, chapter, article etc.) and the journal of publication in the case of scientific articles drastically reduces the number of potential bibliometric indicators derivable from Scholar. It is therefore problematic to develop an analysis of the different types of publication and to define scientifically relevant indicators (IF, SCImago and the like) that are based on the characteristics of the publication venues.

With regard to evaluation, the main issues are those of Scholar coverage and data quality at the contribution level. Using as a benchmark SUA-RD data referring to the years 2011–2013, Scholar covers about 15% of the scientific production compared with 5% of Scopus and 3% of WoS. What is most striking is the Scholar absolute value. Although it certainly has a large degree of underestimation, the analysis shows that Scholar still processes a significant selection of the work produced by the scientific community. It also shows how Scholar is certainly more comprehensive than commercial repositories while far from being exhaustive. This is a first result which shortens the distance between Scholar, Scopus and WoS.

Second, it has been shown that the data collection process for the various sources – in particular the correct attribution of the contributions to the authors – is hampered by similar problems. It was also found that the quality of the data is certainly higher in commercial repertoires than in Scholar, but it is still not absolute. The three databases seem to be more similar than one might think. In other words, for all three sources, you have a trade-off between coverage and disambiguation, which, however, is magnified in the case of Scholar for the lower quality of the dataset.

If the validity and reliability of bibliometric indicators drawn from Scholar prove satisfactory at the aggregate level, there remains the problem of data quality at the level of single publications. Data quality is a critical point for any evaluation system, and the tolerable error threshold varies depending on the object of evaluation. If you evaluate a scientific structure (university, department, research lab etc.), you can appeal to the requirement that a mistake should not be such as to distort the results of the assessment. If you evaluate an individual, even this criterion crumbles because the error will induce a deep crisis of confidence in the system. The result is thus a suggestion for caution in the use of Scholar data, especially with respect to individual assessments.

Finally, it should be noted that these considerations are made in a changing environment, where great efforts have been undertaken to solve the thorny problem of correct attribution of publications to the authors. One example is the ORCID system adopted by ANVUR for the VQR 2011-2014. Commercial repertoires are also developing systems to improve disambiguation. It is therefore plausible that in just a short time the landscape may change dramatically by simplifying the combination of bibliometrics and evaluation. Another phenomenon to consider is the growing diffusion of Open Access in the Italian academic system. This will favour the presence on the web (and therefore on Scholar) of Italian scientific production.

References

ANVUR. (30 June2013). *Valutazione della Qualità della Ricerca 2004-2010 (VQR 2004–2010). Rapporto finale di area. Gruppo di Esperti della Valutazione dell'Area 14 (GEV14)*, Rome.

Baccini, A. (2010). *Valutare la ricerca scientifica. Uso e abuso degli indicatori bibliometrici.* Bologna: Il Mulino.

Clemens, E. S., Powell, W. W., McIlwaine, K., & Okamoto, D. (1995). Careers in print: Books, journals, and scholarly reputations. *American Journal of Sociology, 101*(2), 433–494.

Cole, S., & Cole, J. R. (1967). Scientific output and recognition: A study in the operation of the reward system in science. *American Sociological Review, 32*(3), 377–390.

Corbetta, P. (2003). *Social research. Theory, methods and techniques*. Thousand Oaks: Sage.

Corbetta P., Gasperoni G. and Pisati M. (2001). *Statistica per la ricerca sociale*, il Mulino, Bologna.

Franceschet, M. (2011). PageRank: Standing on the shoulders of giants. *Communications of the ACM, 54*(6), 92–101.

Hamilton, L. C. (2012). *Statistics with Stata: Updated for version 12*. Belmont: Brooks/Cole, Cengage Learning.

Hicks, D. (2004). The four literatures of social science. In H. F. Moed, W. Glaenzel, & U. Schmoch (Eds.), *Handbook of quantitative science and technology research* (pp. 473–496). Dordrecht: Kluwer Academic Publishers.

Hirsch, J.E. (15 Nov 2005). An index to quantify an individual's scientific research output, in *PNAS. Proceedings of the National Academy of Sciences of the United States of America*, *102*(46).

Martin, B. R., & Irvine, J. (1983). Assessing basic research. Some partial indicators of scientific progress in radio astronomy. In *Research policy*, vol. 12, (pp. 61–90); partial Italian translation in Viale, R. and Cerroni, A. (2003) (eds.) *Valutare la scienza.* Soveria Mannelli: Rubettino.

Menard, S. (2002). *Applied logistic regression analysis* (2nd ed.). Thousand Oaks: Sage.

Merton, R.K. (1973). *The sociology of science.* Chicago: The University of Chicago Press; Italian translation: *Scienza, religione e politica*, Bologna, Il Mulino, 2011.

Moed, H. F. (2005). *Citation analysis in research evaluation*. Dordrecht: Springer.

Nederhof, A. J. (2006). Bibliometric monitoring of research performance in the social sciences and the humanities: A review. *Scientometrics, 66*(1), 81–100.

Research Evaluation and Policy Project. (2005). *Quantitative indicators for research assessment – A literature review*, Discussion Paper 05/1, 30 March.

Is the Diffusion of Books in Library Holdings a Reliable Indicator in Research Assessment?

Maria Teresa Biagetti, Antonella Iacono, and Antonella Trombone

1 Analysing the Diffusion of Books in Libraries As Evaluation Indicator

This paper addresses the complex issues of using the diffusion in Italian and foreign library holdings of books published by Italian scholars as a complementary indicator in SSH research assessment. An analogy was established years ago between the diffusion of books in libraries and the citations received by articles published in scientific journals.

Henk F. Moed (2005) pointed out the problems that arise with the use of bibliometric counts based on citations, obtained through tools like Web of Science (Thomson Reuters)[1] or Elsevier's Scopus.[2] First, the structure of the indexes for citation analysis and the coverage of scientific journals for each discipline are critical to allow a good citation analysis: according to Moed, the adequacy of

Maria Teresa Biagetti is the author of Part 1.
Antonella Iacono is the author of Part 2.
Antonella Trombone is the author of Part 3.

[1] http://wokinfo.com/products_tools/multidisciplinary/webofscience/. Presently, the *Web of Science* platform is managed by *Clarivate Analytics*.

[2] http://www.elsevier.com/online-tools/scopus

M.T. Biagetti (✉)
Sapienza University of Rome, Rome, Italy
e-mail: mariateresa.biagetti@uniroma1.it

A. Iacono
Biella Public Library, Biella, Italy
e-mail: antonella.iacono@fastwebnet.it

A. Trombone
University of Basilicata, Central University Library, Potenza, Italy
e-mail: antonella.trombone@gmail.com

© Springer International Publishing AG 2018

A. Bonaccorsi (ed.), *The Evaluation of Research in Social Sciences and Humanities*, https://doi.org/10.1007/978-3-319-68554-0_14

ISI-Thomson coverage is excellent in Biology, Chemistry, Physics etc., and moderate in Humanities and Social Sciences. Analysing the use of citations, significant differences between SSH authors and authors in STEM clearly appear. In the former case the use of the bibliographic citation is more bound to specific schools of thought than in the latter. Another difference lies in the tendency of SSH scholars to publish their research in monographs, as they are somewhat more reluctant to publish preliminary results of their research in scientific journals.

Starting from 2011, Web of Science Core Collection included in WoS indexes the Book Citation Index (BKCI) that allows the search of books and book chapters.[3] BKCI covers only scholarly books that present original research, or reviews of literature, handbooks with full bibliography, books in series, dissertations published by a reputable scholarly publisher, translations of non-English contents. To include a book in BKCI, the priority is given to books published in series that received high numbers of citations. Thomson captures book citations in scientific journal articles and proceedings, identifies those books that have greater impact and put them in the BKCI. As of 2012, BKCI covers 150 highly cited monograph series in SSH. Thomson adopted its own selection criteria. The basic criterion is the currency of a publication, a copyright date of at most 5 years preceding the inclusion in BKCI, and the complete information of authors. Very few Italian publishers participate, and among them we can find Gangemi and Edizioni della Scuola normale superiore, so this tool is not useful for an analysis involving Italian authors.

Daniel Torres-Salinas and Henk F. Moed (2009) and, during the same time period, Adrianus J. M. Linmans (2010) and Howard D. White et al. (2009) autonomously, suggested that an alternative to counting citations was to count the number of copies of books in library holdings. Linmans proposed the *Library holdings analysis*, the count of the monographs owned by libraries, as a complement of citation analysis and of author's productivity, to assess the SSH research. He highlighted that the count of monographs could correct the Anglo-Saxon bias, which is evident in citation analysis.

White and co-authors (2009) suggested a new bibliometric approach for SSH evaluation, the *Libcitation count*, an assessment based on the count of books in libraries, verified through online union catalogues. According to White, the measure is similar to bibliographic citations, and would complement, not replace, peer evaluations. Torres-Salinas and Moed proposed the *Library Catalog Analysis* (LCA) as a new type of bibliometric analysis, mainly to describe, on the basis of books owned by libraries, the structure of SSH scientific fields (in this case, Economics), the difference of distributions of books in Europe and USA, the distributions of monographs considering languages, and the publishers productivity. Moreover, LCA may help in the assessment of individual researchers and university departments. The

[3] The Book Citation Index contains 60,000 books in Science, Social Sciences and Humanities, and every year it adds 10,000 books. In the year 2014, the monographs in Social Sciences, Arts and Humanities accounted for 61%, and monographs in Natural Sciences 39%; Social Sciences presented a broader coverage than Arts and Humanities. http://wokinfo.com/media/pdf/BKCI-SelectionEssay_web.pdf

Libcitation count was recently adopted in a study by Zuccala and Guns (2013) to verify the relationship between the citations received by books in scientific journals in History and Literature, covered by Scopus, in a certain time span, and the presence of these books both in research library holdings and other library types, querying them through WorldCat.[4]

Although checking the diffusion of scholars' books could not replace peer review, it might still constitute a real help in evaluating scientific prominence. Therefore, we decided to carry out a deeper examination of the hypothesis of considering the inclusion of copies of books published by Italian scholars in library catalogues for the purpose of evaluation of the scientific research in SSH. The survey, sponsored by the Italian Agency for the evaluation of the University and research (ANVUR), was carried out in 2014 and 2015. In particular, we considered highly significant the collections of academic libraries, as they could be better representative of the measure of the interest of a scholarly community towards a specific monograph and thus reveal its usefulness.

As far as the methodology is concerned, it is worth mentioning that Torres-Salinas and Moed, with the aim of evaluating the inclusion of books on Economics in library catalogues, selected only academic libraries specialised in Economics. In our research, instead, we selected libraries not specialised in disciplines that we wanted to examine. Of course, specialised libraries in one single discipline should own all that is published in that field, therefore the inclusion in catalogues is predictable. Moreover, the authors declared that they selected mainly Anglo-Saxon libraries: 70% of them in USA, UK, Australia and Canada. This caused Anglo-Saxon bias in the procedure of research. On the contrary, in our work, we chose a well-balanced selection of libraries from Italy, France, Spain, UK and USA.

Regarding the selection of libraries, in his research Adrianus J. M. Linmans used as source the unit catalogue WorldCat (OCLC), by way of which he found data concerning the libraries from USA, UK and Netherlands. White, on the other hand, used the Australian National Bibliographic Database, and WorldCat in part. In the last two cases, there were unit catalogues. For our work, we preferred to verify the presence of books in a set of public online catalogues we chose as representative of different environments. They are national and academic library catalogues, namely 9 foreign and 13 Italian. We selected particularly significant and eminent library catalogues, specifically: 7 National Italian Libraries (the Rome National Central Library and the Florence National Central Library were excluded, as mandatory repositories of copies provided according to the Italian legal deposit act) and 6 academic library systems (that participate in SBN). Moreover, we selected 4 eminent

[4]The statistical examination of the relationships between book citations in the scientific journals covered by Scopus and the presence in libraries made it possible to highlight a significant connection between citations and *libcitation*, when there is a portion of books present in many libraries but with a low number of citations. The opposite case (high level of citations and low libcitation level) is less frequent. Analysing the type of the books, the authors found that a significant share of books with few citations but high presence in libraries consists in dictionaries and reference books. This indicates that citations and the presences of books in library holdings quantify different aspects.

foreign national libraries and 5 libraries of prestigious European and North American universities. Therefore, our work was not restricted to offering only quantitative results, but rather acquired a qualitative dimension, allowing us to evaluate better if the diffusion in library holdings could be a significant indicator of scientific relevance.

For the purpose of consulting the bibliographic information sources, we set up a database of monographs in two scientific fields, *History of books, Bibliography*, *Library Science* and *History of political institutions*, for a total of 497 editions published between 2000 and 2014. However, in order to check the presence of books in libraries, we also had to consider the presence of reprint copies. Therefore, the *corpus* of monographs at the end was made up of 563 units: 279 in *History of books, Bibliography and Library Science,* and 284 in *History of political institutions*. (For the analysis of data and results, see Part 2).

We adopted a methodology different from that of Torres-Salinas and Moed, first of all considering the strategy used with the aim of setting up the database of monographs. In 2009, the two authors extracted the monographs from 42 catalogues of specialised academic libraries, selected from seven countries. In our work, instead, we preferred to select the Italian authors belonging to the scientific fields we had chosen, using the authority database of the Ministry of Education, Universities and Research (MIUR). Then, we selected their monographs from the Italian SBN collective catalogue (Servizio Bibliotecario Nazionale) and refined the data excluding a small number of works not adequate for the research. Adrianus J. M. Linmans used a similar criterion. He selected a group of scholars from the members of the Faculty of Human Sciences in Leiden University (Nederland), using a sample of 80 authors (and 1135 books). White selected a group of professors from six Australian university departments.

Some considerations can be expressed and some important conclusions can be drawn from our survey. First, we have to underline that key factors conditioning the inclusion of monographs in library holdings are numerous. Factors that can influence the purchase of a monograph vary according to the type of library. In academic libraries the purchasing decision is certainly influenced by the need to have copies available if the monograph is included in university exam programs. At least for a part of monographs in one of the fields of our survey – *History of books, Bibliography and Library Science* – since there is a strong link with the professional activity, in Italy the purchases by libraries are influenced by the authors' reputation in the professional environment, in addition to learning needs. Textbooks in Library Science and monographs that examine relevant topics of the discipline in depth constitute an important area of interest for the professional improvements of librarians. In this case, librarians are involved in a dual role: the professional librarian, who must select acquisitions considering scientific criteria, and the user, who addresses *desiderata* to the library. Moreover, it is to be noted that libraries have often pursued the policy of acquiring all monographs published by editorial series of prominent publishers. This is a factor to take into consideration when we count the inclusion of monographs in library holdings.

Second, but of equal importance, with the aim of evaluating the inclusion of books in library holdings, since the act of purchasing has been defined by White (2009) as

an "indicator of esteem", we need to verify within the total number of presences in catalogues the share of monographs that were not purchased but rather donated to libraries. As with scientific publications, university professors and researchers, who are authors of books, frequently donate a copy of their monographs to the library of the department. Other copies can be donated to libraries specialised in the discipline of the topic of books. The libraries accept the donation, as the topics exactly match their holdings.

It is important to keep in mind that none of the authors here mentioned, Torres-Salinas and Moed, Linmans, White, who proposed to analyse library holdings to define a reliable indicator of diffusion and the scientific level of monographs, considered the incidence of gifts in library holdings. It is necessary to note here that it is impossible to find the information about whether a work was a purchase or a gift in online public access catalogues, both Italian and for the most part foreign,[5] and therefore the survey requires the cooperation of libraries.

During our study we could verify the incidence of gifts in only seven libraries and library systems[6] out of the 13 in the sample. It was a partial survey of gifts, and we offer it as an example of the incidence of donations in huge size holdings. In the scientific field of *History of books, Bibliography and Library Science*, 92 items out of 2165 (editions and reprints), i.e., 4,2%, were donated to seven libraries and library systems whose gift data we could verify. In the field of *History of political institutions*, 88 items out of 1097 (editions and reprints), i.e., 8,2%, were donated to seven libraries and library systems that we could verify.

It is convenient to underline the limitations of the survey in verifying gifts. It is necessary to notice that in the case study of the Sapienza Rome University library system, embracing 59 libraries grouped in six disciplinary areas, we counted a gift if the only copy included in a library of the system had been a gift. As concerns the Venice Cà Foscari University library system, it was possible to verify the gifts included only between 2010 and 2014 in the libraries of Humanities and of Linguistics; and as regards Milano State University library system we could analyse the data concerning the libraries of Humanities and Law studies, of History and of Political Sciences. Regarding the two latter University library systems, we could control the inclusion of gifts without verifying the inclusion of purchased copies of the same monograph.

[5] The Library of Congress OPAC shows the data concerning the acquisitions of books in the MARC fields 925 and 955 (see Chapter *Quality evaluation of online library catalogues, advanced discovery tools and linked data technologies*). After our survey, all the monographs of *History of books, Bibliography and Library Science*, were found to have been acquired by the Library of Congress. We did not consider it necessary to continue the verification with the monograph of *History of political institutions*.

[6] Milan Braidense National library, Turin University National library, Venice Marciana National library, Naples National library; Rome Sapienza University library system, Venice Ca′ Foscari University library system, Milan State University library system. The survey of the data concerning gifts was made possible thanks to the courtesy and cooperation of the Directors and some librarians of the mentioned libraries. It was not possible to carry out a survey of donated copies in the foreign libraries of our sample.

A further clarification is needed. In any case, it is impossible to count all monographs donated to libraries. When catalogue data of a library are recovered and managed in a new system, in fact, the data concerning the acquisition of items, including data on donations, frequently are not recovered. The type of acquisition, by gift, bequest, loan, purchase, or legal deposit can no longer be known. Moreover, if the library transfers the data from a management system to another, the manager can decide to not recover the data concerning the gifts.

This critical element in the survey of catalogue data, along with the difficulty to assess the respective extent of the purchases and of the gifts, as they are not data obtainable through public OPAC interfaces, makes the survey of the inclusion of authors' monographs in library catalogues a truly complex activity, not well grounded in certainty.

We highlighted that our examination revealed a very mixed picture and a substantial difference in the presence of the monographs in the libraries in which the surveys were completed. All the same, the analysis of the holdings of Italian and foreign national and academic libraries allowed us to have a wide view and to compare different realities. The results offer the opportunity to make an articulate analysis.

Considering the field of *History of books, Bibliography and Library Science*, three subsets can be established. The first consists of books, mainly historical topics, with very high presence in both Italian (SBN) and international word (WorldCat, for foreign libraries only), also in foreign academic libraries. The second subset presents a very high inclusion of monographs, mainly of professional topics, in Italian libraries (SBN), whereas the presence in international libraries (WorldCat) is lower. The third consists of books that present, on one side, low level of presence in Italian SBN, and on the other a high presence in foreign libraries, particularly in academic libraries, in both WorldCat and in the sample of libraries we selected.

With regard to the field of *History of political institutions*, for foreign libraries we did not check WorldCat data and we used only those of the sample selected. These libraries include a higher number of monographs concerning historical-institutional topics, or dealing with personalities and events of the past, that received greater attention in the foreign libraries than SBN libraries.

To conclude, we do not believe that quantitative criteria based on the count of presence offered by a single source are significant. With regard to Italy, high presence in the only SBN Italian OPAC cannot be considered a sufficiently solid indication to assess the diffusion, the attractiveness and the impact of monographs, and it cannot be sufficient to consider, *ipso facto*, these monographs scientifically eminent. On the other hand, our survey of the foreign libraries, in particular of the academic ones that are less influenced by contingencies in the purchase and without constraints in their choices, provided relevant results: the presence of a monograph in at least half of the foreign libraries of our sample is more significant than a high presence in the only Italian SBN collective OPAC.[7]

The factors that affect the decisions of purchasing and that determine, for a wide part, the presence of monographs in holdings, are numerous, different for the vari-

[7] http://www.sbn.it/opacsbn/opac/iccu/free.jsp

ous types of libraries, and often influenced by contingencies. Academic libraries, for instance, must purchase books used in the university curricula. Besides, a critical element is the library adoption of *approval plans* that is an acquisition method under which a library automatically receives new books with the right to return what it decides not to buy. On the contrary, Howard White (2009) believes that in spite of the automatic procedures of approval plans, library collections are often decided by the librarians. Approval plan methods were recently introduced in Italy and adopted mainly by academic libraries, which carry out cooperative policies in selection of purchases. Among the criteria to select online library catalogues with the purpose of doing quantitative analysis of their holdings, therefore, we should consider the explicit announcement of the policy of libraries in the development of the collections. (For issues regarding the development of collections, see Part 3).

Among the variables that must be considered in the assessment of the diffusion of monographs in libraries, two require particular attention: the number of gifts on the one hand, the relations of authors with foreign university libraries that may influence the distribution of books on the other. Finally, it is to be noted that the quality of the library catalogues also influences the analysis of books in libraries. Only the adoption of authority control criteria of authors and titles safeguards the reliable disambiguation of people having the same name, allowing for the certain identification of the author and title of the work for the purpose of an efficient evaluation and disambiguation of new editions from reprints.

In short, our survey found that:

1. The criterion of the presence of monographs in libraries cannot be adopted to establish the scientific value of monographs and, therefore, of their authors. We do not believe that it would be possible to demonstrate with certainty a connection between the high numbers of presence in libraries of a sample, even if chosen using objective and meditated criteria, and the high scientific value of authors.
2. However, it may be claimed with reason that verifying the diffusion of monographs in prestigious libraries, selected on the basis of well-balanced criteria, can constitute a reinforcing element of a positive or negative judgement that has already been formulated, after a well-managed *peer review*.
3. Our survey presented the results of a search for books in two union catalogues, Italian SBN and WorldCat, and in a sample of significant online library catalogues from USA, UK, Italy, France and Spain. Authors who carried out surveys in this field of interest, usually used only a union catalogue, often WorldCat, as their source and did not present their results broken down by the different types of libraries. Moreover, they apply the bibliometric analysis to data exclusively from WorldCat. In contrast, in our research we analysed the presence of monographs in a range of unit catalogues, like SBN and WorldCat, and online public access catalogues belonging to national and university libraries, Italian and foreign, and we showed that, thanks to a deeper analysis carried out on different library types, it is possible to offer a broader picture and to better verify the diffusion and draw conclusions about the possibility of using that indicator for the evaluation.

4. We demonstrated the effect of gifts of monographs in library holdings. Although the count of gifts could not be completed in all the libraries of the sample, and the foreign libraries could not be examined, the data retrieved show that it is necessary in both quantitative and qualitative analysis to know the percentage of donated monographs. Online public access catalogues rarely allow verification of the administrative status of an item, that is, if the monograph was purchased, donated, or if it was acquired due to the deposit law. It would be helpful if data concerning acquisitions and donations could be made available in public catalogues. As far as Italy is concerned, it would be advantageous if, starting from now, the data concerning gifts, changes and legal deposit copies could be retrievable by the public in MARC fields in the Italian SBN OPAC.

2 A Detailed Examination of the Survey: Methodology and Results

Our empirical research began with the decision to select a large and representative sample of monographic publications produced in Social Sciences and Humanities to be used for searching in online catalogues, in order to extract information about monographs and make some assessments of the reliability and quality of these online bibliographic tools.

The choice focused on the two scientific fields *History of books, Bibliography and Library Science* (limited to the bibliographic and Library science area) and *History of political institutions*, including monographs published between 2000 and 2014.

The first step was the identification of the authors that deal with research areas included in the two selected disciplines. The source of these names was the official and authoritative source of the Ministry of Education, University and Research (MIUR),[8] which collects all the names of professors, associate professors and researchers. Moreover, it was considered appropriate to add to the list of active authors in the disciplinary areas selected for our survey and found in the university staff official ministerial source, the authors who have obtained National Scientific Qualification (art. 16 of law no. 240 dated 30 December 2010[9]) in the call for candidates in 2012 and 2013 but were not already part of the university staff.

The group of authors thus identified counted 151, of which 71 belong to the *History of books, Bibliography and Library Science* field and 80 to the *History of political institutions* field. We selected all the monographs published between 2000 and 2014 by these authors.

[8] http://cercauniversita.cineca.it/php5/docenti/cerca.php

[9] This regulation defines a new procedure for the recruiting of university professors based on scientific qualification criteria. A national commission evaluates and assesses the candidates' scientific qualification.

A characteristic element in the methods used to create our database of monographs to be considered for the survey was a dual data collection channel. First the SBN National Union Catalogue OPAC[10] was consulted, and then the retrieved monographs compared to the Casalini *I libri* database.[11]

At this preliminary stage of the project we noted numerous discrepancies in the presence of the monographs in SBN OPAC and Casalini database *I libri*: monographs not in SBN were retrieved by comparing with the database Casalini *I libri* and, conversely, it was found that some monographs present in SBN were not registered in the Casalini database.

The detection of discrepancies in the two databases allowed us to investigate more thoroughly the specificity and to strengthen some insights of our investigation. In particular, our experience emphasised the usefulness of having a comparison between the two bibliographic tools for the verification of the details about authorship and for the detection of editions and reprints.

The work of creating the database of monographs for the two sectors was then refined by a careful selection of the publications to be included, which favoured the criterion of the scientific evaluation.

For the selection of the monographs to be included, we considered the submission of books by scholars in the case of scientific evaluation. So we chose the edition as the bibliographic category to build the database.[12]

This helped us to identify some guidelines for creating the database, which led us to some necessary methodological choices:

1. The choice to adopt a quantitative criterion, objectively identifiable, by which to establish the minimum acceptable consistency for the scientific assessment of a monograph.[13]
2. The decision to establish the selection of the authors of monographs on the category of edition, thus excluding reprints (which are counted in the phase of the detection of the monographs in the online catalogues, as certainly indicative of the diffusion of the works).
3. The decision to exclude collected works assembled by editors.
4. The option not to limit the database of monographs to only those with an ISBN, as the presence of an ISBN is not to be considered a guarantee of the scientific nature of a research project, but rather a commercial tool for the publishers to identify the editions, facilitate management of rights and monitor sales.

[10] http://www.sbn.it/opacsbn/opac/iccu/free.jsp

[11] http://www.casalini.it/ilibri/index.asp

[12] In our research we were inspired by ANVUR criteria expressed on the occasion of VQR 2004–2010 (and only partially by those announced for VQR 2011–2014), the 2012–13 ASN, and partly also by the criteria adopted for SUA- RD. The major difference in relation to the criteria adopted in the ministerial assessment is that in our project we decided to take into account both printed and digital books (e-books).

[13] In the absence of recent international standards, we decided to refer to ISO 5127:2001 Information and documentation - Vocabulary, (revised in 2010 and recently revised in 2015).

The database includes:

1. Books, both in print or digital format (e-books), by author or more authors, published between 2000 and 2014 with or without ISBN, preferring printed publications in the cases of monographs present both in print and in digital format.
2. New editions published between 2000 and 2014, even considering new editions of works published before 2000 that are, therefore, treated as assessable publications.

The types of documents considered in the survey include monographs, scientific papers, publications of sources, bibliographies, university textbooks, indexes, etc. Excluded from the database were publications having fewer than 49 pages, edited compilations, PhD dissertations,[14] private and limited editions.

The data collection included around 497 editions: 229 from *History of books, Bibliography and Library Science* and 268 from *History of political institutions*.

The second stage of our research sought to verify if widespread publication is a meaningful indicator, comparable to bibliometric indicators such as the number of citations received by a publication. Our methodological choice considered the need to refer to multiple sources and not just a single source to provide qualitative considerations in addition to quantitative results.

The previous monograph database, created for the purpose of assessing the quality of online library catalogues, was supplemented with the addition of reprints found in the selected libraries reaching the number of 563 bibliographic units (279 for *History of books, Bibliography and Library Science* and 284 *History of political institutions*) and this was used for the identification of the presence in bibliographic sources.

In order to check the spread of monographs in online library catalogues, a sample of representative and prestigious national and university libraries in Italy and abroad was selected.

The sample of Italian libraries was created by choosing seven national libraries (Cosenza National Library, Potenza National Library, Turin National University Library, Naples National Library, Bari National Library, Venice Marciana National Library, Milan Braidense National Library excluding the two National Central Libraries of Rome and Florence that receive obligatory copies under the provisions of the law on legal deposit) and six university library systems participating in the SBN catalogue (Cagliari University libraries, Venice Ca′ Foscari, Rome Sapienza University libraries, Turin University libraries, Macerata University libraries, Milan State University libraries).

In order to create a sample of foreign libraries four national libraries were chosen (Biblioteca Nacional de España, Bibliothèque nationale de France, British Library, Library of Congress) and five prestigious academic libraries (Universidad Complutense, University of California, Los Angeles, Princeton libraries, Oxford University Libraries, Harvard Library).

[14] The decision to exclude the bibliographic records about PhD thesis, originated from the observation that the presence on SBN OPAC of this type of monograph is sporadic and uneven.

We choose also to query the OPAC of the largest union catalogue WorldCat, for the *History of books, Bibliography and Library Science* only, in order to detect presence in foreign libraries, to conduct a test that would be the pair of the total amount of occurrences in the National SBN OPAC, thus reaching a total of 10 foreign and 14 Italian OPAC.

The procedure for collection and analysis of data included automatic extraction of data through dedicated software where possible, so in the end the data were collected through a combination of automatic and manual surveys.

To arrive at the cumulative results of the monographs' presence in the selected libraries, we manually counted the occurrences in national libraries (not including the two National Central libraries of Rome and Florence) and selected library systems in order to detect the partial and total circulation data in Italian and foreign libraries, national and academic; the total number of locations in SBN including the legal deposit copies; the total number of units present in the individual country/university libraries in the sample; and, finally, the total number of locations of reprints in the Italian and foreign OPAC.

With regard to the presentation of the research data, these were provided either in detail both in the aggregate form in tables which concisely present all research data concerning the assessment of the presence of the monographs in the Italian and foreign libraries in which the surveys were performed. The bibliographic data for the authors, the titles of the works and the elements that identify the publication have been obscured and rendered unrecognisable.

The analysis for the scientific field *History of books, Bibliography and Library Science*, along with circumstances of widespread monographs in Italy and abroad, and monographs more widespread in Italy and less abroad, found numerous cases of monographs with a very small occurrence in Italian libraries (totally detected in the Italian Union Catalog SBN) and a strong presence in foreign libraries. This could be an indicator of the impact of scientific monographs due to the fact that their presence in foreign libraries is not conditioned by the presence of the monographs in the university curricula, however, it shows the usefulness of further investigation of the terms and deployment strategies abroad of Italian publishers.

In the *History of political institutions* field we detected a greater dissemination of historical monographs or publications abroad devoted to biographies of characters, or to particular historical events. Particularly significant in this case was the comparison with the data about diffusion in the Italian libraries detected on SBN OPAC, where there are some cases of very modest presence in Italy compared to significant presence in foreign libraries.

More interesting is the comparison of the data in the two scientific areas surveyed. A first comparison concerns the absolute frequency of monographs that are part of the two scientific fields (Fig. 1).

The data, which refer to the total of monographs including the reprints, are influenced by two main elements:

1. A greater occurrence of reprints in the *History of books, Bibliography and Library Science* field (279 units, including 229 editions and 50 reprints), compared to *History of political institutions* (284 units, of which 268 are different editions and only 16 are reprints).

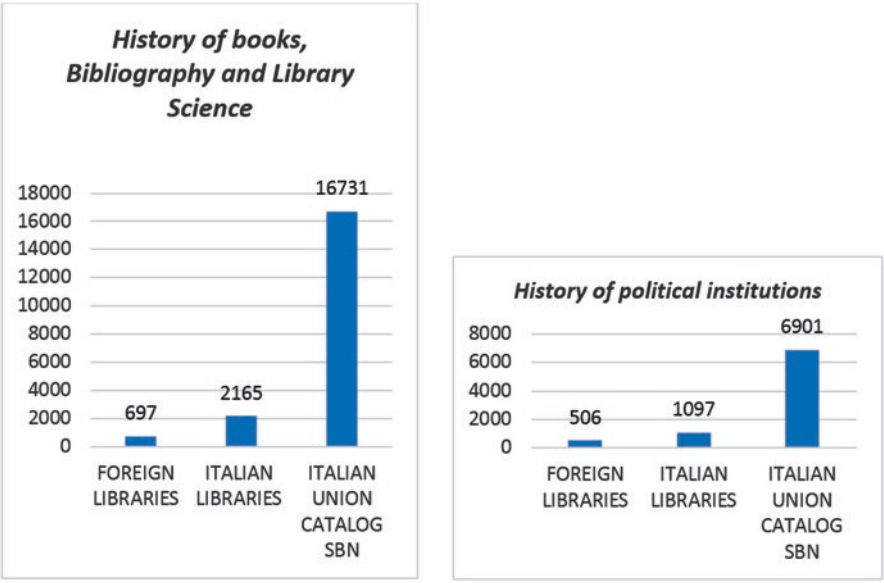


Fig. 1 Diffusion of publications in the two selected scientific fields

2. A greater presence in the Italian libraries participating in the Italian Union Catalogue SBN (36.9% of editions and reprints of monographs within the *History of books, Bibliography and Library Science* field exceed 50 occurrences in Italian libraries in SBN, while this percentage drops to 13.4% in the *History of political institutions* disciplinary field).

The comparison of data collected in the tables relating to the dissemination in foreign libraries (Figs. 2 and 3) reveal the presence of a large number of publications in Harvard University Libraries and Library of Congress, which stood out as a point of reference on an international level for the richness of the collections. Of the national libraries in our country the Milan Braidense National Library had the highest number of presences. The distribution of frequencies that relate to the presence in the Italian academic libraries show a rather large spread of publications in academic libraries of the two universities of Milan (State University Library) and Macerata (Figs. 4 and 5).

Carrying out the analysis on multiple sources turned out to be an appropriate methodological choice, providing the opportunity to offer elements of qualitative judgement that can be accompanied by quantitative assessments.

For while there are elements that can lead us to suppose a scientific impact, there are other thorny and problematic elements that instead reveal that the presence of monographs in libraries is not always to be considered a sufficiently reliable

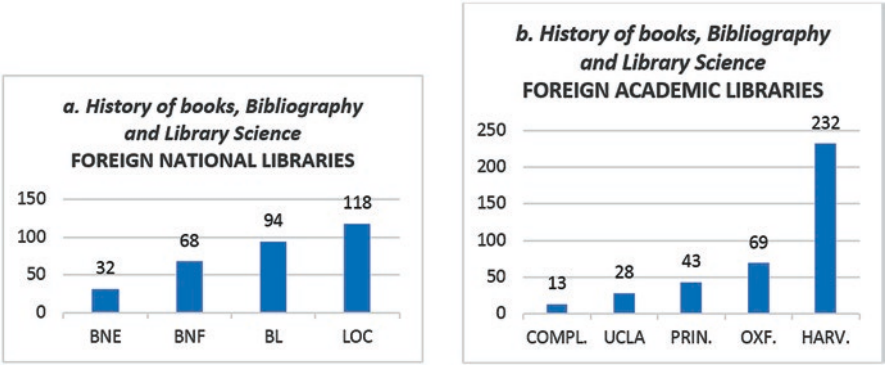


Fig. 2 *History of books, Bibliography and Library Science* (**a**) Presence in foreign national libraries (*BNE* Biblioteca Nacional de España, *BNF* Bibliothèque nationale de France, *BL* British Library, *LOC* Library of Congress) (**b**) Presence in foreign academic libraries (*COMPL* Universidad Complutense, *UCLA* University of California, Los Angeles, *PRIN* Princeton libraries, *OXF* Oxford University Libraries, *HARV* Harvard University Libraries)

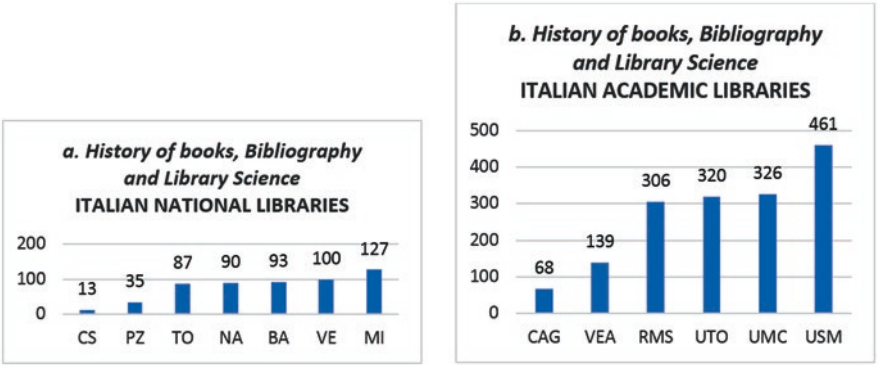


Fig. 3 *History of books, Bibliography and Library Science* (**a**) Presence in Italian national libraries (PZ Potenza National Library, CS Cosenza National Library, BA Bari National Library, VE Venice National Library Marciana, TO Turin National University Library, NA Naples National Library, MI Milan Braidense National Library) (**b**) Presence in Italian academic libraries (VEA Venice Ca′ Foscari, CAG Cagliari University libraries, UTO Turin University libraries, RMS Rome Sapienza University libraries, UMC Macerata University libraries, USM Milan State University libraries)

indicator to determine the scientific impact of a research publication. We found that these occurrences are not always the result of conscious choices, but are influenced by several factors, often contingent, like certain special purchasing strategies such as cooperative acquisition policies of publications, from purchasing plans (approval plans) or particular forms of giving and, in the case of foreign academic libraries, from relationships between university institutions.

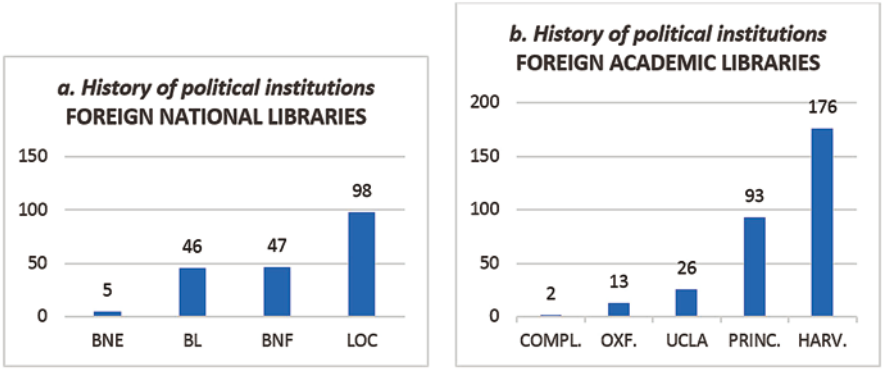


Fig. 4 *History of political institutions* (**a**) Presence in foreign national libraries (BNE Biblioteca Nacional de España, BL British Library, BNF Bibliothèque nationale de France, LOC Library of Congress) (**b**) Presence in foreign academic libraries (COMPL. Universidad Complutense, OXF. Oxford University Libraries,UCLA University of California, Los Angeles, PRINC Princeton libraries, HARV. HarvardUniversity Libraries)

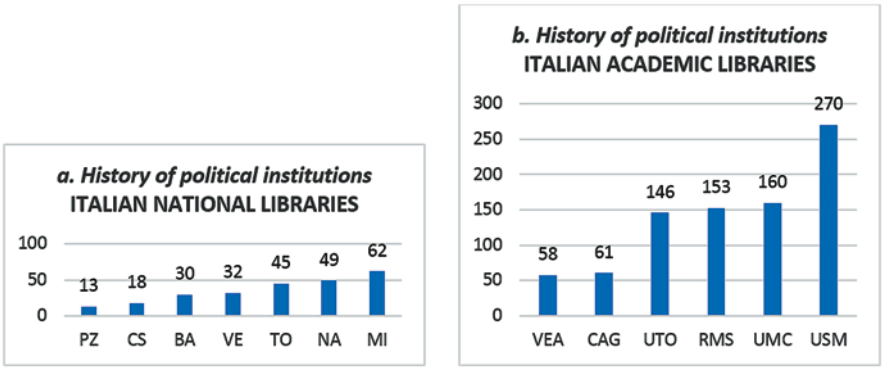


Fig. 5 *History of political institutions* (**a**) Presence in Italian national libraries (*PZ* Potenza National Library, *CS* Cosenza National Library, *BA* Bari National Library, *VE* Venice National Library Marciana, *TO* Turin National University Library, *NA* Naples National Library, *MI* Milan Braidense National Library) (**b**) Presence in Italian academic libraries (*VEA* Venice Ca′ Foscari, *CAG* Cagliari University libraries, *UTO* Turin University libraries, *RMS* Rome Sapienza University libraries, *UMC* Macerata University libraries, *USM* Milan State University libraries)

3 Library Collection Management in Research Evaluation Processes

An evaluation of scientific monographs conducted through bibliographic and item data available from library catalogues should also analyse the models and methods of collection development applied by the libraries themselves. In theory, every

library should adopt a library science canon[15] to build up its collections that meets its institutional objectives, the profiles of its target users, the composition and characteristics of the demographic, economic, social and cultural environment in which the library is found. The indicators of the attention paid to collection management are essentially two: the resources dedicated to it by libraries, which also implicate the budget allocated to this activity; the space dedicated to this subject by professional literature. The great attention paid to these issues by Anglo-American libraries corresponds to a lower reflection of the Italian professional literature.

The collection management procedures of a library depend on the adhesion to a reference model. One of the scenarios for the future of academic libraries indicates the end of traditional management models of collection growth and foresees selection procedures made exclusively by users. The review of the literature on this issue made by Linda Phillips and Sara Williams, relating to the period 1997–2003, shows that the main issues dealt with at that time concerned the selection and purchase of electronic resources. Hence, we are in the first phase of a change in the nature of academic collection, during which new assessment methods for purchasing are considered and cooperative acquisition systems, shared among several universities, are explored. A review of the literature published between 2004 and 2008 on the same subject shows that digital asset management and the issues arising from cuts suffered by library budgets are recurring topics. The need for the permanent evaluation of the use of electronic resource collections arises, then, in a period of economic difficulty, wherein consortium purchasing patterns establish themselves, directly related to the offer of scholarly literature in electronic format, in packages of publications, as an alternative to individual titles (Phillips and Williams 2004; Bullis and Smith 2004).

A review of the literature on collection development covering the years immediately following reveals that, in addition to the problem of budget cuts and a tendency to shift attention from the purchase of printed publications to electronic resources, academic libraries have in common two further approaches in this field: the growth of collections driven by patron demand and the digitisation of unique library collections (Thomas 2012). The report confirmed that the academic library is increasingly disintermediated from the discovery process. The radical transformations that libraries are facing require a redefinition of the roles of librarians and library functions even with regard to purchasing. Public and academic libraries must continue to ensure the quality of their collections to satisfy the needs of their users and to maintain high quality standards, comparable to other preeminent institutions. Management and growth of the collection are quality indicators for academic and research libraries, in particular.

The concept of bibliographic collection is the starting point for the organisation of a library's structure, and also for the definition of the relationship with the user. Although in the age of digital information the concept of the collection's centrality wavers, it remains a central point in the reflections on performance of library services. Currently, the meaning and the need of the collection are questioned: the

[15] The principle of the "canon" is expressed by Solimine (1999b).

reference to the historical roots of the idea of library as a collection, according to Klaus Kempf, serves precisely to redirect the service concept.[16] The origins of the collection, as we know it today, dates back to the late Renaissance and the modern age, when from the economic, functional, political and religious value of the medieval "treasure", they begin to consider the aesthetic aspects of the art collections. Once separated from the museum, the library collection assumes the character of a lasting and stable investment, an aspect of the library that has survived the ages. Today, the competition of market and of paid service providers invade every typical library activity. The library acts as an intermediary and care for the local collection is no longer considered a priority for user information needs. Printed publications are purchased as per the budgetary possibilities of the library and should theoretically be linked with the rest of the collection, with a view of the overall value. In fact, the choice of titles and purchasing decisions are increasingly shifted to users through a Patron Driven Acquisition service, according to which they can propose the order of monographs, printed or electronic, through bibliographic data previously entered in online catalogues. In this context, the librarians responsible for purchases have an increasingly irrelevant role.

The scientific literature pays great attention to the buying patterns defined as purchase-on-demand (POD) and patron-driven acquisitions (PDA). The tendency to purchase publications only when users require them serves to limit the spending budget, while at the same time increases the user involvement in collection building. Some studies document an increase in the circulation of books directly requested by users, even if they express concern about the effects, in the long term, of purchasing driven only by users. The problem that arises is to understand what impact this acquisition model could have on academic publications and how monographs (especially in the humanities) might be disseminated in relation to less known research areas. The PDA dominates the scholarly literature on the collection management (Levine-Clark 2010). It is a kind of acquisition guided by users, which enables them to request the purchase of bibliographic material through the selection of a link activated from the catalogue, or through similar techniques. Among the main advantages that it offers, it is effective in the management of costs, due to the fact that only publications that users require are purchased. In addition, there is a noted increase in the circulation of such requested materials, perhaps due to some sort of dragging effect that the use seems to produce. Some research also observed that PDA integrated with other forms of acquisition managed by librarians helps the development of a balanced collection (Shen et al. 2011; Schroeder 2012). However, librarians can manage data resulting from these transactions in many ways. Theoretically, the PDA method transfers the users' *desiderata* into digital environment. The risk arising from the uncritical adoption of these tools could be the exclusive purchase of what is requested, but that is not necessarily the best choice.[17]

[16] An analysis of library collections from the modern age to the digital library can be found in Kempf (2013) and in the following Kempf (2014).

[17] On the application of PDA in academic libraries see Vivarelli (2015). See also Vivarelli (2007), Bridges (2016), Goedeken (2015), Walters (2012).

Data-driven collection management is not a new concept for libraries. Indeed, it is a fundamental aspect of collection care. In recent years, librarians are implementing complex and cooperative methods that exploit data obtained from library transactions to establish more precise and shared management methods. The monitoring of data is useful both as a decision-making support and as an evaluation tool of changes that, over time, collections have suffered. This "analytical librarianship" can be used to find a compromise between two opposing forces: the budget cuts on the one hand, the continued growth in demand for collections and services on the other.[18] While the purchasing and circulation of e-books are topics extensively explored by professional literature, the fact that academic monographs continue to be used in hard copy text format, especially for humanities, social sciences and mathematics, is held in low regard.[19] For a long time, in academic libraries the collection management of printed books has been overlooked because of the predominance of digital resources. Today we are witnessing a shift in the management of the printed collection, also considered part of a larger collection, on a local and national basis, then there is a local collection repositioning in the national context.[20] Printed books were often excluded from such national and consortium management. In the United States they are also beginning to use data coming from the shared monograph collections of Hathi Trust for the print-on-demand policy, based on the preferences expressed by students and researchers (Kay et al. 2014).

The hybrid model of library collections, where printed materials coexist with electronic ones, remains predominant, with a clear decreasing trend in the purchase of printed resources. The increasingly frequent articulation of academic libraries in consortium models usually also involves new methods of selection and purchase of electronic and printed publications. In these contexts, the nodal point of the agreements concerns the creation of structures that buy and catalogue in a participatory way. In the academic field, shared purchasing plans aim mainly to avoid duplication and are achieved through the approval plan tool. The case of the shared purchases of printed collections is more controversial, because it involves accounting and logistical assessments related to the movement of users in the consortium's territory of reference.

3.1 Programmatic Working Documents

The collection development policy is a document that should contain an accurate description of collections and development policies that the library intends to adopt. Typically, it also specifies how the users' reports are evaluated in relation to

[18] For an analysis of data obtainable from the interactions between users and the catalogue of the New York Public Library, see Trombone (2016), post-print published on Academia.edu.

[19] Two reports, the UK Survey of Academics 2012 and 2015, conducted by Ithaka S + R, Jisc, and Research Libraries UK (RLUK), report in detail data for the printed monographs compared to ebooks; the exact opposite situation occurs for e-journals, which have now replaced traditional newspapers: Jisc et al. (2012, 2015).

[20] The use of printed monographs in academic collections is analysed in Showers (2015).

purchases and what is the attitude of the library towards gifts and exchanges.[21] This document takes on particular importance when libraries adhere to the increasingly frequent cooperative procedures of coordinated collection development, today highly adopted both on a territorial basis and for purposes of institutional cooperation. The purchasing policy cannot be managed by relying on the selection of individual publications to be included in collections; therefore, a policy document should state which are the enhancement criteria a library adopts, those of material selection, benchmark parameters, the level of depth that is attributed to the collections in the different subject areas. The collection development policy should, therefore, contain detailed analysis and development policies, as well as directions on how the library manages and evaluates the user *desiderata*, the bibliographic gifts and exchanges.[22]

In 1979, the American Library Association started a process of formalising the purchasing policy, and, in 1983, in the United States the experience of the *Conspectus* project began, the development of a collection planning method that thenceforth spread around the world (American Library Association, Collection Management and Development Committee 1979; Wood and Strauch 1991; International Federation of Library Associations and Institutions, Section on Acquisition and Collection Development 2001). The *Conspectus* evaluation grid is an operational instrument connected to the development and programming of collections. It provides clarifications concerning documentary coverage levels for individual topics and analysis carried out through the use of numerical codes defining the depth of the coverage level of a collection. It is based on a coding grid that marks the various fields of knowledge, consisting of 24 subdivisions articulated in 7000 subjects, defined according to the classification of the Library of Congress. For each subdivision, a numeric code indicates both the size of the collection (Existing collection strength), and the commitment in the current acquisitions (Current collection intensity), assessing the percentage of ownership compared to authoritative bibliographies. Conspectus also provides an encoding of the language coverage of the subjects. The project was born in US academic libraries, in the late seventies, also as an online inventory aimed at sharing information on the collections. In that period the Research Library Group[23] developed a system of collecting levels known as the RLG Conspectus, intended primarily for the uniform evaluation of collections in research libraries. The use of these collecting levels evolved from a tool for evaluation into a meaningful set of descriptors employed in library collection policy

[21] On the contents of the planning documents and on gifts see also Solimine (1999a, b) p. 15–17; in IFLA section Acquisition and collection development see also Cassell et al. (2008).

[22] On the political value of the collections card in library science reference should be made to Solimine (2008).

[23] The Research Libraries Group (RLG) was founded in 1975 by the New York Public Library and Columbia, Harvard and Yale universities. RLG grew to over 150 research libraries and worked to provide information discovery services, develop and operate collaborative programs, and create and promote relevant standards and practices. RLG became a part of OCLC in 2006.

statements. The same method was later adopted in other countries to promote the coordinated development of collections both in individual libraries, and in territorial library systems.[24]

3.2 Selection and Acquisition Procedures

For the analysis of publications they intend to buy, libraries can adopt evaluation grids like that of Whittaker, a tool that qualitatively analyses the bibliographical and bibliological characteristics of documents. It examines people involved in the realisation of the work, the author's and publisher's authoritativeness, the content, the design and production techniques.[25]

The approval plan is a method of streamlining acquisition procedures, applied to library purchases especially in the Unites States. It is based on various types of indicators, derived from classification schemes, from authoritative terminology sources, from publishing typologies. In compliance with a series of semi-automatic selection criteria, the supplier sends publications to libraries, with the possibility of restitution of materials not considered suitable for the library collection. The approval plan is, therefore, a form of outsourcing of a function normally carried out by libraries: the selection and the acquisition of publications. The characteristic of this procedure lies in the agreement between the library and the sending provider of new books. The library defines the acquisition criteria, while the vendor selects the book to send. The library can return the publications that it does not intend to buy within a percentage agreed. This service can be activated for the supply of books of a particular subject, country or linguistic area, but it is also possible for a library to acquire all its publications through an approval plan. The reasons leading to the definition of approval plan methods are, usually, related to the reduction of library staff and, consequently, the absence of subject bibliographers exclusively dedicated to the collection development. Moreover, an outsourcing of service can allow speeding up the supply and the processing procedures of publications.[26] The fundamental principle is to establish a workflow of automatic and systematic orders, defined by a relevancy profile, which should be agreed with all the library stakeholders.

---

[24] Three relevant English texts were translated in Italian: Association of research libraries (1993); the IFLA Guidelines of 2001 translated by Commissione Nazionale biblioteche delle Università e della Ricerca AIB ed (2001) and Bushing et al. (2008).

[25] It takes its name from Kenneth Whittaker, author of *Systematic evaluation. Methods and sources for assessing books*, London, Clive Bingley, 1982 (Whittaker 1982), also available in the Italian edition: Whittaker (2002).

[26] Reasons and results achieved with the approval plan are shown in Kempf (2006); the English version is also available on the website of the Associazione Italiana Biblioteche, *Outsourcing projects and approval plans. Ten years of experience in a large research library: a case study of the Bayerische Staatsbibliothek* <http://www.aib.it/aib/commiss/cnur/boekempf.htm3>. See also Brantley (2010), Mueller (2005).

Communication between libraries and vendors aimed at acquisitions may also be managed through automated procedures, the best-known example being EDI (Electronic Data Interchange), based on standardised buying patterns, also integrated with library software and capable of converting data in MARC bibliographic formats (Guerra and Pellizzari 2001; Sandy 2013; Morriello 2006). It is a way of implementing the paperless trade, in which the information flow between supplier and customer takes place through the Internet. EDI benefits can be summarised in three main points: the data directly received by electronic systems are more accurate than those manually transcribed; electronic transmission and the subsequent data acquisition by the computer is faster than the traditional paper-based sending methods, and also than the following manual transcription. EDI, then, increases the speed and accuracy of business transactions and reduces transmission costs and processing of information. It is generally adopted jointly with approval plan methods.[27]

The Italian Library Association (AIB) website dedicates a special section to collection development and management, where concrete examples of collection development policies are shown, approval plans and rules for the management of national and international gifts.[28]

The possibility of verifying the presence of the monographs in library catalogues for the purposes of bibliometric assessment is closely dependent on a series of necessary requirements that libraries and their catalogues should meet to be selected as valid targets for this kind of research. One of these is the transparency of criteria of collection development in libraries identifiable as query targets.

Libraries often don't declare their own methods of collection growth and selection. There are institutions that provide, among their quality standards, the adoption of guidelines and collection development policy for selecting purchases, gifts and exchanges of bibliographical resources. These assessment methods are frequently supported by deliberations of technical and scientific committees, appointed or elected within their organisations, which direct the library policies also regarding the further growth of collections. Particular attention must also be paid to libraries that use approval plan tools, which determine automatic purchases and returns of publications among libraries and publishers. The approval plan methods are newly established, they are administratively integrated in electronic management systems of libraries and adopted increasingly in library consortia, mainly academic, implementing cooperative policies in the selection and management of purchases. One of the criteria to select library catalogues as target for quantitative analysis on their data should take into account, therefore, the explicit institutional declaration of the methods adopted for collection development: the adoption of a publicly available policy; the presence of an institutional regulation for the selection and acceptance of gifts and legacies, this also freely accessible; the diffusion of the composition and tasks of the technical-scientific committees of libraries, if they exist; finally, the pos-

[27] An Italian academic experience is described by Di Girolamo and Pistelli (2005).

[28] The web page is edited by Rossana Morriello <http://www.aib.it/aib/lis/gest.htm>

sible adoption of approval plan methods and the setting of criteria for this instrument, if any.

Finally, the transparency of library collection criteria should also involve the statement of adhesion to patron driven acquisition models, if they exist. This popular method of purchasing publications entails the direct involvement of users, who suggest their purchase requests to the library, usually through tools available in next generation catalogues. The method is the digital equivalent of the traditional *desiderata* expressed by library users, so it falls under the formulas of direct involvement of users, highly adopted in public and academic libraries. The library should openly declare its adoption of patron driven acquisition policies, a condition that suggests a preliminary definition of the criteria and methods adopted, shared within the institution.

As a result, in order to analyse the presence and dissemination of monographs in library collections, the choice of a library catalogue would also involve the preliminary evaluation of the selection procedures and of rules adopted for the collection's development. Moreover, the analysis should be extended to the period of reference of evaluation activity. The clarity in communication of their own increasing collection methods is a factor of fundamental importance for the selection of libraries used in this kind of assessment.[29]

References

American Library Association, Collection Management and Development Committee. (1979). In D. L. Perkins (Ed.), *Guidelines for collection development*. Chicago: American Library Association.

Association of research libraries. (1993). *Manuale conspectus*. Roma: Biblioteca Nazionale Centrale di Roma.

Brantley, J. S. (2010). Approval plans, discipline change, and the importance of human mediated book selection. *Library Collections, Acquisitions & Technical Services, 34*(1), 11–24.

Bridges, K. (Ed.). (2016). *Customer-based collection development*. London: Facet.

Bullis, D. R., & Smith, L. (2004). Looking back, moving forward in the digital age: A review of the collection management and development literature, 2004–8. *Library Resources & Technical Services, 55*(4), 205–220.

Bushing, M., Davis, B., & Powell, N. (2008). *Il metodo Conspectus*. Roma: Associazione italiana biblioteche.

Cassell, K. A., Johnson, S., Mansfield, J., et al. (2008). *Gifts for the collections: Guidelines for libraries*. The Hague: IFLA. Available at http://www.ifla.org/files/assets/hq/publications/professional-report/112.pdf. Accessed 16 Jul 2016.

Commissione Nazionale biblioteche delle Università e della Ricerca AIB. (ed) (2001). *Linee guida per una carta delle collezioni usando il modello Conspectus*. Available at http://www.aib.it/aib/commiss/cnur/iflacons.htm3. Accessed 16 Jul 2016.

[29] In the Italian context, for a case of clarity in the adoption of an approval plan see the Economics and Law Library, University of Brescia, <https://www.unibs.it/ateneo/amministrazione/servizio-sistema-bibliotecario-di-ateneo/uoc-acquisizione-risorse-bibliografiche>

Di Girolamo, M., & Pistelli, M. G. (2005). EDI for acquisitions in libraries: a solution (still) for few? The case of the Library of the Università di Milano Bicocca. In: *Current issues in collection development: Italian and global perspectives, International conference on collection development, Bologna, 18th February 2005*. Available at http://www.aib.it/aib/commiss/cnur/boedigir.htm3. Accessed 16 Jul 2016.

Goedeken, E. A., & Lawson, K. (2015). The past, present, and future of demand-driven acquisitions in academic libraries. *College & Research Libraries, 76*(2), 205–221.

Guerra, L., & Pellizzari, E. (2001). Approval Plan in EDI: il futuro delle acquisizioni in biblioteca? *Bollettino AIB, 41*(1), 21–30.

International Federation of Library Associations and Institutions, Section on Acquisition and Collection Development. (2001). *Guidelines for a collection development policy using the Conspectus model*. Available at http://archive.ifla.org/archive/VII/s14/nd1/gcdp-e.pdf. Accessed 16 Jul 2016.

Jisc, RLUK, Ithaka. (2012). *UK survey of academics 2012*. Available at http://www.sr.ithaka.org/publications/ithaka-sr-jisc-rluk-uk-survey-of-academics-2012/. Accessed 16 Jul 2016.

Jisc, RLUK Ithaka. (2015). *UK survey of academics 2015*. Available at http://www.sr.ithaka.org/publications/uk-survey-of-academics-2015/. Accessed 16 Jul 2016.

Kay, D., Stephens, O., & DeNoyer, A. (2014). *Last copy services: what are the opportunities and benefits of collaboration?* Available at https://www.iii.com/sites/default/files/IIILastCopyServicesConversation3.pdf. Accessed 16 Jul 2016.

Kempf, K. (2006). Progetti di outsourcing e approval plans. 10 anni di esperienza in una grande biblioteca di ricerca. Il caso della Bayerische Staatsbibliothek. In: *Current Issues in Collection Development: Italian and Global Perspectives. Atti del convegno internazionale sullo sviluppo delle raccolte, Bologna* (pp. 137–148).

Kempf, K. (2013). *Der Sammlungsgedanke im digitalen Zeitalter. L'idea della collezione nell'età digitale*. Fiesole: Casalini libri.

Kempf, K. (2014). Bibliotheken ohne Bestand? Bestandsaufbau unter digitalen Vorzeichen. Bibliothek. *Forschung und Praxis, 38*(3), 365–397.

Levine-Clark, M. (2010). Developing a multiformat demand-driven acquisition model. *Collection Management, 35*(39), 205–206.

Linmans, A. J. M. (2010). Why with bibliometrics the humanities does not need to be the weakest link. Indicators for research evaluation based on citations, library holdings, and productivity measures. *Scientometrics, 83*(2), 337–354.

Moed, H. F. (2005). *Citation analysis in research evaluation*. Dordrecht: Springer.

Morriello, R. (2006). Acquisizioni tramite approval plan. Outsourcing o nuova opportunità per i bibliotecari? *Biblioteche Oggi, 24*(1), 22–31.

Mueller, S. (2005). Approval plans and faculty selection: Are they compatible? *Library Collections, Acquisitions & Technical Services, 29*, 61–70.

Phillips, L. L., & Williams, S. R. (2004). Collection development embraces the digital age: A review of the literature, 1997–2003. *Library Resources and Technical Services, 48*(4), 273–299.

Sandy, J. H. (Ed.). (2013). *Approval plans: Issues and innovations*. New York, London: Routledge.

Schroeder, R. (2012). When patrons call the shots: Patron-driven acquisition at Brigham Young University. *Collection Building, 31*(1), 11–14.

Shen, L., Cassidy, E. D., Elmore, E., et al. (2011). Headfirst into patron-driven acquisition pool: A comparison of librarian selections versus patron purchases. *Journal of Electronic Resources Librarianship, 23*(1), 203–218.

Showers, B. (Ed.). (2015). Data-driven collections management. In *Library analytics and metrics. Using data to drive decisions and services* (pp. 23–46). London: Facet.

Solimine, G. (1999a). *Le raccolte delle biblioteche*. Milano: Editrice Bibliografica.

Solimine, G. (1999b). Riflessioni sull'esigenza di un canone biblioteconomico per la gestione delle collezioni. *Accademie e Biblioteche d'Italia, 67*(1), 5–18.

Solimine, G. (2008). *La biblioteca. Scenari, culture e pratiche di servizio*. Roma-Bari: Laterza.

Thomas, M. L. (2012). Disruption and disintermediation. A review of the collection development and management literature, 2009–10. *Library Resources and Technical Service, 56*(3), 183–198.

Torres-Salinas, D., & Moed, H. F. (2009). Library catalog analysis as a tool in studies of social sciences and humanities: An exploratory study of published book titles in economics. *Journal of Informetrics, 3*, 9–26.

Trombone, A. (2016). *La gestione del catalogo partecipato: il ruolo dei bibliotecari nelle dinamiche d'interazione tra dati e utenti. In: Relazioni del Convegno Stelline 2016. Bibliotecari al tempo di Google. Profili, competenze, formazione* (pp. 81–86). Milano: Editrice Bibliografica. Postprint available at Academia.edu.

Vivarelli, M. (2007). La costruzione delle raccolte. Teorie e tecniche per lo sviluppo e la gestione delle collezioni. In G. Solimine & P. G. Weston (Eds.), *Biblioteconomia: principi e questioni* (pp. 39–59). Roma: Carocci.

Vivarelli, M. (2015). Formazione, sviluppo, integrazione delle collezioni documentarie. In G. Solimine & P. G. Weston (Eds.), *Biblioteche e biblioteconomia. Principi e questioni* (pp. 205–227). Roma: Carocci.

Walters, W. H. (2012). Patron-driven acquisition and the educational mission of the academic library. *LRTS, 56*(3), 199–213.

White, H. D., et al. (2009). Libcitations: A measure for comparative assessment of book publications in the humanities and social sciences. *Journal of the American Society for Information Science and Technology, 60*(6), 1083–1096.

Whittaker, K. (1982). *Systematic evaluation. Methods and sources for assessing books*. London: Clive Bingley.

Whittaker K. (2002). *Metodi e fonti per la valutazione sistematica dei documenti* (P. Lucchini & R. Morriello (Eds.). Manziana (Roma): Vecchiarelli. English edition: Whittaker K (1982). *Systematic evaluation. Methods and sources for assessing books*. London: Clive Bingley.

Wood, R. J., & Strauch, K. (Eds.). (1991). *Collection assessment. A look at the RLG Conspectus*. New York: Haworth Press.

Zuccala, A., & Guns, R. (2013). Comparing book citations in humanities journals to library holdings: scholarly use versus "perceived cultural benefit". In: *Proceedings of the 14th International Conference on Scientometrics and Informetrics. Vienna, July 15–20* (pp. 353–360).

The Social Impact Assessment in Social Sciences and Humanities: Methodological Issues from the Italian Experience

Luca Lanzillo

1 Introduction

In the *knowledge economy* the evaluation of research institutions is an important element, in particular with respect to universities. The need to verify, and justify to the public, the results of public funding of research activities has prompted several countries to develop evaluation procedures at a central/governmental level. For a long time the attention has been mainly focused on scholarly production – i.e. publications – and on its *academic* impact. Over the years the interest has begun to shift to the impact outside the scientific community, that is the economic, environmental, social, and cultural effects produced by research (Donovan 2008). While academic impact assessment has resorted to tools developed over time by the scholarly community (peer review and bibliometrics), *external* impact assessment is an unexplored field presenting some difficulties, especially the lack of methods and tools for its analysis and measurement. In this sense, a particularly thorny issue is the impact of Social Sciences and Humanities (SSH), whose outcomes are not easily identifiable (especially economic ones).

This chapter focuses on the assessment of external impact – hereinafter *impact* – of Social Sciences and Humanities within the policies of university performance assessment. The aim of this contribution is to offer some suggestions for an assessment that is more aware of the research impact, with reference to the Italian experience.

Although the peculiarity of the Italian socio-economic context – Italy is indeed an atypical country compared to other ones – the choices made in terms of research impact assessment can provide an interesting contribution even on an international level.

L. Lanzillo (✉)
Dipartimento di Scienze documentarie, linguistico-filologiche e geografiche, Università di Roma La Sapienza, Rome, Italy
e-mail: lu.lanzillo@gmail.com

© Springer International Publishing AG 2018

A. Bonaccorsi (ed.), *The Evaluation of Research in Social Sciences and Humanities*, https://doi.org/10.1007/978-3-319-68554-0_15

This chapter is structured as follows: Sect. 2 offers a brief overview of main issues related to research impact, in particular in SSH. Sect. 3 explains some peculiarities of the Italian context that are necessary to better understand the environment in which universities find themselves and Sect. 4 describes the Italian experience and ANVUR's[1] choices about assessment procedures. The following section reflects on some issues related to impact assessment in SSH research, including a brief comparison with the British experience of *Research Excellence Framework* (REF 2014). The last section offers some ideas and suggestions for future investigation. The whole chapter refers to research carried out in universities, but could be extended (with some changes) also to other research institutes, such as Public Research Organisations (PROs).

2 A Brief Overview

Scholarly literature on science-society relationships has generally referred to *technology transfer* activities (Bozeman 2000; Etzkowitz and Leydesdorff 1998; Geuna and Muscio 2009), whose recipients are essentially companies in the context of a new form of scientific research (Gibbons et al. 1994; Nowotny et al. 2001). However, this concept is rather narrow, because it refers to only one dimension of the broader *knowledge transfer* process. Over time, these processes have been gradually integrated in universities' activities and today they have reached a formal institutionalisation (Predazzi 2012). Together with its two traditional missions – education and research – the university has assumed a further task, generically called *third mission* that represents (just as generically) "all the direct interactions between universities and society, in addition to traditional missions of education […] and research" (ANVUR 2014, p. 559).

The terminological evolution from *technology transfer* to *knowledge transfer* and then to *third mission* has probably not coincided with an actual conceptual evolution. Therefore, there has been an erroneous overlapping between *technology transfer* and *third mission* concepts. Indeed, talking about third mission in most cases refers to economic impact of research, also known as *valorisation of research results*. Moreover, this type of impact is much more recognisable, quantifiable, evaluable than other ones (Benneworth and Jongbloed 2010; Olmos-Peñuela et al. 2014).

It is realistic to conclude that this misunderstanding could be one of the reasons for a lack of presence of SSH in scholarly literature about research impact. Yet recent literature in both Humanities and Social Sciences has started to examine this issue more closely, trying to identify distinctive ways in which this research contributes to society, without necessarily being associated with economic or pragmatic values. Some studies have begun to reflect on knowledge transfer in SSH in the

[1] The Italian National Agency for the Evaluation of Universities and Research Institutes.

context of specific countries (e.g. Gascoigne and Metcalfe 2005; Landry et al. 2001; Munro 2016).

SSH disciplines have always established interactions with society, although these relationships are often *hidden connections* if the attention is centred on economic impact: they obtain visibility only when we pay serious attention to "people-based, problem-solving and community orientated activities" (Hughes et al. 2011, p. 54). Indeed, SSH share a common interest about human phenomena, and their research results can impact on society in the form of invisible innovations, intervening in goals, ideas, values of people and their communities (British Academy 2004; European Commission 2007).

The institutionalisation of the third mission also requires appropriate policies that provide all the actors involved with the tools needed to reach the largest possible impact. In this sense, the European Commission strongly encourages science-society relationships, scientific information circulation and knowledge transfer processes as engines for the European area (Basili 2013; Soriano and Mulatero 2010). This encouragement is put into effect by EU research funding programmes: Horizon 2020 requires that project proposals show their expected impact, in order to also highlight the sustainability of projects over the years. The European Commission seeks to fund projects that are able to extend their effects beyond the funding period, after which they have to continue independently. Unfortunately, this idea is often related to a simple request to challenge unemployment and to produce new job opportunities, more often supporting economic impact having an immediate effect rather than social impact that is visible only after some years.

Although they may seem disadvantaged, SSH could however obtain an important space in terms of interdisciplinary research. The Commission fully recognises SSH's crucial role in societal development, especially according to major societal challenges (European Commission 2015b). In this sense, Horizon 2020 changes approach compared to previous framework programmes, because it has placed its "focus […] around 'Challenges' rather than disciplinary fields of research" (European Commission 2014, p. 18). Unfortunately, collaboration is an activity still uncommon in SSH, especially due to the predominantly solitary nature of their research (Research Information Network 2011).

Lastly, the more evident consequence of the institutionalisation process of the third mission is its gradual integration within research assessment programmes. Some countries have integrated, in different ways, external impact assessment within their national research assessment exercises (RAND 2010): the United Kingdom has integrated *impact case studies* within its Research Excellence Framework (REF2014), the Netherlands evaluates *societal relevance* within its Standard Evaluation Protocol (SEP 2015–2021), Italy considers *third mission* within its Valutazione della Qualità della Ricerca (Research Quality Assessment, VQR 2004–2010 and current VQR 2011–2014).

However before continuing it is appropriate to offer to the reader a brief description of the Italian context, an atypical country with regard to research and development policies.

3 Historical Remarks on the Evolution of the Higher Education System

Even though Italy is a relatively young state, a little more than 150 years old, it is also the cradle of the university: in Bologna, at the end of the eleventh century, the nucleus of the oldest European university (probably also the oldest in the world) was formed. Aside from its long history, it is important to highlight that at the time of Italian national Unity (1861) the new Italian state inherited from the previous pre-unification states a large number of universities, each with its own internal organisation and a special relationship with central power (in Italy there were lots of *centres*).

The new state thus began to establish a unitary university model, promoting the model established in Europe during the nineteenth century: theorised by Wilhelm von Humboldt, it is the model of a university targeting a small group of elites – the future leaders of a *classic* industrial society – seeking to fulfil two missions – research and education – while the entire system is firmly based on two pillars of freedom of the scholarly community and the loneliness of the scholar (Garbini 2010). It is a model in which all disciplines revolve around philosophy.

Despite the great efforts of Vito Volterra (1860–1940), persuaded that scientific research might be able to transform the economy and society, promoter of the International Research Council[2] and of the Italian National Research Council (CNR), Italy soon became a victim of the highly centralised and nationalist politics of Fascism. A very careful education policy, promoting a reform of school and university that created very strong imbalances between the Humanities and scientific disciplines – giving a much greater weight to the former and strengthening the university's idealistic vision, based above all on philosophy and law – and further increased the centralised statism in which scholars are treated as state officials from whom absolute loyalty is required (Garbini 2010).

At the end of World War II the influence of the United States in Europe gradually led to the crisis of the Franco-Central European model of society, a crisis that also involved the scholarly community. US influence also meant *importing* the idea of development proposed by Vannevar Bush (1945). But while at the legislative level the Humboldtian model was being phased out, the previous idealistic setting continued to persist, especially at a cultural level.

After the initial economic boom – during which the policies for scientific research (especially nuclear, electronics, biochemistry and industrial chemistry) altered the structures of traditional power in the scientific community – there was a series of events that strongly hindered the country's scientific-technological and economic development (Paoloni 2010).

Italian policies started to lose interest in the model adopted by other Western countries (based on a close integration between scientific research and production system), opting for an opposite model called *development without research* (Greco

[2] Precursor of the current International Council for Science (ICSU).

and Termini 2007). This approach focuses on small- and medium-sized enterprises, whose weight had steadily grown in the Italian industrial system, but that were not able to offer a significant contribution in terms of R&D-based innovation. The Italian landscape of small- and medium-sized enterprises was indeed characterised by a constellation of family businesses – often with a long tradition – in which new generations were employed at the end of compulsory education (or at the end of secondary school). The founders often came from independent or artisanal jobs or from industries characterised by a relatively simple production system (e.g., in the local agricultural sector). This landscape was set in a political context characterised by a strong administrative-territorial fragmentation – provinces, small municipalities, regions – further *stiffened* by an exasperated centuries-old socio-cultural *particularism* of the Italian peninsula.

In the mid-Sixties, despite its economic boom, Italy was the poorest among Western countries and chose to focus entirely on the production of goods with low knowledge content, giving up a serious investment in higher education and research. The objective of that choice was to be much more competitive than other Western countries in the production of this type of goods (like food, furniture, fashion industries, common white goods). In a country where there was no economic demand for R&D-based innovation, there was not even a demand for universities and research centres. Italian society would not know how to use highly qualified people in a productive way. In this type of society, with no one to talk, the university could only be a weak element regardless of the quality of the scientific community or its tradition.

By the end of the 1980s and the 1990s, with the start of the globalisation process and the emergence of new countries in global competition, this strategy revealed its weakness: hence began a long productivity crisis, a structural crisis of its productive specialisation. While other countries were increasingly immersed in the *knowledge economy*, the Italian system, to use the words of Solimine, was instead substantially "lacking in knowledge and skills, a country 'without knowledge'. We are so ignorant that we do not even realise how serious and dangerous our level of ignorance is, and the need to resolve the matter as quickly as possible" (Solimine 2014, p. 3). This ignorance also generated a somewhat negative perception of the value of education and training.[3]

In this context, external research impact assessment is certainly a difficult task, requiring the development of appropriate methods and instruments. This is true even more for SSH, which, because of their peculiarities, are more exposed in the delegitimisation process that has taken place over the years based on the concept of *usefulness*.

Assessing how universities work in the third mission is indeed a way to verify the quality of the actions undertaken and the results achieved, but it is especially an opportunity to show what universities already do for society and what potentially they could do (Donovan 2011). On the one hand it can therefore be an important

[3] The fact that *Bush Report* was translated in Italy only in 2013 is probably an evidence of some changes.

instrument for *advocacy* of universities; on the other hand, however, it could be a harmful weapon – both for the university and society – if it resorts to *recycled* methods and instruments or otherwise not appropriate to investigate all dimensions of the impact.

4 The Assessment of the Third Mission in Italy: History, Methods, Instruments

4.1 Regulatory Framework

Italian legislation has recently recognised the third mission as the institutional duty of universities. The Ministry of Education, University and Research (MIUR) established periodic assessment of the third mission (ministerial decree 47/2013) as part of the Self-Assessment, Periodic Assessment, and Accreditation system (*Autovalutazione, Valutazione periodica e Accreditamento*, AVA) (ANVUR 2013a).

There was yet some experimentation within the first national research assessment exercise (VQR 2004–2010): each institution submitted data about technology transfer activities (patents and spinoff companies) and their results were classified by eight indicators that were used as a starting point for following developments. After this experience, ANVUR highlighted the difficulty of describing the Italian context in detail using only these eight indicators and the impossibility of using them for funding allocation (ANVUR 2015). The Agency has worked hard on these subjects, organising regular public workshops (2013–2016) to foster debate and sharing best practices, also at an international level.[4]

Meanwhile the Annual Unique Form for data collection of Departmental Research (formally *Scheda Unica Annuale della Ricerca Dipartimentale*, SUA-RD)[5] was established and used within the AVA system. The aim of SUA-RD is the standardised and comparative collection of data regarding academic activities at a departmental level: it is structured into three parts, the third of which is dedicated to third mission activities. Data collected by SUA-RD – supplemented by additional sources (ANVUR 2016) – are evaluated within VQR 2011–2014 by an Experts' Committee (formally *Commissione di Esperti della valutazione della Terza Missione*, CETM), which, on the basis of ANVUR's Manual (2015), will compile a general final report and one final report for each structure evaluated. This procedure is carried out at the same time as ordinary VQR and still does not contribute to the funding allocation. This follows a shift in the administrative ministerial framework, which has moved the assessment of the third mission from the activity aimed at

[4] Workshops were an opportunity to not only meet experts from other countries (e.g., Sweden, France, the Netherlands) and international organisations (European Union, OECD), but also to talk to representatives of Italian society (local authorities, industry, journalism).

[5] See previous chapters in this volume for its most comprehensive description.

self-evaluating and evaluating the educational activities of universities (the AVA system) to the activity aimed at research assessment (the VQR system).

How are third mission activities defined? ANVUR established that only formally approved experiences (e.g., by an official agreement between department/university and an external subject) can be submitted for evaluation. The Agency does not choose to evaluate activities managed in an informal way because: i) third mission is an institutional duty, not an individual duty; ii) institutions have to prove their commitment to activities, that must be part of a clear and coherent institutional strategy (at a department and/or university level).

4.2 Third Mission Activities Classification and Assessment Methods

Olmos-Peñuela et al. (2014) list a different model of categorisation related to university-society interactions. ANVUR implemented a model described in its *Third Mission Assessment Manual* (2015)[6] and used by the panel responsible for evaluation (ANVUR 2016).

Third mission activities are classified in two main areas related to economic impact (*Valorisation of research*) and social impact (*Production of public goods*), respectively. Both areas are split into four types of activities for a total of eight different types.

The four types of activity related to Valorisation of research (Group A) are:

(a) Intellectual property management
(b) Academic entrepreneurship (spinoff companies)
(c) Third-party research
(d) Intermediaries

Knowledge transfer, in this case, consists in transforming scientific knowledge into useful knowledge for productive purposes. How does this transformation take place? The crux consists in the transformation of a public good (i.e., knowledge produced by the public research system) into a private good available to someone who exploits its potential for commercial purposes, with a number of regulation issues.

The four types of activities related to Production of public goods (Group B) are:

(a) Creation and management of cultural heritage
(b) Clinical research and training (health protection)
(c) Lifelong learning
(d) Public engagement

[6]The first version of the Manual (February 2015) was made public for comments on possible changes/additions, following which the current version was drafted (April 2015).

In this case the public good (knowledge) is converted into another good or service of a public nature. Their recognition and measurement are made difficult by the fact that they are often carried out on a personal level by individual scholars and not at an institutional level by universities. These activities could be related to the concept of *Popularisation of Science* or *Science Communication*, although the term *science* is usually linked to STEM disciplines. Furthermore, the concept of *science communication* is unpopular among the SSH community, very often interpreted as generalisation (in a negative sense). This type of third mission should actually be its essence (and its purpose): thus the principal purpose of academic institutions should be their concrete participation in the building and dissemination of *scientific citizenship* (Greco 2007). This does not mean that they should work to establish a technocracy, but rather they should disseminate knowledge among citizens in order to help them to better understand their reality and participate consciously in democratic life. A reality characterised by increasingly complex challenges (social, cultural, environmental, technological) that require appropriate critical thinking.

The current Italian classification is the result of a long reflection, begun in 2011 within the VQR 2004–2010 – the first assessment of third mission activities in Italy – through a survey based on eight generic quantitative indicators. Five indicators were associated with economic impact,[7] other three with socio-cultural impact.[8] In this first assessment – the results of which were published by ANVUR (2013b, 2014) – one must take into account: (i) the economic landscape described in Sect. 3; (ii) a strong imbalance in favour of activities with an economic impact; (iii) the issue of data availability, based on the experience gained by universities. On that latter point, it should be emphasised that the exploitation of research activity through contract research was introduced in the Italian academic system 30 years ago[9] and some kind of codification already exists. In contrast, socio-cultural activities are still a largely unknown area.

Starting from data collected within the first VQR, ANVUR established a working group in order to define more structured indicators able to consider all areas of action. Furthermore, it announced two calls for experts for both areas of the third mission in order to establish a register of reviewers. From this register, 28 experts have been nominated for CETM, organised into two panels (CETM A and CETM B) that evaluate data collected through the SUA-RD[10] for the years 2011–2014.[11]

[7] Namely: Number of patents granted, number of accredited spinoffs, participation in incubators, participation in consortia aimed at technology transfer.

[8] Namely: Number of activated archaeological excavations, presence of museum centres managed or co-managed by the university, number of *other* third mission activities (except third-party activities).

[9] Introduced through Presidential decree DPR 382/1980, which allowed universities to carry out third-party research.

[10] As integration for data provided by institutions and collected by ANVUR, panels have also used: institutional portals, other institutional sources, data specifically requested by individual institutions, direct hearings with delegates of institutions.

[11] For Group B activities and third party activities of departments, only the years of 2013–2014 are considered.

The final CETM report is independent from the final VQR report and has essentially exploratory purposes (ANVUR 2016).

The panel assesses not only activities related to these two groups but also a third area, the institutions' third mission strategies. It is important to highlight that assessment must take into account the specific nature of each institution and the context in which they operate. In particular, the three main factors considered are (ANVUR 2016):

(a) Specialisation of institutions, i.e., their disciplinary profile.
(b) Their social-economic and territorial context.
(c) Their size in relation to their human resources.

The need to evaluate data characterised by an extreme heterogeneity, in addition to the experimental nature of the evaluation, determined the choice of a mixed evaluation method (qualitative and quantitative) on the basis of an *informed peer review*, i.e., the analysis of quantitative data through the judgement of experts. In this sense it is important to emphasise ANVUR's choice – even with differences between the two groups of activities – with respect to different objects within the same evaluation system and of the different maturity of indicators used. Even though this situation might suggest a lack of consistency, on the other hand a differentiated approach makes it possible to investigate the individual issues more thoroughly, an essential element when there are evolving objects.

The adoption of an evaluative mix is significant especially when the aim is to evaluate the full range of knowledge domains, that is, when the object is not restricted to economic and technological impact, but extended also to socio-cultural returns that cannot be counted, i.e. those of SSH. As in the context of research assessment, it is perhaps important to briefly consider how the inclusion of these disciplines requires special attention.

5 Research Impact Assessment: Some Reflections About SSH

The recent introduction of the concept of third mission and its assessment in the context of the Italian academic system is a change of pace compared to the scenario described in Sect. 3. In Italy, where the science-society dialogue is structurally very weak, the idea of assessing the status and the health of these exchanges can be an important opportunity to provide a picture of what universities do every day for their local communities but, for various reasons, they do not communicate clearly. This is even more true if focus is shifted to the SSH, whose impact is much more socially widespread albeit often much less noticeable. Given their great heterogeneity, SSH are a difficult case and require a multidimensional analysis.

Some SSH research topics are related to broad issues of collective interest, not only for experts: e.g., historical research, studies on communication, current events in culture and politics, contemporary society's dynamics, issues about economic development. Science communication and public orientation activities do not

exclude scientific rigor, which is their foundation: the concept of *active citizenship* and a conscious involvement of citizens in public life depend on the deep difference between mass media's *idle chatter* and accurate information.

For this reason, especially in SSH, it is necessary to pay attention to the possible relationships between research quality and assessments of the third mission. Although they should be evaluated in different and separate ways on the basis of suitable criteria, the insufficient consideration towards third mission activities has compromised the perception of their importance: "qualified scholars should not spend time on these type of activities". In this way, scholars have totally neglected the transfer of research results to a circle that reaches beyond their scientific community. An example is the relationship between current research and school textbooks, often characterised by an embarrassing delay and by a simplistic rehash using clichés and dated structures. The university as a whole has been damaged by a context that has emphasised its separation from society. Through a correct valorisation of editorial products for public consumption, it is possible to endorse scholars' commitment in knowledge transfer activities.

Furthermore, with regard to the Italian situation, we cannot underestimate the boost to the sustainability of research due to EU policies that are oriented towards inclusion of citizens in the innovation process. This trend started with the last part of the 7th Framework Programme (European Commission 2015a), which produced a large variety of research outputs. Joining this type of funding criteria and encouraging research project that are strongly public-oriented strengthen the production of editorial and popular outputs.

In this sense, we suggest that there are three elements to be considered in an appropriate assessment of the research impact of SSH disciplines. First, we should examine the relationship between research, education and third mission and the role of the monograph/book. While in STEM disciplines these three missions are clearly identified and somewhat separated, in SSH their boundaries are not evident. With regard to the production of public goods (Group B), their potential overlap in research and teaching activities is easily noticed: the main differences, indeed, concern the specific target audience, not the type of activity. Furthermore, books are still a good channel for knowledge dissemination because of their history and *familiarity* within society. For this reason (and depending on specific disciplines), it could often be difficult to distinguish between a research monograph, an academic textbook and a science communication monograph. The elaboration of criteria for assessment activities must considers these issues. We find here an interesting overlap between the issue of third mission in SSH and the call for an appropriate framework for the evaluation of books, which is the object of several chapters in this volume.

Second, the *effects* of the third mission in SSH have an undeniable local dimension because "scientific knowledge impact manifests itself probably more in the geographical vicinity" (ANVUR 2015, p. 6). It means that these activities, especially those related to social and cultural impact, could *drive* scholars towards an opposite direction from internationalisation, one of evaluative criteria used for research products (in particular, in scientific publications). Besides, unlike research

in STEM disciplines, research in SSH usually has a very low level of *scalability* that does not allow it to effectively extend the results obtained to a wider audience than that for which they were produced. Therefore, a balance between research and its valorisation is indispensable.

Third, third mission activities pertain to universities, not to scholars. Research and education are part of the duties of each scholar; third mission, on the contrary, is an institutional duty whose deployment is left to a variable combination of initiatives at organisational, department, team or individual levels, according to their own history, competencies, interpretation of the needs of society, and strengths. Possible lack of experiences in some scientific areas does not entail a negative assessment (ANVUR 2015). For this reason, SSH scholars could decide to retreat and universities (in particular those that are generalist and large) could be satisfied with economic results reached by STEM disciplines, recognisable and quantifiable in assessment activities. On the other hand, ANVUR's choice to evaluate the strategic lines of the third mission of the institution is a great opportunity for SSH to carve out a significant space within the policies of their universities, taking advantages of this situation (in a positive sense).

It has been said above that Italy is an atypical country in the global landscape, in which research still does not play a central role in the economy and the development process. For this reason, the decision to proceed with long-range experimentation on the third mission could allow institutions to gradually equip themselves to deal with these new issues, encouraging their advocacy towards society.

Therefore it is clear that it was not possible to import a model already used in other contexts. Choices made by ANVUR are diametrically opposed to those made in the UK within the Research Excellence Framework, in which the impact is one of the three research quality assessment criteria[12] (REF 2014; 2012). The British REF asked universities to submit a number of case studies, proportional to the number of their scholars, showing the impact of research carried out since 1993 and identified in the period 2008–2013.

A first observation is that very often the impact of research is not an intentional effect but, on the contrary, it is a process of *knowledge spillover* in society: the impact of the university is much more than just the third mission, but it is the overall set of effects, intentional and unintentional, generated by research, education, and third mission activities.[13] As evidenced by Piazza and Matteucci (2015), the REF has generated a wide debate, and criticism by the SSH community revolves in particular around two aspects: (i) the difficulty in identifying socio-economic impact and, therefore, the risk of a substantial cut in research funding; (ii) the lack of clarity in the selection and evaluation criteria of the case studies presented.

Regarding the REF, Penfield et al. (2014) also identify five challenges that show the difficulty of comparing the impact of different disciplines:

[12] Criteria are: Outputs (65% of weight), Impact (20% of weight), Environment (15% of weight).

[13] Reflection proposed by Andrea Bonaccorsi at the workshop *The evaluation of the third mission in the VQR 2011–2014: a comparison with universities and research institutions*, organised by ANVUR (Rome, 28 June 2016).

1. Time lag between research and impact.
2. Changes in impact over time.
3. Difficulties in linking impact to specific research.
4. Possible lack of adequate evidence to demonstrate research impact.
5. Sometimes it could be difficult to collate evidence.

The British context is certainly different from the Italian one, as the degree of maturity in the field of research evaluation is higher. The REF model, albeit with a weight balance attributed to the evaluation criteria, merging impact and quality, bear the risk of communicating the idea that research could work towards the first aspect at the expense of the second. If they are separated, as in the Italian VQR, departments will be forced to seek both of them: the idea is that if the quality is high, this also favours society.

Moreover, especially in the production of public goods, the monitoring of activities has a fundamental relevance in order to increase accountability of institutions. The aim of the assessment process is therefore not only the impact – that could become manifest after many years – but above all the strategies adopted by institutions, as well as the actions to verify, maintain and expand them over time.

6 Concluding Remarks

Over the last few years, in Italy there has been an increasing interest in the socioeconomic impact of research and the university's third mission, well behind other Western countries. Talking about social impact within SSH disciplines should be simple, almost expected. But often the emphasis is on economic impact, not cultural.

Moreover, it is important to highlight that in Italy the issue of third mission received its legitimisation and institutionalisation mainly in the context of evaluation, rather than from an autonomous political process (see the chapter by Blasi, Romagnosi and Bonaccorsi in this volume). This has accelerated the creation of awareness in universities, but has also certainly made manifest the need for a serious and extensive debate among scientific communities. Before thinking about the importance of assessment – which is essential – they should reflect on the object of assessment itself.

This also creates an interesting research agenda for scholars in evaluation. For example, in investigating the perception of scholars about the notions of interdisciplinary research, public orientation and cooperation, Ochsner et al. (2016) find that they are not always perceived as positive elements in socio-humanistic communities. Scholars who dedicate themselves (also) to dissemination of their results are considered as time-wasters, stealing time from real research. According to this view, research (and education) is the main activity of scholars, and their scientific quality should be evaluated only through these, regardless of their personal inclination

towards dissemination. However it is not acceptable – especially in the current context – that this inclination could negatively affect their scientific curriculum.

Nevertheless, in addition to a more conservative position, the number of SSH researchers interested in social impact is increasing, although they often have to face a public that does not really trust the importance of their research (Scamuzzi 2015).

In conclusion, there are four issues that should be considered in the reflection on the research impact assessment within SSH:

1. *Reception issue*: it is necessary to begin a serious debate within scientific communities about the importance of social impact of research and of the dissemination of results. Debate cannot be focused only on assessment, but it must be originated from a large and shared reflection on the object of assessment itself.
2. *Democratic issue*: SSH can offer an essential contribution in building *scientific citizenship* among society. Taking part in local development of their local regions, universities cannot renounce socio-humanistic contributions: critical thinking, mental flexibility and the ability to understand surrounding reality are just some of the skills necessary for citizen awareness (Nussbaum 2010).
3. *Exposure issue*: SSH disciplines establish lots of relationships with society, but they are often *hidden connections* as results of non-institutional interactions, frequently at a personal level. Studying the in-depth social impact of SSH research makes it possible to increase exposure, overshadowed by relevance given to economic impact of research.
4. *Social acceptance issue*: studying and evaluating SSH research impact combats the popular stereotype of their futility and of their subordination to hard sciences (Olmos-Peñuela et al. 2015), as well as more clearly highlighting the role of the university and research as driving forces for the global development of society.

Acknowledgement The author would like to thank Professor Andrea Bonaccorsi and Professor Alesia A. Zuccala for their useful comments on the first draft that greatly improved the manuscript.

References

ANVUR. (2013a). *Autovalutazione, valutazione e accreditamento del sistema universitario italiano*. www.anvur.org/attachments/article/26/documento_finale_28_01_13.pdf. Accessed 9 Sept 2016.

ANVUR. (2013b). *Valutazione della Qualità della Ricerca 2004–2010 (VQR 2004–2010): Rapporto finale ANVUR: Parte Prima: Statistiche e risultati di compendio*. http://www.anvur.org/rapporto/. Accessed 10 Sept 2016.

ANVUR. (2014). *Rapporto sullo stato del sistema universitario e della ricerca 2013*. http://www.anvur.org/attachments/article/644/Rapporto%20ANVUR%202013_UNIVERSITA%20e%20RICERCA_integrale.pdf. Accessed 10 Jul 2016.

ANVUR. (2015). *La valutazione della terza missione nelle Università e negli Enti di Ricerca: Manuale per la Valutazione.* http://www.anvur.org/attachments/article/26/Manuale%20valutazione%20terza~.pdf. Accessed 10 Sept 2016.

ANVUR. (2016). VQR 2011–2014: *Criteri per la valutazione delle attività di Terza Missione delle Università e degli Enti di Ricerca.* http://www.anvur.org/attachments/article/878/Documento%20criteri%20CETM_15~.pdf. Accessed 10 Sept 2016.

Basili, C. (2013). Oltre la comunicazione scientifica: riflessioni sul trasferimento della conoscenza nelle scienze umane. *AIB Studi, 53*(2), 111–117. https://doi.org/10.2426/aibstudi-9019.

Benneworth, P., & Jongbloed, B. W. (2010). Who matters to universities? A stakeholder perspective on humanities, arts and social sciences valorization. *Higher Education, 59*, 567–588.

Bozeman, B. (2000). Technology transfer and public policy: A review of research and theory. *Research Policy, 29*, 627–655.

British Academy. (2004). *"That full complement of riches": The contributions of the arts, humanities and social sciences to the nation's wealth*. London: British Academy.

Bush, V. (1945). *Science. The Endless Frontier.* United States Government Printing Office, Washington, DC. http://www.nsf.gov/od/lpa/nsf50/vbush1945.htm. Accessed 9 Sept 2016.

Donovan, C. (2008). The Australian research quality framework: A live experiment in capturing the social, economic, environmental, and cultural returns of publicly funded research. *New Directions for Evaluation, 118*, 47–60. https://doi.org/10.1002/ev.260.

Donovan, C. (2011). State of the art in assessing research impact: Introduction to a special issue. *Research Evaluation, 20*(3), 175–179. https://doi.org/10.3152/095820211X13118583635918.

Etzkowitz, H., & Leydesdorff, L. (1998). The endless transition: A "triple helix" of university-industry-government relations. *Minerva, 36*(3), 203–208.

European Commission. (2007). *Report of the Expert Group on Humanities: Positioning Humanities Research in the 7th Framework Programme.* Directorate-General for Research, Brussels. https://ec.europa.eu/research/social-sciences/pdf/policy_reviews/egh-report_en.pdf. Accessed 10 Sept 2016.

European Commission. (2014). *Horizon 2020 in brief: The EU Framework Programme for Research & Innovation.* ec.europa.eu/newsroom/horizon2020/document.cfm?action=display&doc_id=4752. Accessed 10 Sept 2016.

European Commission. (2015a). *Commitment and Coherence: essential ingredients for success in science and innovation: Ex-Post-Evaluation of the 7th EU Framework Programme* (2007–2013). http://ec.europa.eu/research/evaluations/pdf/fp7_final_evaluation_expert_group_report.pdf. Accessed 10 Sept 2016.

European Commission. (2015b). *Integration of social sciences and humanities in horizon 2020: Participants, budget and disciplines.* Publications Office of the European Union, Luxembourg. doi: https://doi.org/10.2777/382488. Accessed 10 Sept 2016.

Garbini, R. (2010). Scendere dalle spalle dei giganti. La parabola dell'istituzione universitaria in Italia. In C. Basili (Ed.), *Sinergie invisibili. Ricerca e informazione scientifica nell'Economia della conoscenza* (pp. 285–306). Rome: Consiglio nazionale delle ricerche.

Gascoigne, T., & Metcalfe, J. (2005). *Commercialisation of research activities in the humanities, arts and social sciences in Australia.* Department of Education, Science and Training, Canberra. http://www.chass.org.au/wp-content/uploads/2015/02/PAP20050501TG.pdf. Accessed 10 Sept 2016.

Geuna, A., & Muscio, A. (2009). The governance of university knowledge transfer: A critical review of the literature. *Minerva, 47*, 93–114. https://doi.org/10.1007/s11024-009-9118-2.

Gibbons, M., Limoges, C., Nowotny, H., et al. (1994). *The new production of knowledge: The dynamic of science and research in contemporary societies*. London: Sage.

Greco, P. (2007). La terza missione dell'Università. *AIM magazine, 62*, 50–52.

Greco, P., & Termini, S. (2007). *Contro il declino: una (modesta) proposta per un rilancio della competitività economica e dello sviluppo culturale dell'Italia*. Turin: Codice.

Hughes, A., Kitson, M., Probert, J., et al. (2011). *Hidden connections: Knowledge exchange between the arts and humanities and the private, public and third sector*. Swindon: Arts &

Humanities Research Council. http://www.cbr.cam.ac.uk/fileadmin/user_upload/centre-for-business-research/downloads/special-reports/specialreport-hiddenconnections.pdf. Accessed 10 Sept 2016.

Landry, R., Amara, N., & Lamari, M. (2001). Utilization of social science research knowledge in Canada. *Research Policy, 30*, 333–349.

Munro, E. (2016). Illuminating the practice of knowledge exchange as a 'pathway to impact' within an Arts and Humanities Research Council 'creative economy knowledge exchange' project. *Geoforum, 71*, 44–51.

Nowotny, H., Scott, P., & Gibbons, M. (2001). *Re-thinking science: Knowledge and the public in an age of uncertainty*. Cambridge: Polity Press.

Nussbaum, M. C. (2010). *Not for profit: Why democracy needs the humanities*. Princeton/Oxford: Princeton University Press.

Ochsner, M., Hug, S. E., & Daniel, H. D. (2016). Humanities scholars' conceptions of research quality. In M. Ochsner, S. E. Hug, & H. D. Daniel (Eds.), *Research assessment in the humanities: Towards criteria and procedures* (pp. 43–69). Cham: Springer. doi: 0.1007/978-3-319-29016-4.

Olmos-Peñuela, J., Castro-Martínez, E., & D'Este, P. (2014). Knowledge transfer activities in social sciences and humanities: Explaining the interactions of research groups with non-academic agents. *Research Policy, 43*, 696–706.

Olmos-Peñuela, J., Benneworth, P., & Castro-Martínez, E. (2015). Are sciences essential and humanities elective? Disentangling competing claims for humanities' research public value. *Arts & Humanities in Higher Education, 14*, 61–78. https://doi.org/10.1177/1474022214534081.

Paoloni, G. (2010). Il sistema della ricerca nell'Italia del secondo Novecento: riflessioni istituzionali e storico-politiche. In C. Basili (Ed.), *Sinergie invisibili: ricerca e informazione scientifica nell'economia della conoscenza* (pp. 233–262). Rome: Consiglio nazionale delle ricerche.

Penfield, T., Baker, M. J., Scoble, R., et al. (2014). Assessment, evaluations, and definitions of research impact: A review. *Research Evaluation, 23*, 21–32. https://doi.org/10.1093/reseval/rvt021.

Piazza, S., & Matteucci, M. C. (2015). Assessing the impact of research on society, economy and culture in the humanities and social sciences: A case study on a UK university campus. *RT: A Journal on Research Policy & Evaluation, 3*(1), 1–18. 10.13130/2282-5398/4318.

Predazzi, E. (2012). The third mission of the university. *Rendiconti Fisiche Acc Lincei, 23*(1), S17–S22. https://doi.org/10.1007/s12210-012-0182-4.

RAND Corporation. (2010). *Capturing research impacts: A review of international practice.* http://www.rand.org/pubs/documented_briefings/DB578.html. Accessed 10 Sept 2016.

REF2014. (2012). *Panel criteria and working methods.* http://www.ref.ac.uk/media/ref/content/pub/panelcriteriaandworkingmethods/01_12_1.pdf. Accessed 10 Sept 2016.

Research Information Network. (2011). *Reinventing research? Information practices in the humanities.* http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1859267. Accessed 10 Sept 2016.

Scamuzzi, S. (2015). Umanisti e scienziati: ancora due culture? In S. Scamuzzi & G. Tipaldo (Eds.), *Apriti scienza: Il presente e il futuro della comunicazione della scienza in Italia tra vincoli e nuove sfide* (pp. 235–252). Bologna: il Mulino.

Solimine, G. (2014). *Senza sapere: il costo dell'ignoranza in Italia*. Roma-Bari: Laterza.

Soriano, F. H., & Mulatero, F. (2010). Knowledge policy in the EU: From the Lisbon strategy to Europe 2020. *Journal of the Knowledge Economy, 4*, 289–302. https://doi.org/10.1007/s13132-010-0020-9.

Do SSH Researchers Have a Third Mission (And Should They Have)?

Brigida Blasi, Sandra Romagnosi, and Andrea Bonaccorsi

1 Introduction

Do researchers in Social Sciences and Humanities (SSH) have a third mission or is the third mission idea an instrument to subjugate researchers' academic freedom to meet the needs of the market, so it would be better that they do *not* try to demonstrate their impact?

The debate on the relationship between research assessment and the evaluation of the impact of SSH research is recent. Two factors triggered it: the discussion on the so-called *decline of Humanities* (mainly regarding government funding), and the decision, by the UK Research Excellence Framework, to include an explicit dimension of *impact* for all departments, including SSH, as a necessary component of the overall research assessment, with direct and large consequences on funding.

The impact of research is a longstanding issue in research policy (Davies et al. 2005; ESRC 2009; Donovan 2011; Bornmann 2013). There have been efforts to build up comprehensive impact assessment methodologies, particularly in the Dutch context, such as the ERIC methodology (Spaapen et al. 2007) and the SIAMPI approach (Spaapen et al. 2011; Wolf et al. 2013). Its prominence has increased after the REF exercise preparation and the publication of its impact results (Smith et al. 2011; Penfield et al. 2014; King's College and Digital Science 2015; Samuel and Derrick 2015; Chowdhury et al. 2016; Digital Science 2016). At the same time, the debate echoes a larger, and highly controversial, debate on the role of SSH in contemporary universities and society, and on the evolution of academic systems.

B. Blasi (✉) • S. Romagnosi
Agenzia Nazionale Valutazione del sistema Universitario e della Ricerca (ANVUR), Rome, Italy
e-mail: brigida.blasi@anvur.it; sandra.romagnosi@anvur.it

A. Bonaccorsi
DESTEC, University of Pisa, Pisa, Italy

IRVAPP-FBK, Trento, Italy
e-mail: a.bonaccorsi@gmail.com

© Springer International Publishing AG 2018

A. Bonaccorsi (ed.), *The Evaluation of Research in Social Sciences and Humanities*, https://doi.org/10.1007/978-3-319-68554-0_16

In this chapter, we review the main positions in the debate. Next, we illustrate the Italian path to evaluate the third mission of universities, including SSH, and we offer preliminary data on the activities carried out by SSH scholars. The Italian experience is peculiar, since the law has introduced a large scale assessment of the third mission, but it has been kept separated from the research assessment.

This chapter is a companion to Lanzillo's chapter in this volume which is devoted to methodological issues related to the Social Impact Assessment in SSH. This chapter takes an internal view, reporting on the Agency choices and the controversies surrounding them. Lanzillo's chapter takes an external and critical view, from the perspective of a social science scholar. Together, they offer an articulated view on an open issue.

2 Controversies over the Notion of Impact in Social Science and Humanities

The interest of the literature for the third mission has focused almost exclusively on Science, Technology, Engineering and Mathematics (STEM) disciplines and their potential to contribute to the economic exploitation of research. More precisely, the focus has traditionally been on commercialisation, following the successful model of research-patenting-licensing-spinoff companies that were created in the last quarter of the twentieth century in the USA (the so-called Silicon Valley model) (Molas-Gallart et al. 2002; Bonaccorsi 2016).

Some recent contributions have enlarged the definition of impact in several directions. It is now recognised that the impact that universities and, more generally, public funded research institutions have on society and economy is not limited to commercialisation. A useful concept suggested in the recent UK literature is the "knowledge exchange pathway", which suggests a multiplicity and variety of channels of interaction (Abreu et al. 2009). Hughes and Kitson (2012, 2013) classified the knowledge exchange pathways between universities and firms in four categories: *people-based activities* (attending conferences, participating in networks, giving invited lectures, sitting on advisory boards, student placement, standard setting forum, employee training, curriculum development, enterprise education), *problem-solving activities* (informal advice, joint research, joint publications, consultancy services, contract research, research consortia, hosting personnel, prototyping and testing, external secondment, setting of physical facilities), *commercialisation activities* (formed/run consultancy companies, licensed research, patenting, spinout companies), *community-based activities* (lectures for the community, schools project, community-based sports, public exhibitions). They showed that, contrary to the canonical Silicon Valley model, a small minority of people in universities were engaged in regular commercialisation activities (from 4 to 14% according to a survey of UK academics), while a much larger proportion was engaged in problem-solving and people-based activities.

Academics engage in a variety of activities, which are only partially covered by the university institutional umbrella, and registered in administrative records

(Perkmann et al. 2015). This is even more true for SSH researchers (Olmos-Penuela et al. 2013) who often engage into informal collaborations with external actors.

It is also recognised that university-business interactions are only part of the broader impact of universities in society. University-business interactions are certainly crucial to the transformation mechanism of turning knowledge into economic growth opportunities, as a consistent literature has shown in recent years (Hughes and Martin 2012; Perkmann et al. 2013; Bonaccorsi et al. 2013, 2014a, b; Haskell et al. 2014). Other knowledge exchange interactions take place outside the business sector. In summarising the evidence about the knowledge exchange between university and society in the UK, Hughes and Martin (2012, 2) notes that

> Outside the STEM subjects, only a relatively small proportion of academics have been involved in 'narrower' forms of commercialisation such as patenting, licensing or forming a spin-out company. In general, far more have engaged in wider or 'softer' people-based forms of interactions such as consultancy, attending meetings, giving talks, and helping with problem-solving.

This is consistent with the view of a study on knowledge exchange in Arts and Humanities in the UK (Hughes et al. 2011, 1):

> Once we move beyond that narrow perspective [of technology transfer through patents, licenses and spin-offs] on the role of impact to include mechanisms which include people-based, problem-solving and community orientated activities, the Arts and Humanities display as rich and diverse a set of connections as other disciplines and a particularly rich set of the third sector and community interactions.

A more recent study of 4059 UK-based academics has compared their engagement between two surveys, in 2008–2009 and 2015 (Lawson et al. 2016). An interesting finding from the study is that among the identified pathways, Public Engagement (including three activities: Community Lectures, Public exhibitions, and Projects with schools), increased during the period. Another interesting result is that academicians with a higher propensity to be active in Public Engagement declare that their primary motivation is to improve the university outreach as part of their mission.

It must be recognised at the beginning that we are dealing with a substantively different problem from the well-known valorisation or research commercialisation issue. It must be acknowledged that the impact of research on SSH on society has received much less interest and visibility than similar STEM studies. One obvious reason is measurability: research in STEM takes shape in several well-identified and measurable objects (indexed articles, patents, spin-off companies, licensing contracts), while research in SSH has an impact which is mediated by people and takes a variety of social forms which are often poorly measurable. Measurability should not obscure the substance of things (Gingras 2014; Wilsdon et al. 2015). One important negative consequence of the focus on more measurable activities is a distortion in universities' behaviour. As Molas-Gallart and Castro-Martinez (2007) already noted, there is significant ambiguity in the notion of the third mission, which requires a continuous process of clarification and discussion with academic communities. If only easily measurable results are targeted, there might be unintended

adverse outcomes regarding behavioural biases (Olmos-Penuela et al. 2015; Rosli and Rossi 2016).

It must be also recognised that SSH communities have been much less engaged in demonstrating a commitment to impact. There is a heated debate surrounding the notion of the impact of universities in non-STEM disciplines and non-commercial pathways. Several arguments have been put forward in the international debate on the controversial concept of "impact" of SSH. While there is a distinction between impact and the third mission (to which we return later in this chapter), the impact debate is crucial to articulate the issue of the appropriate assessment of the third mission.

We discuss below the spectrum of intellectual positions on the issue of the impact of SSH research and whether a notion of the third mission makes sense in these fields.

2.1 Threat to Academic Freedom

This thesis is probably the most diffused in the SSH literature that has dealt with research evaluation. A full range of positions are found here, depending on the theoretical foundations, but the convergence on the argument that academic freedom of researchers is at risk is substantial.

According to an influential line of thinking, inspired by Michel Foucault (1978), evaluation is a device to control academicians' behaviour, by using apparently objective measures that are ultimately accepted as if they were "natural objects". Given the supposedly objective nature of numbers, researchers become slaves to a system which has an overall aim of limiting the freedom of thinking rather than increasing accountability on public resources. During the French debate about research evaluation, the argument that evaluation was a form of "voluntary slavery" was often raised (Abelhauser et al. 2011; Gori 2011, 2013; Del Rey 2013). This line of thinking uses the notion of *governmentality* introduced by Foucault in a series of celebrated lectures (Dean 2010). This idea implies that governments in modern societies no longer control the behaviour of citizens using explicit orders (which, given the diffusion of culture and autonomy, would be ineffective). Instead, governments use an articulated system of technical devices, such as classifications, standards, labels, social statistics, which are legitimised by their scientific nature. The French word used by Foucault for this device was *dispositif*.

According to Supiot (2015), the numbers adopted during an evaluation have a power which is similar to law, but not justified in democratic terms. He draws an interesting analogy between a judge and an evaluator. The judge examines a concrete case and subsumes it into a general legal discipline. He deliberates on the case and makes a public deliberation. The subjects can appeal and offer counterarguments. An evaluator also examines a concrete case, such as an article or book. He subsumes it into a set of research quality criteria and gives a score. This score, contrary to the judge's sentence, is not justified in detail. The apparent objectivity of

numbers makes the evaluation untouchable. While the judge and the evaluator both follow a stringent logic of subsumption of the particular under the general, the evaluator escapes rational justification, protected by the opacity of numbers.

Another line of thinking uses the sociological theory of fields, proposed by Pierre Bourdieu (1984). According to Bourdieu, social actors compete in the social arena by leveraging their capital. The dynamics of social competition creates a "field" (*champ*, in French), or a social structure which is not defined by institutional boundaries but by struggle and power relations among the actors. The academic world is no exception. Here researchers strive to build the legitimation of their research interests and enlarge their impact on society. One of the primary resources they control is the power to give credentials, in particular, to certify the content of university degrees for students. According to this line of thinking, in a modern society in which higher education degrees grant access to the upper echelons of the job market and to move up the social stratification, the joint control of education and research gives academicians more power. Consequently, evaluation is nothing objective. It is a form of academic struggle, in which the winners try to impose a discipline on the losers. The task of social sciences is to reveal the academic battle's hidden dimension.

These arguments, mostly adopted in continental Europe, have a certain parallel in the American debate on academic freedom. Here the situation is institutionally different: academic freedom is not constitutionally protected but embedded into the practices of universities. While in most European countries the State manages the academic career, in the United States it follows the decentralised practices of universities. Since the 1940s, there has been a process of formalisation and institutionalisation of practices adopted by all universities in hiring and retaining academic staff with full respect for their personal beliefs, across all dimensions, and their freedom to choose research interests. The relationship between academic freedom and SSH research is problematic, due to the structural movement of the US academic system towards non-tenured positions. It is well documented that the share of tenured positions out of the total academic staff decreased significantly in the last part of twentieth century. This is true not only in those fields as in Life science where staff had been burgeoning but also in Humanities. In these disciplines, a steady flow of PhD graduates has created an abundant supply of temporary researchers and undergraduate teachers. According to some authors, the lack of protection of untenured researchers may limit academic freedom, particularly the researchers' ability to maintain critical lines of research (Bousquet et al. 2004). In this direction, the emphasis on the demonstration of impact by Humanities researchers is interpreted as a pressure to accept market-like, or instrumental values, while sacrificing the essential non-instrumental nature of research and culture in Humanities.

These lines of thinking converge on the argument that universities risk losing their autonomy and being subject to the domination of market forces (Bollier 2002; Bailey and Freedman 2011; Collini 2012). This is even more so for the Humanities since they are more remote from the market than STEM (Citton 2010; Cassin 2014; Bérubé and Ruth 2015).

Does evaluation really limit academic freedom? The answer is not evident. We are not aware of any large-scale demonstration that evaluation is used to punish individual

researchers based on their personal beliefs or lines of research. It is possible that such use might be difficult to be documented in the scientific literature, or even in published sources. What is sufficiently clear, is that there should be a clear distinction between the freedom to conduct academic research and the duty to contribute to society.

2.2 Non-instrumental Value of Research in Humanities

Another argument, most prominently suggested by Martha Nussbaum (2012), is that research in Humanities preserves its value only insofar as departs from the practical values that other fields of study may accept. According to this argument, research in Humanities has a time horizon which encompasses the whole of human history. Scholars in these fields make use of texts, or pieces of art, dating back centuries as if they were contemporary. The main aim of the research is not instrumental but truth oriented.

At the same time, research in Humanities is a necessary ingredient in the creation and formation of critical thinking. Critical thinking is only possible if people are educated to appreciate the historical depth of events and phenomena and to retrieve antecedents of the current issues and opinions from the cultural past. Research in Humanities teaches people to examine arguments, compare them, see connections, make a well-grounded decision and to communicate effectively in society. In summary, Humanities are necessary for democracy.

This powerful argument is at the root of the request that research in Humanities is treated differently from other fields, by respecting the longer time horizon and the critical value of its achievements. Some authors have criticised this argument. They claim that there is no substantive difference between being instrumental for the market or being democratically conducive. What is at stake is the intrinsic non-instrumentality of research in Humanities. A defence based on the subordination to democratic values, while remarkable *per se*, is not enough.

This debate suggests a crucial point for the evaluation of impact: the time horizon by which it should be defined and possibly measured is a critical variable. Asking research in Humanities to show short-term impact is intrinsically wrong. If the main result is found in the educational process, it deploys the effects over generations, rather than in the short-term. As recently argued by Molas-Gallart (2015), there is a need for an evaluation model which respects the peculiar nature of SSH research.

2.3 Bureaucratisation

Advocates of this position do not negate the importance of accountability but question the process by which it can be achieved considering administrative burden, cost, and bureaucratic mentality. Ben Martin formulated the most prominent position by describing impact assessment as a Frankenstein monster (Martin 2011). This

metaphor means that it is a cumbersome process, demanding a significant amount of documentation and data collection. It is also a process that may spiral out of control and have implications that go beyond the regulators' intentions. Pressuring researchers to demonstrate impact may create distortions in the allocation of time among the three areas of research, teaching, and community service, often at the expense of educational activities.

From an empirical point of view, it is true that the time budget allocated to bureaucratic activities has increased for academic staff worldwide (Bess 1988; Bowen and Tobin 2015). Several surveys have documented an alarming growth trend in the number of hours allocated to activities such as completing forms, writing documents, drafting proposals, and running the meetings. It is unclear, however, to what extent this increase in bureaucratic burden is due to a need to meet accountability requirements, or to other factors. Among other factors that contribute to the load, a prominent role is played by the increase in the share of competitive research grants. Competitive grants require a full-scale process of writing proposals *ex-ante,* giving an accurate accounting and *ex-post* content-related documentation. In many countries, a sharp decrease in the success rates of research proposals has been documented, which translates into a higher level of competition and a need for more professional draft proposals.

Another source of bureaucratic burden comes from the collegiate tradition. Academic staff have a longstanding tradition of collective decision-making. It has been shown by the venerable tradition of organisation theory that academic decision-making is shaped by a sense of belonging to a community, and the need to carefully control the behaviour and decisions of colleagues in other disciplines. This requires various forms of group or collective decision-making, with a limited degree of delegation of decisions.

Whatever the underlying factors, academic staff perceive an increase in their burdens and report emotional discomfort with the situation. While it is difficult to blame competitive research funding or collegiate decision-making (at least in public), it is easy to blame the evaluation process. The main argument is that evaluation is an external constraint, something that adds complexity and unnecessary reporting needs for all researchers.

2.4 Road to Relevance

Some learned societies and scientific academies have embarked on a definition process and measured the impact of SSH research, accepting the accountability challenge to provide value. This process went ahead in the UK, following institutional pressure from the government to demonstrate impact.[1] In 2008 the British National

[1] According to Hughes and Martin "It has been estimated that between 2000 and 2010 the UK research councils produced 96 documents dealing with impact, of which 35 were impact studies and 19 internal evaluation reports (de Campos 2010, Table 3)" (Hughes and Martin 2012, 1, footnote 2).

Endowment for Science Technology and the Arts (NESTA) and the Arts and Humanities Research Council (AHRC) published an influential report which has demonstrated that the view that innovation is only rooted in scientific and technological knowledge is misleading (Bakshi et al. 2008). Reeves (2007) carried out a survey of methods to evaluate the Arts impact. Despite these efforts, it was well known that the UK government adopted a selective approach to research funding, by asking all scientific fields to be explicit on the impact of research and cut SSH funding but not STEM research. This institutional shock (Martin 2012) created an environment where some actors accepted the challenge and tried to demonstrate the impact. As the Council of University College in London (2011) wrote

> Today, in an era when the arts, humanities and social sciences are perceived as being under threat from the Government's funding changes, UCL remains committed to maintaining and investing in them. We need to counter a trend towards instrumentalist attitudes towards higher education in the new funding environment.

A striking example of this trend is the Arts Council of England series of reports on the economic impact of Arts. The most recent edition (Centre for Economic and Business Research 2015) stated that

> Overall, the arts and culture sector is more than paying its way in the returns it brings to the Treasury. Every pound of public funding going to the Arts Council's national portfolio organisations pays back £5 in tax contributions from the sector as a whole. The most recent figures show an annual return of £2.35 billion to the Treasury.

A similar trend is visible at a European level. Here there is not an institutional pressure to show impact, but the overall narrative on societal challenges which is an overarching framework for the Horizon 2020 initiative, implicitly goes in that direction. As a response, several stakeholders in the SSH field have started to build up articulated arguments to show the impact of research.

Science Europe, the leading association of funding agencies and Public Research Organizations, convened a Committee on Humanities which published several documents on the implications of societal challenges (Science Europe 2013). In another report, Science Europe (2015, 15) it was argued that

> Europe is now being undertaken by other nations who seek to transform their cultures of reproduction through innovation-driven progress. They are doing this by bringing the humanities to the centre of technological and scientific developments. At the same time Europe's policies and funding programmes, in particular, Horizon 2020, see to be relegating the humanities to the research landscape peripheries.

Similarly, the League of European Research Universities (LERU) has published a collection of position papers on the potential impact of SSH research on each of the challenges identified under the H2020 initiative. In 2012 LERU issued a position paper where it was argued that the European Commission proposal for the Horizon 2020 programme (European Commission 2011) and the plan's articulation regarding Societal challenges was welcome (League of European Research Universities 2012). LERU warned against a narrow definition of Societal challenges,

giving relevance only (or chiefly) to scientific and technological research. In 2013, it addressed each of the challenges by showing the crucial role of SSH research (League of European Research Universities 2013a, b). Taking an example, based on widely-cited scientific research on climate change (Cash et al. 2003; Adger et al. 2012), it argued that without state-of-the-art research on behavioural aspects of energy use and consumption there was no way to manage the sustainability transition.

Several authors have followed the same reasoning and empirically demonstrated the importance of cultural activities, primarily based on Humanities originated content, for several dimensions of social and economic impact. These areas included the psychological well-being of people (Grossi et al. 2011), competitive production of symbolic meanings of companies (Dell'Era 2010) and urban renewal and growth (Florida 2002). Buekers and Nugteren (2012) proposed that a link between research in Humanities and the scientific and technological fields can be identified in the role of creativity.

An entirely new branch of policy interest has been created on top of the concept of the "creative economy" (UNDP-UNCTAD 2010). The creative economy is defined as an "emerging concept dealing with the interface between creativity, culture, economics and technology in a contemporary world dominated by images, sounds, texts and symbols". (ib.). Another stream of policy interests was generated around the notion of "creative cities" (UNESCO 2016).

These initiatives are valuable. They increase the visibility of SSH and offer arguments which can be used in the dialogue with research funding agencies and decision makers. However, it is not obvious that they solve the problem of the funding devoted to Humanities and to the overall higher education sector. An objection to the above argument is that the instrumental value of Humanities does not come from research but from education, or more loosely, culture. In a UK study of cultural organisations, Hughes et al. (2014) noted that the disciplines that are most involved in the collaboration with the private sector business are Creative and Performing Arts, Humanities, and Education. The most used forms of interaction were taking part in networks, providing expert and informal advice, and supervising student projects. In fact, most of the activities currently examined under the heading of "culture"[2] certainly had a relationship with *research* in Humanities, but an indirect and lagged one. Why should governments fund high-level research in challenging issues for Humanities, to obtain short-term results, say, for museums or cultural heritage? This is not an easy question to answer.

[2] As an example, the cultural activities which, according to Grossi et al. (2011) have a positive impact on psychological well-being of people are as follows: Jazz music concerts; Classical music concerts; Opera/ ballet; Theatre; Museums; Rock concerts; Disco dance; Paintings exhibitions; Social activity; Watching sport; Sport practice; Book reading; Poetry reading; Cinema; Local community development. In turn, creative cities under the UNESCO initiative deal with: Crafts and Folk Art; Design; Film: Gastronomy; Literature; Music; Media Arts (UNESCO 2016).

2.5 Relevance and Intrinsic Value

According to this argument, asking SSH to demonstrate its impact on society is not only legitimate but also enhances the status of research in these fields. What is needed, is an evaluation process which respects the peculiarities of SSH research and adopts a sophisticated impact model. This is the road taken by the Italian experience. The conceptual underpinnings of this approach are as follows.

First, there must be an explicit recognition of differences in various forms of knowledge production activities. Several recent contributions have started to examine the epistemic value of research in Humanities, with the goal of showing its intrinsic scientific method (Bate 2011; Belfiore and Upchurch 2013; Bod 2013; Brooks 2014). The value of research in SSH does not provide short-term results but has an essential role in the intelligibility and reflexivity of human society. This role goes much beyond any specific or immediate spillover to society, or any role in maintaining other socially valuable activities, such as cultural activities or cultural heritage. The essential reason for governments to fund SSH research is that it is a condition to maintain countries's or people's historical continuity. Research in Humanities supports and augments the ability, at any given point in time, to interpret the meanings of the symbolic reality by which humankind has lived since its origin. If research in Humanities was interrupted, the ability to understand the meanings of languages, texts, images, artworks, or material artefacts, would be irreversibly damaged. Research in Social Science is essential to maintain the intelligibility of social linkages. Eventually, if research in Humanities were interrupted, society would collapse as a distinct, human enterprise.

Second, research in SSH has a long-term, indirect, profound and hard-to-measure impact on the advancement of society and economy. Any attempt to measure the direct impact of research is bound to fail because time scales do not overlap. There is no point in trying to show that Humanities research has the same short-term impact of research in Technology and Engineering. It is sufficient to put your hand into your pocket to find the evidence of technological achievements in the form of your smartphone. Even if it is true, it is much harder to demonstrate that the smartphone would not work without an accumulation of research in linguistics, semiotics or philosophy of language.

Third, and consequently, there must be a distinction between individual academic freedom and societal impact. Individual researchers in SSH must be left free to identify the most promising research directions. It is pointless to ask individual researchers to demonstrate the effect, because it may manifest in a varying time scale, take place in different locations and sectors, or even remain unknown. This distinction can be seen in the following statement: individual researchers are personally responsible (and must be accountable) for research and teaching; the impact on society is instead a collective outcome, for which institutions must be responsible. Individual researchers are engaged in activities that magnify, accelerate and make more effective, the impact of research on society. The accountability for this

impact is on the side of institutions because they have the time horizon and the scope to encompass the large variety of forms of impact.

There is a need for methodological developments on the impact evaluation of activities at the university level. Most existing methods deal mainly with the impact at the programme level (Molas-Gallart et al. 2000; Molas-Gallart and Tang 2011; Morton 2015).

Fourth, it is expected that involvement in third mission activities, particularly those involved in Public Engagement, is not detrimental to the other two missions of universities – research and teaching. While some level of tradeoff can be identified, overall it is a synergy and complementarity relationship (Hughes et al. 2010).

Finally, under these assumptions, it may be useful to introduce the notion of the third mission for scholars in SSH. This is the road taken by the Italian Agency. The next sections illustrate the Italian experience, as a first step in the direction of a comprehensive assessment of the contribution of SSH to society.

3 The Institutional Path of the Evaluation of Third Mission

3.1 The Preparation Process

As for many other aspects of the life of universities and public research in Italy, legislative or administrative initiatives trigger changes. In the case of the third mission, there have been three oscillating positions in Ministry of University and Research's actions:

(a) considering the third mission as part of the periodical research evaluation exercise (VQR 2004–2010), but limiting it to a restricted set of indicators (Ministerial decree launching the research assessment 2011);
(b) including the third mission as part of the annual self-evaluation of departments, together with teaching and research (Ministerial decree launching the accreditation, evaluation, and self-evaluation system 2012);
(c) re-establishing the evaluation of the third mission within the framework of the periodical research evaluation exercise (VQR 2011–2014) but without an impact on the performance-based funding formula (Ministerial decree launching the second research assessment 2015).

Faced with this oscillating administrative framework, the Agency made a bold decision to develop a robust and consistent framework for concepts, evaluation criteria, indicators, and assessment methodology, to be implemented as soon as the administrative framework would make it compelling. In other words, the preparatory process was started well in advance with respect to the administrative deadlines and was flexible to the changes in the ministerial orientations. Therefore, the framework was initially created under the provision of point (b) above but eventually shifted into the second research assessment exercise (VQR 2011–2014) outlined in

point (c). It is currently pursued as part of the research assessment exercise, but with separate panels and a dedicated methodology (see below).

The preparation of the evaluation framework followed a participatory approach. A first workshop was held in April 2013 in Rome and involved vice-rectors, university administrative staff and PROs. The workshop generated high expectations for the implementation of the evaluation. A major decision that came from the workshop was that it was important to ensure that the third mission conceptual definition should be comprehensive and not limited to traditional indicators of industrial valorisation and commercialisation of research. After this event, an Expert group was created to review the state-of-the-art of the international debate and outline a Manual for the evaluation of the third mission. The group included some of the most reputed academic scholars in social sciences with experience in the qualitative and quantitative analysis of the third mission.

After a year of work, the Expert group released some methodological recommendations to the Agency. These recommendations were formally approved in 2014, and the group was confirmed for a second year with a mandate to draft the Manual. The Agency adopted the draft of the Manual in 2015 and submitted it to a dual process of consultation. Firstly, it was made available on the Agency website with an invitation to submit comments and recommendations. More than a hundred suggestions were received. Approximately half of the recommendations were accepted and led to modifications of the draft. A detailed document addressing all comments and explaining the acceptance or rejection of the suggestions was published on the website. Secondly, it was presented in another workshop in Rome, this time with the contribution of international scholars that gave comments on methodological and substantive issues. After these events, the Agency Board formally approved the Manual. The next steps implemented the evaluation methodology suggested by the Manual. A call for experts was issued in 2015, generating about 200 candidates. Out of these, a panel of 28 experts was created and started work. The evaluation report was delivered in 2017.

3.2 The Main Choices

There have been some substantive choices, which make the evaluation of the third mission in Italy an interesting case study in the international context. We discuss them in turn.

3.2.1 Definition Comprehensiveness

The first fundamental choice was to enlarge the third mission's definition and to institutionalise a conceptual shift, from limiting the perimeter to the industrial valorisation and commercialisation of research to include all the dimensions of universities' and PRO's impact. Note that this choice was *not* strictly mandated by the legislation or the Ministerial decrees, which explicitly mentioned only a few

traditional third mission indicators such as patents and spin-off companies. The Agency interpreted the legislative and administrative texts as making these indicators mandatory but did not limit to them the definition of the third mission.

The Manual formalised this conceptual shift by defining the third mission as being based on two separate dimensions. It used a precise and formal language derived from economic theory to make the distinction sharp and well-grounded. The first dimension is the *valorisation of research* or the transformation of research into economic value. This aspect by definition involves the transformation of goods supported by public funding (public research) into private goods (Intellectual Property Rights, spin-off companies, contract research). The rationale for this dimension is that in market economies the transformation of publicly funded research into marketable goods can valorise the full potential of knowledge, generating economic opportunities and employment. Managing the boundaries between public knowledge and private appropriation of opportunities is the challenge that all universities and PROs have been asked to address in the last decades.

From an operational point of view, this dimension is articulated into four areas of evaluation:

1. IPRs Management
2. Spinoff companies
3. Contract research
4. Collaboration with territorial intermediaries

The second dimension is the *production of public goods*. This aspect considers that universities and PROs produce knowledge spillovers that are not appropriated by economic actors but are made available to the whole of society. These knowledge spillovers have a different time scale than those that are the object of commercial valorisation, they are more diffused, less visible, more heterogeneous, but nevertheless have the same importance than those that are traditionally considered part of the third mission. To make this importance symbolically visible, the second dimension has been operationalised with the same number of areas of evaluation than the first. From an operational point of view, the public-good aspect is also organised into four areas of evaluation:

1. Production and maintenance of cultural heritage (museums, archaeological excavations, historical buildings)
2. Contribution to public health (clinical experimentation, biobanks, lifelong training in Medicine)
3. Lifelong Learning (adult and permanent education)
4. Public Engagement.

These areas do not exhaust the full range of the third mission activities geared towards society but cover most of them. The area of Public Engagement received a comprehensive definition, one that has allowed departments and universities to submit an enormous diversity of contributions.

The formal definition of these two areas has produced some significant consequences. First, they have been the object of separate chapters of the final Report.

This will make the richness of the third mission institutionally visible and overcome the narrow definition which is still so prevalent in the literature and in practice. Second, the expert panel has been divided into two sub-panels of the same size. This has ensured that adequate effort has been placed on evaluating not only those areas where there is more consolidated experience, such as patents or spin-off companies but all areas of impact.

3.2.2 Informed Peer Review and Maturity Model

The assessment method chosen was a mix between indicator-based evaluation and peer review. A collection of indicators started in 2014 and was replicated in 2015 and 2016. However, the evaluation is not indicator-driven but is left to the peer review of a panel of 28 experts, recruited after a competitive CV submission. This is another original contribution to the international debate and practice. It has been considered that the maturity of the field does not permit the operationalisation, measurement and standardisation of all the relevant indicators. Only in a few areas (namely, patents, spin-off companies and contract research) there is a sufficiently large literature on conceptual, methodological and practical issues, that can be used to design a system of indicators. It was strongly felt that pushing the evaluation of the third mission in the direction or (prematurely) standardising the scales, to produce a unique final score, would have been a major methodological and substantive mistake. It would have distorted the behaviours of universities and PROs and generated counter-reactions. On the contrary, a "maturity model", to borrow a term from the computer science field, was suggested. According to this model, experts in each of the (4 + 4) areas of the third mission adopt the metrics that they consider most appropriate given the status of the international debate and the quality of existing indicators.

The adoption of a maturity model has an important consequence: it is possible that the areas of the third mission will be evaluated with different methodologies and will produce non-comparable scales and scoring outcomes. In other words, in those areas in which the maturity of indicators is high, the international methodological reflection is advanced, and the assessment practice is rooted, it will be possible to use cardinal scales, to aggregate the sub-items and possibly to produce rankings (if useful). However, in other areas, the conceptual discussion on the content of the third mission is still problematic and the level of standardisation of practices is low. In these cases, the evaluation will not be based on cardinal scales, but will most likely use qualitatively-assigned ordinal values, perhaps with a coarse level of granularity. Consequently, it will be impossible to aggregate all evaluations into a unique score at a department or university level. It is well-known in the theory of evaluation that the aggregation of ordinal measures is inappropriate if the underlying scales are inconsistent. This means that there will not be a national ranking of the third mission, but a collection of evaluations that reflects the heterogeneity of institutions' strategic choices.

This is a significant departure from the preferred methodology adopted by the Agency in the assessment of research. In the 2004–2010 VQR, all scores were expressed as a unique discrete ordinal scale, irrespective of the underlying assessment methodology (either peer review or bibliometric indicators). By normalising the indicators of departments and universities against the national average (to eliminate size effects), it was possible to generate a unique ranking in each of the disciplines. By aggregating these rankings at the level of universities, weighting disciplines by the respective number of staff led to a unique national ranking of universities. The legislation mandated the rankings' production and aimed at informing the Ministry of Universities and Research of the allocation of funding using a performance-based formula. Nevertheless, the production of rankings had a major and prolonged impact on the media, raising the visibility of the research assessment exercise much beyond the boundaries of the interested audiences (Blasi et al. 2017a, b). With the adoption of the informed peer review approach, coupled with a maturity model, there will not be such a ranking. The expected sacrifice in visibility, however, was more than compensated by the correctness of the methodological choices.

3.2.3 Autonomy of the Evaluation of the Third Mission

Although the third mission evaluation is part of the research assessment exercise (VQR), it does not contribute to the final score of departments and universities. Instead, it is carried out with a dedicated methodology and does not produce a score which is included in the final VQR score. Consequently, the evaluation of the third mission will *not* enter into the performance-based formula funding for universities, which allocates up to 30% of government block granting. Only the research assessment scores will count.

The Agency anticipated this orientation after the 2004–2010 exercise, in which a small number of indicators of the third mission were experimentally introduced. The analysis of the resulting data suggested that the maturity of the indicators were not high enough to form the basis for funding. The group of experts then confirmed this orientation, but with a different motivation. The experts asked that any performance-based funding aimed at taking stock of the evaluation of the third mission should be by a construction *additional* to the existing government resources. In other words, the experts strongly advised against the risk that the third mission could be seen as competing with research and teaching, rather than as a complementary and strategic action. Based on these recommendations, the Manual states that the third mission assessment was inappropriate as a funding basis.

Even more precisely, the Manual indicated that, while research and teaching are formal obligations of all staff in universities, the third mission is not an obligation for all individuals, it should be rather considered an *institutional* mission. This meant that institutions were to be held accountable and *not* individuals. In other words, individuals enjoy academic freedom, which is protected (as in other European countries after the totalitarian experience of the twentieth century) by the

Constitution. Academic freedom means that they should not be considered responsible for the use of the knowledge they produce.

This choice contrasted with the choice of the UK Research Excellence Framework, in which the impact of research (in all disciplines) contributed to the final score, entered into the formula funding and received a weight of 20% (it will be 25% in the next exercise). It also contrasts with the orientation of the Dutch Research Assessment Protocol, in which the impact of research is explicitly included.

It was another contribution of the Italian experience to the international debate. Why did the Italian Agency make this choice? The UK and Netherlands have a long experience in research assessment, started in the 1980s. Their academic communities have absorbed the research quality criteria. Bibliometric indicators and peer review are (almost) universally accepted. In these systems, the notion that the impact of research may become an intrinsic part of its quality has been suggested as an evolutionary process, improving, and refining a commonly accepted definition of quality. In the Italian case, the experience with research assessment is less mature. The VQR 2004–2010 was the first large-scale exercise, covering more than 180,000 products. In early 2000 another exercise, called VTR (*Valutazione Triennale della Ricerca*) was launched, but was based on a small sample of products (approx. 18,000, an order of magnitude less) and did not involve all researchers. Also, the VTR did not have the funding formula as an explicit goal. Summing up, the acceptance of the research quality criteria in the academic community could not be taken for granted. Including the notion of impact within the first large-scale research, would have been detrimental to the assessment.

There is a second argument. In the REF formula, the three most critical dimensions (impact, environment, and quality) enter the final formula with a separate percentage weight – since the formula is additive, the overall measure is compensatory. A poor research quality performance might be compensated by an excellent performance in research impact and vice versa. This was a non-trivial choice. It meant that it is possible to "trade" impact with quality, that is, to define an exchange rate at which one can give up quality in return for impact. In some sense, this approach seems to consider inevitable that quality and impact are, to a certain extent, a substitute among themselves. If on the contrary, they could be regarded as entirely complementary and positively correlated, then the inclusion of impact would be redundant. Given the attention of the UK government to the issue of value from public expenditure, it is unlikely that such an orientation would have been dominant. In the Italian context, the Agency has chosen to make the two dimensions (quality and third mission/impact) the object of two separate and non-comparable exercises. This prevents weaker institutions from claiming that, although they do not perform well in research quality, they nonetheless have a valuable impact on society. This argument has repeatedly been made in the public debate after the first VQR, particularly for a few universities and PROs that performed poorly on quality. This argument reduces the institutions' accountability. In a system in which the research assessment still needs complete acceptance, there was a risk in compromising pure research quality factors. The separation between the two dimensions and

the exclusive role of research quality on university funding create enormous pressure on institutions.

Finally, there is an important epistemic argument, one that is not made explicit in official documents but may be reconstructed from the evaluation practice. The argument is that research quality cannot be defined by criteria *external* to the scientific communities. The pressure for accountability of universities, which is inevitable and positive in democratic societies, should not be addressed by giving up the fundamental principles of the autonomy of the scientific sub-system and the self-regulation of the scientific community. Accountability and independence should be kept together and not compromised.

This principle is preserved in the Italian experience by the notion that accountability on the third mission is placed at the door of institutions, not of individuals. A researcher in Ancient Greek should not be kept individually responsible for the impact of his research in society, because this result is highly unpredictable, lagged in time, displaced in space, and often unobservable. Placing pressure on researchers for an impact of each piece of their research is not useful and is often devastating. This is even more so in SSH, particularly in Arts and Humanities. This does not preclude the institutions (departments and universities, and PROs) from defining an institutional strategy to increase and improve the visibility of the impact on society and economy, using voluntary interests and capabilities of researchers, providing the appropriate institutional or career incentives, and managing the inevitable tensions and trade-offs.

3.2.4 Reflexivity

The evaluation is not a neutral process, exclusively based on technically necessary procedures. It is a social process, in which the measurement interacts with the measured reality in a highly nonlinear way. Those actors that are subject to the evaluation adapt their strategies to maximise (or at least, to increase) their standing among others and in the competition for resources. The adaptation may be functional (i.e. in the direction predicted and desired by the evaluation) but may also be dysfunctional (i.e. in the wrong direction, or in no direction at all).

If a different conceptual framework is used, the evaluation can be treated like a game played by the institution against the actors. The actors receive a payoff from their behaviour according to predefined rules and will engage in the strategic games to beat the rules. The extent to which strategic games can be played depends on the nature of the rules.

This is even more so in the case of indicators. Once they are created, they are used in the social system as if they had internal legitimation and without any consideration of their limits and validity. As it is often said, indicators “take on a life of their own”. In the philosophical language of applied ontology, they become social realities. The assumption that people use indicators with the attached instruction sheet is unwarranted. The instruction sheet is rapidly lost, and the indicators circulate without restrictions.

In the case of the third mission, this means that there is the risk that institutions target specific indicators as the goal of the activity, rather than as approximations to be used in a wider informed peer review framework. With this concern in mind, the Manual developed some research evaluation criteria, aimed at investigating to what extent this was an issue.

For example, in the field of IPRs, what should be evaluated is not the number of patents *per se*, but the strategic role of patents in the relationship between a university and its industry partners, including the spin-off companies. This is much more complicated to evaluate. On the contrary, focusing on the number of patents, or even on the revenues derived from licensing (a better indicator, but still too limited in scope), would induce a perverse incentive to increase the patenting activity, without controlling the quality of patents and the cost effectiveness of patent protection. This is an unintended and adverse outcome since patenting is a strategic action in which the total number of patents is not necessarily the most important factor.

In the case of SSH, a critical issue was the definition of the unit of analysis. As it will be shown, Public Engagement initiatives from SSH take a tremendous diversity of forms and extend over time in substantially variable ways. For example, the public engagement of researchers into scientific controversies may have a peak of events in a short period (meetings, press releases, media coverage, TV, parliamentary hearings), but it depends on long-term interaction and trust building between researchers and stakeholders. If for the sake of standardisation of indicators, the evaluation systems place emphasis on discrete, individual events, it may distort the universities' behaviour. They will "slice" events into ever smaller countable units, and the overall picture of the interaction between research and society will be lost. Also, the comparability across institutions will be compromised, because they will choose the unit of analysis that better serves the strategic goal of "gaming" with the evaluation.

Inspired by the contributions of social sciences on evaluation, the Manual takes care to avoid unintended consequences in distorting universities and PROs' behaviour.

3.2.5 Evolutionary Nature of the Evaluation Process

The Manual explicitly states that the evaluation of the third mission is a dynamic and evolutionary process and that new areas of evaluation can be considered in the future, in addition to the eight (4 + 4) areas implemented in the current exercise. This means that the indicators used for evaluation can be modified after the initial exercise. This contrasts with the theory of evaluation (particularly the realist approach to evaluation), which recommends that the framework of evaluation be stable.

The contradiction is only apparent. The third mission is intrinsically dynamic because it depends on the interaction with society and economy and its variable needs. There is a significant difference between the evaluation of research and

teaching and the evaluation of the third mission. In the case of research, the fundamental rules of scientific communication are relatively stable over time. In the STEM disciplines the publication process, role of scientific journals, peer review, conference organisation, format of scientific contributions, were all created more than three centuries ago. They were diffused in advanced societies between the 19th and 20th centuries and measured and standardised in the last third of the twentieth century with bibliometrics and scientometrics. The meaning of the count of journal articles and citations, and the possible indicators that have been built on top of these is relatively stable and accepted. This is not the case, as it is well known, in SSH. This does not mean that any research quality criteria are absent – simply that their measurability and standardisation level is lower. In research assessment, it is highly recommended that evaluation criteria and indicators are stable over time, to induce an adjustment process among institutions and researchers.

When it comes to teaching, the situation is less clear, due to the complex interaction between teaching effectiveness and *ex-ante* selection of students. Existing methodologies place great emphasis on the *process* of design, planning, implementation and monitoring of the higher education cycle, rather than on specific learning outcomes.

Overall, there is a significant amount of international research assessment and education quality evaluation experience, which offers a relatively stable framework.

With the third mission, it must be admitted that this is not the case. There are many third mission definitions, often with a purely nominalist approach. Most of the literature reports on case studies in particular disciplines or narrowly defined indicators. We know a lot about patents and spin-off companies because they are more observable than, say, contract research (universities keep the names of contractors confidential), or engagement in popular science (media coverage is not reported in CVs or official documentation).

Faced with this complexity, the Agency took a deliberately evolutionary approach. After the first evaluation (2011–2014) there will be a full-scale revision. If there is a need for including other areas, this will be done. A prominent candidate is the recently launched field of student entrepreneurship. It has been documented in the literature that new companies created by students, before their degree, or by graduates immediately after the degree, outnumber spin-off businesses and grow more in turnover and employment. This new area, however, creates huge problems in the collection and standardisation of data at the university level. The Agency initiated two experimental studies, commissioned to the University of Bologna and to Milan Polytechnic, to examine the issue, quantify the phenomenon, and suggest guidelines for future evaluation.

A similar approach will be taken for some of the areas currently under the broad definition of Public Engagement.

4 The Public Engagement of Scholars in SSH. A Preliminary Look at the Italian Research Assessment Evidence

The evaluation of Public Engagement (PE) is among the third mission activities under evaluation in the 2011–2014 VQR framework. PE can be defined as the set of non-profit activities with the aim of educational, cultural and social development. The PE activities and related benefits can be communicated and shared with their audience in several ways. An illustrative list is included in the Manual.[3]

Data on Public Engagement were only collected for 2013 and 2014 and do not cover the entire VQR period because the SUA-RD (*Scheda Unica Annuale della Ricerca Dipartimentale*) data collection system was designed for a different purpose[4] and only for those years.

Universities (and departments) were called to select up to five (three) remarkable PE initiatives and for each to provide:

(a) Date
(b) Title
(c) PE Category (choosing in the list proposed in the Manual)
(d) Brief description (500 words)
(e) Budget and external funds
(f) Estimated impact (e.g. number of participants for events, number of views for web resources, number of copies for publications; estimated audience for TV/radio events, etc.)
(g) Website

Also, each university (and department) should provide information on the presence of a PE monitoring system and dedicated resources.

The expert panel carried out a qualitative analysis of the initiatives. In this chapter we report a collection of data on the PE initiatives in 2013 and 2014 as described by the departments of 94 Universities and shown by disciplinary area. In particular, we compare social sciences and humanities (SSH) with other sciences, i.e. life sciences, engineering and natural sciences.

The total number of expected PE forms was calculated by summing up the number of forms by departments (three per year) and the number of forms per university (five per year). With respect to the expected number, departments have submitted, on average, 75.4% of the total. Departments in SSH have submitted more initiatives: 78.5% of the expected number compared to 73.2% for STEM disciplines.[5] Also, as

[3] See https://www.anvur.it/attachments/article/880/Manuale%20di%20valutazione%20TM~.pdf

[4] The SUA-RD data collection system (the Italian acronyms can be translated as Departmental Research Information System) annually collects data from departments and is used for the ongoing accreditation of universities and academic curricula. The requirement of accreditation of curricula, is that they are carried out by a faculty which is active in research. Data collected via the SUA-RD information system are made available to the expert panel in charge of carrying out the evaluation of the third mission.

[5] The disciplinary classification of departments was made by the CETM expert panel based on the prevalent set of disciplines.

Table 1 Public Engagement initiatives at department level in 2013–2014 by disciplinary macro-area

PE initiatives	SSH	Other disciplines	Total
Expected PE forms[a]	2307	3393	5700
% Expected PE forms	*40,5*	*59.5*	*100*
Filled PE forms	1812	2485	4297
% Filled PE forms	*42.2*	*57.8*	*100*
% Filled/Expected PE forms	*78.5*	*73.2*	*75.4*
N. of departments	382	541	923
N. of valid departments[b]	336	454	790
% of valid departments	*88%*	*84%*	*86%*

Source: VQR 2011–2014

[a]Expected PE forms refers to the maximum number of forms that departments could submit to the SUA-RD information system, i.e. three for each year

[b]Departments which completed at least one form in 2014

many as 88% of departments have submitted at least one initiative, against 84% of departments in STEM. This preliminary finding is interesting since it contradicts the conventional wisdom that Public Engagement, as part of the third mission, is something closer to the interests of people in STEM disciplines (Table 1).

For each form/initiative, universities have chosen one or more PE categories according to the list provided in the Manual. Slightly more than half of the initiatives fall into just one category of the list (53%), while others fall into multiple categories.

Only 15% of the forms offered information on the budget that has been used. In the remaining cases, no data was available. Most of the initiatives report a small budget – less than 10,000 euro, with no differences across disciplines (Fig. 1).

The most diffused initiatives, in order of frequency of occurrence, are the following:

1. organisation of public events (e.g. Researchers' Night, Open Day).
2. interaction with high schools and vocational guidance.
3. active participation in public meetings organised by other parties (e.g. Science cafes, Festivals, Science fairs).
4. organisation of concerts, exhibitions and other public utility events open to the community.
5. popularisation activities aimed at children and young people.
6. interactive websites and/or scientific blogs.
7. educational material published by the teaching staff at national or international level.

In data interpretation, the classification of each initiative under categories was made with no limits on the number of categories to be assigned. This was due to the need to cover as many dimensions of impact as possible. For example, if a natural sciences department organises an exhibition and a programme which involves children or schools in an exhibition, it is reasonable to classify the initiative under two

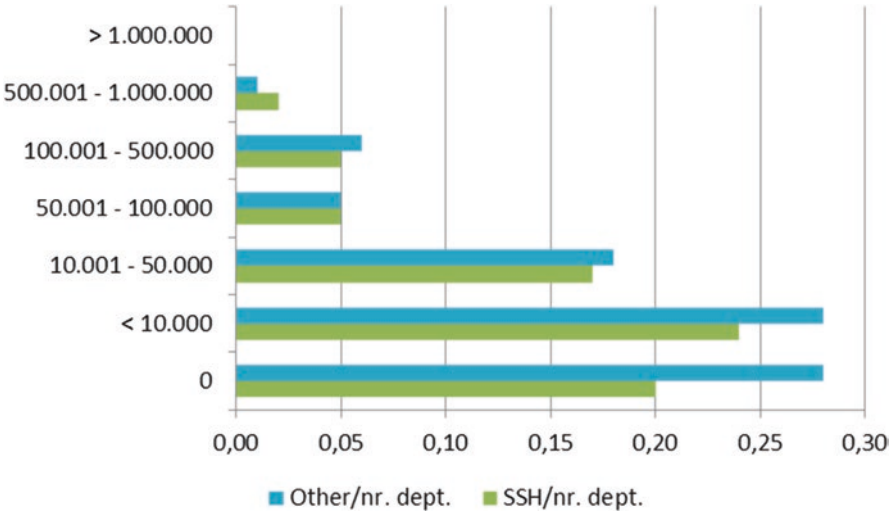


Fig. 1 Budget of initiatives of Public Engagement by disciplinary area. Data in euro. (Source: VQR 2011–2014)

categories. At the same time, the number of categories had no effect on the evaluation because all actions are evaluated under the same broad category of Public Engagement. SSH Departments submitted 1812 initiatives with 3517 categories (1,94 labels per initiative), while STEM departments submitted 2485 with 5410 categories (2,17 labels per initiative).

SSH and STEM departments share a similar pattern on the most important initiatives, that are represented by events, either organised internally or in which they participated (rank 1 and 3 in both fields). On average, all departments submitted one or two initiatives of this kind. We can conclude that the involvement of universities in events is a broadly diffused practice across the entire university system, with no differences between fields.

Immediately afterwards, we observe a slight difference. In SSH departments the item ranked as second is represented by cultural initiatives (concerts, exhibitions, other public events), while in STEM departments it is the interaction with schools and the vocational guidance. The diffusion of cooperation with schools initiatives reflects the awareness of the scientific community that the early involvement of children and students into the life of laboratories is one of the most important factors in their decision to enrol into scientific studies at university level. It is also an important way to make science closer to society, starting with the young population. In SSH the interaction with schools is only slightly less important, ranked fourth.

Interestingly, in STEM departments the initiatives ranked fifth are the popularisation activities aimed at children and young people. Combined with the active involvement with schools it seems that departments in STEM disciplines are more heavily involved in the education of younger generations. It is possible that STEM researchers have a better culture and experience in popularisation, and perhaps

more material for youth. In turn, the initiative ranked fifth in SSH is "popular science publications."

Surprisingly, immediately below the top five, we find in both SSH and STEM departments involved into interactive websites and scientific blogs activities. At least one initiative based on websites and blogs is mentioned in almost 600 cases. This is an interesting finding, given the relative novelty of social network technologies.

Below this rank there is some divergence between SSH and STEM: the latter engage more actively in health-related initiatives, popular science publications and radio and TV programmes, while the former engage in efforts towards children, collaboration with policy making and university magazines.

There are discipline-related initiatives: STEM departments are involved in health-related activities and participate in standardisation boards. Not surprisingly, SSH scholars are much more involved in policy making and much less in health-related initiatives and standardisation committees.

Summing up, the following remarks can be made.

First, scholars in SSH *do* have a third mission, which is carried out with no less engagement than their STEM colleagues. This was the first significant finding from the Italian experience. Asking departments to illustrate their initiatives produces a list of projects in which SSH was no less active.

Second, there is no difference in the engagement of SSH and STEM in public events, popular science, interactions with schools and the use of social media.

Third, the most diffused initiatives are organised into discrete events, such as Open days, Researchers' nights, Science festivals and Science fairs. An interesting (and open) question is to what extent the exposure to discrete events is a sufficient condition for the population or for specific targets such as children and school students, to enter into an awareness and maturation path for science and culture. A related issue is whether discrete events have the potential to engage people actively, and not as a passive audience of shows and exhibitions produced by researchers. On these matters, there is scope for further research and evaluation.

Finally, initiatives such as participatory democracy or citizen science are much less diffused. Only a few dozen experiences were mentioned.

5 Evaluating the Third Mission of SSH Departments on a Large Scale: Preliminary Methodological and Substantive Lessons

The data presented above are descriptive statistics aimed at highlighting areas of interest in the analysis of the engagement of SSH communities. The data show the significant involvement of SSH departments in the submission of dossiers that describe structured initiatives. The overall participation structure is roughly similar for discrete events and for science communication and popularisation, and

interactions with schools and social media. The structure differs significantly in all discipline-related public engagement initiatives, as one would expect. The engagement is much less robust in new policy instruments, such as participatory tools.

Now comes the challenge. How is it possible to evaluate the third mission, given the descriptions offered by the departments?

We try to illustrate a few lessons from this preliminary experience, starting from methodological issues.

In the third mission field there is a visible tension between ecological validity and comparability. Ecological validity requires that the concepts, definitions and metrics closely reflect the differences found in the field. This is more easily found in the evaluation of research, in which a clear distinction between bibliometric tools and expert-based tools is sufficient to map the field. After this classification is introduced, there are many specialised tools and techniques to refine the concepts and metrics, as it happens in the field normalisation of publications and citations. In the third mission field there is a substantial heterogeneity of objects, so that validity can only be the result of an evolutionary and institutional process of approximation and learning. This tension is visible in the use of classifications. The classification used in the SUA-RD data collection system is rather comprehensive, being based on 16 categories. These categories have been defined using a multistep process including:

– A semantic clustering of the information collected in the VQR 2004–2010, under the heading "other third mission activities."
– A comparison with a large-scale survey (ISAAC – Italian Scientists multi-technique Auditing and Analysis on Science Communication) administered in 2014 by a large consortium of Italian universities.
– A comparison with international literature, particularly with the support of the Expert group on the third mission (in which the expert for Public Engagement was prof. Massimiano Bucchi)
– A pre-test with several volunteer departments in 2014.

All this preliminary work notwithstanding, reality exceeds classifications. On the one hand, departments regularly use multiple classifications, as an indication of the complex underlying multidimensionality of the initiatives undertaken. On the other hand, the qualitative dossiers submitted by departments contain much richer and articulated initiatives than those contemplated in the classification.

This richness creates a huge problem for the evaluation methodology, because the degrees of standardisation, replicability and control decrease dramatically.

Another methodological issue refers to the choice of the evaluation framework. Given the heterogeneity and the lack of experience in submitting the dossiers, it was decided to use an informed peer review approach. This failed to solve the issue. It has proved necessary to engage in a calibration exercise, to align the judgments formulated by the experts. The difficulty was that the calibration could not be designed *ex-ante* because of the unknown quality and form of the submitted material. The Manual for the Evaluation of the Third Mission develops some clear quality criteria, but its concrete implementation was not straightforward.

The informed peer review produced an output in which departments have been ordered in broad bands. The resulting output is not a ranking but a collection of good practices. It was felt that the granularity of qualitative judgments was not fine enough to introduce cardinal scales. Yet other options could be considered to enhance the visibility of the third mission, for example, a prize for the best initiative in each area of evaluation.

At the same time, the experts in the CETM report that, with a replication of the assessment and the improvement of the data collection, as well as with feedback from universities and the academic community, the evaluation could make significant progress in a few years.

The Expert group convened by the Agency, called Committee of Experts on Third Mission (CETM), made some recommendations to improve the data collection process and the evaluation of departments and universities. First, there is a discrepancy between the classification offered in the data collection forms and reported in Table 2, and the richness of dossiers submitted. Second, the collection of dossiers is independent of the size of departments and universities, and this does not permit the appropriate evaluation of the relationship between the effort and resources employed and the results on PE. In the next evaluation exercise information on the total number, and related budget, of PE initiatives yearly carried out by universities and departments will be asked. Third, the initial experience shows that the compilation of dossiers is, in some cases, poorly articulated. It is recommended to build a template to increase the evaluation of dossiers, as an inadequate description may penalise good initiatives.

Summing up, the evaluation of Public Engagement is a highly experimental, learning-driven exercise. By its very nature, it must not become a bureaucratic exercise. It is rather a support for institutional learning and a tool to recognise and valorise the potential of human and intellectual engagement of researchers in society.

From a substantive point of view, several remarks can be made.

It is important to reflect on the SSH community's impact and the third mission, by examining the identity and values of their way of conducting research and education. It is a prejudice, often repeated by policymakers and not opposed enough by scholars, that only STEM disciplines contribute to the third mission. There are historical reasons why the attention of administrators and policy makers are focused only on commercialisation of research as the main component of the third mission. One factor is the simultaneity between the technological revolution of the fourth quarter of the twentieth century in life and information science, the changes in US patent regulation, venture capital funding, and the rise of the entrepreneurial university model (Bonaccorsi et al. 2011). These trends created the expectation that universities were in a strong position to enter into the innovation cycle, protecting the intellectual property on their inventions, and obtaining substantial funding from newly created industries and the financial market. All these assumptions proved wrong, as the literature has subsequently recognised.

Yet another important factor is observability. The main elements of the commercialisation are readily observable: patents, licensing contracts, and spinoff companies. These are clearly defined entities. Data are routinely collected (see for example

Table 2 Type of initiatives by discipline

Public Engagement initiatives	SSH	Rank	Other	Rank	Total	SSH per department	Other per department
Popular science publications published by the staff at national or international level	259	**5**	313	**7**	572	0,77	0,69
Participation of staff in radio and TV transmissions at national and international level	183	**10**	302	**8**	485	0,54	0,67
Active participation in public meetings organised by other parties (e.g. Science cafe, Festival, Science fair)	384	**3**	523	**3**	907	1,14	1,15
Organisation of public events (e.g. Researchers' Night, Open Day)	679	**1**	773	**1**	1452	2,02	1,70
Publications (paper and online) dedicated to external audiences (e.g. university magazines)	185	**9**	242	**9**	427	0,55	0,53
Training initiatives on science communication (dedicated to academic or non-academic staff)	100	**14**	110	**14**	210	0,30	0,24
Interactive websites and/or scientific blogs	219	**6**	380	**6**	599	0,65	0,84
Community use of museums, hospitals, sport facilities, libraries, theaters, historical university buildings	124	**12**	159	**12**	283	0,37	0,35
Organisation of concerts, exhibitions and other public utility events open to the community	404	**2**	403	**5**	807	1,20	0,89

(continued)

Table 2 (continued)

Public Engagement initiatives	SSH	Rank	Other	Rank	Total	SSH per department	Other per department
Participation to the formulation of public interest policies (policy making)	216	**8**	204	**10**	420	0,64	0,45
Participation to committees for standardisation and technical norms	58	**16**	153	**13**	211	0,17	0,34
Public health initiatives (e.g. open days dedicated to information and prevention)	65	**15**	380	**6**	445	0,19	0,84
Collaboration with authorities for urban and territory development and valorisation projects	178	**11**	192	**11**	370	0,53	0,42
Interaction with high schools and vocational guidance	371	**4**	760	**2**	1131	1,10	1,67
Popularisation activities aimed at children and young people	217	**7**	444	**4**	661	0,65	0,98
Participatory democracy initiatives (consensus conferences, citizen panels)	109	**13**	72	**15**	181	0,32	0,16
Total	***3751***		**5410**		***9161***	***11,16***	***11,92***

the role of the AUTM Survey in the USA). Data compared in time and space create an entire field of analysis, expectations, and policy making.

The conceptualisation of the entrepreneurial university emphasised the role of universities vis-à-vis the industrial sector. The overall Triple Helix narrative, or Knowledge Triangle rhetoric, gives prominence to the university-business interaction.

It is fair to say that the SSH disciplines accepted that they would be excluded from this narrative and from the conceptualisation of the third mission. It was a mistake. Engaging SSH communities so that they seriously reflected upon their strategies to enter into active and proactive relation with society and the economy did a service to their ability to represent their peculiarities and values.

A second substantive lesson is that words are important. The choice to define the third mission into two distinctly separate areas, and to use a sharp economic

language to characterise the two domains (valorisation and commercialisation, or creation of private goods vs. the creation and maintenance of public goods) was well received and legitimised the evaluation of the third mission. This choice no doubt increased the complexity and difficulty of the task, but it helped to establish a common ground.

It is too early to provide data on the evaluation of third mission activities of SSH researchers – the process is ongoing. In the next few years it will allow large-scale analysis and some comparison with the experience of other countries.

References

Abelhauser, A., Gori, R., & Sauret, M. J. (2011). *La folie évaluation*. Paris: Mille et une nuits.

Abreu, M., Grinevich, V., Hughes, A., & Kitson, M. (2009). *Knowledge exchange between academics, and the business, public and third sectors*. London and Cambridge: UK~IRC.

Adger, W. N., Barnet, J., Brown, K., Marshall, N., & O'Brien, K. (2012). Cultural dimensions of climate change impacts and adaptation. *Nature Climate Change, 3*, 112–117.

Bailey, M., & Freedman, D. (2011). *The assault on universities. A manifesto for resistance*. London: Pluto Press.

Bakhshi, H., Schneider, P., & Walker, C. (2008). *Arts and humanities research and innovation*. Bristol: Arts and Humanities Research Council and National Endowment for Science, Technology and the Arts. Available at http://www.ahrc.ac.uk/documents/project-reports-and-reviews/arts-and-humanities-research-and-innovation/.

Bate, J. (2011). *The public value of the humanities*. London: Bloomsbury Academic.

Belfiore, E., & Upchurch, A. (2013). *Humanities in the twenty-first century. Beyond utility and markets*. Houndmills: Palgrave Macmillan.

Bérubé, M., & Ruth, J. (2015). *The humanities, higher education, and academic freedom*. Houndmills: Palgrave Macmillan.

Bess, J. L. (1988). *Collegiality and bureaucracy in the modern university. The influence of information and power on decision-making structures*. New York: Teachers College Press.

Blasi, B., Romagnosi, S., & Bonaccorsi, A. (2017a). Playing the ranking game: Media coverage of the evaluation of the quality of research in Italy. *Higher Education, 73*(5), 741–757.

Blasi, B., Romagnosi, S., & Bonaccorsi, A. (2017b). Universities as celebrities? How the media select information from a large Research Assessment Exercise. *Science and Public Policy*, forthcoming.

Bod, R. (2013). *A new history of humanities. The search for principles and patterns from antiquity to the present*. Oxford: Oxford University Press.

Bollier, D. (2002). *Silent theft. The private plunder of our common wealth*. New York: Routledge.

Bonaccorsi, A. (2016). Addressing the disenchantment. Universities and regional development in peripheral regions. *Journal of Economic Policy Reform*, Online 31 Aug, 1–28.

Bonaccorsi, A., Calvert, J., & Joly, P. B. (2011). From protecting texts to protecting objects in biotechnology and software. A tale of changes of ontological assumptions in intellectual property protection. *Economy and Society, 40*(4), 611–639.

Bonaccorsi, A., Colombo, M. G., Guerini, M., & Rossi-Lamastra, C. (2013). University specialisation and new firm creation across industries. *Small Business Economics, 41*(4), 837–863.

Bonaccorsi, A., Colombo, M. G., Guerini, M., & Rossi-Lamastra, C. (2014). The impact of local and external university knowledge on the creation of knowledge-intensive firms: Evidence from the Italian case. *Small Business Economics, 43*(2), 261–287.

Bonaccorsi, A., Secondi, L., Ancaiani, A., & Setteducati, E. (2014b). Exploring the role of third-party research. *Journal of Technology Transfer, 39*(2), 169–198.

Bornmann, L. (2013). What is the societal impact of research and how can it be assessed? A literature survey. *Journal of the American Society for Information Science and Technology, 64*(2), 217–233.

Bourdieu, P. (1984). *Homo academicus*. Paris: Les Editions de Minuit.

Bousquet, M., Scott, T., & Parascondola, L. (2004). *Tenured bosses and disposable teachers. Writing instruction in the managed University*. Carbondale: Southern Illinois University Press.

Bowen, W. G., & Tobin, E. M. (2015). *Locus of authority. The evolution of faculty roles in the governance of higher education*. Princeton: Princeton University Press.

Brooks, P. (Ed.). (2014). *The humanities and public life*. New York: Fordham University Press.

Buekers, M., & Nugteren, B. (2012). Creative arts and research-intensive universities: A crucial partnership. LERU Briefing paper no. 2. September 2012. Available at http://www.leru.org/index.php/public/publications/category/briefing-papers/. Accessed Oct 2017.

Cash, D. W., Clark, W. C., Alcock, F., Dickson, N. M., Eckley, N., Guston, D. H., Jager, J., & Mitchell, R. B. (2003). Knowledge systems for sustainable development. *PNAS, 100*(14), 8086–8091.

Cassin, A. (Ed.). (2014). *Derrière les grilles: Sortons du tout-évaluation*. Paris: Fayard/Mille et une nuits.

Center for Economic and Business Research (2015) Contribution of the arts and culture industry to the national economy. An update of our analysis of the macroeconomic contribution of the arts and culture industry to the national economy. Report for Arts Council England, July 2015. Available at http://www.artscouncil.org.uk/media/uploads/documents/publications/340.pdf. Accessed Oct 2017.

Chowdhury, G., Koya, K., & Philipson, P. (2016). Measuring the impact of research: Lessons from the UK's Research Excellence Framework 2014. *PLoS ONE, 11*(6), e0156978.

Citton, Y. (2010). *L'avenir des Humanités. Èconomie de la connaissance ou cultures de l'interpretation?* Paris: La Découverte.

Collini, S. (2012). *What are the universities for?* London: Penguin Books.

Council of University College London (UCL). (2011). *White paper 2011–2021*. London: UCL.

Davies, H., Nutley, S., & Walter, I. (2005). Approaches to assessing the non-academic impact of Social Science Research: Conceptual, methodological and practical Issues. Report to the ESRC. Available at http://www.esrc.ac.uk/. Accessed Oct 2017.

de Campos, A. L. (2010). A study on methodologies for research impact assessment: responses of the UK research councils to the Warry report. *Industry and Higher Education, 24*(5), 393–397.

Dean, M. (2010). *Governmentality. Power and rule in modern society*. Thousand Oaks: Sage Publications.

Dell'Era, C. (2010). Art for business: Creating competitive advantage through cultural projects. *Industry and Innovation, 17*(1), 71–89.

Del Rey, A. (2013). *La tyrannie de l'évaluation*. Paris: La Découverte.

Digital Science. (2016, July). Publication patterns in research underpinning impact in REF 2014. Report to HEFCE. Available at http://www.hefce.ac.uk/pubs/rereports/year/2016/refimpact/. Accessed Oct 2017.

Donovan, C. (2011). State of the art in assessing research impact: Introduction to a special issue. *Research Evaluation, 20*(3), 175–179.

Economic and Social Research Council. (2009). *Taking stock. A summary of ESRC's work to evaluate the impact of research on policy and practice*. Swindon: ESRC. February 2009.

European Commission. (2011). Proposal for a council decision establishing the specific programme implementing Horizon 2020 – The framework programme for research and innovation (2014–2020). Brussels: COM.

Florida, R. (2002). *The rise of the creative class: And how it's transforming work, leisure, community and everyday life*. New York: Perseus Book Group.

Foucault, M. (1978). La « gouvernamentalitè » : Cours du Collège de France, Annéè 1977–1978 : Securitè, territoire et population, 4 leçon, 1 février 1978. In M. Foucault (Ed.), *Dits et écrits II. 1976–1988* (Vol. 2001, pp. 635–657). Paris: Gallimard.

Gingras, Y. (2014). Criteria for evaluating indicators. In B. Cronin & C. Sugimoto (Eds.), *Beyond bibliometrics. Harnessing multidimensional indicators of scholarly impact*. Cambridge, MA: MIT Press.

Gori, R. (2011). *La dignité de penser*. Paris: Les Liens qui Libèrent.

Gori, R. (2013). *La societè des imposteurs*. Paris: Les Liens qui Libèrent.

Grossi, E., Sacco, P. L., Blessi, G. T., & Cerutti, R. (2011). The impact of culture on the individual well-being of the Italian population: An exploratory study. *Applied Research Quality Life, 6*, 387–410.

Haskell J., Hughes A., Bascavusoglu-Moreau E. (2014). The economic significance of the UK science base. A report for the campaign for science and engineering. Available at http://www.sciencecampaign.org.uk/asset. Accessed Oct 2017.

Hughes, A., & Kitson, M. (2012). Pathways to impact and the strategic role of universities: New evidence on the breadth and depth of university knowledge exchange in the UK and the factors constraining its development. *Cambridge Journal of Economics (Special Issue), 36*, 712–750.

Hughes, A., & Kitson, M. (2013). *Connecting with the Ivory Tower: Business perspectives and knowledge exchange in the UK*. London and Cambridge: UK~IRC and NCUB.

Hughes A., Martin B. (2012) Enhancing the impact. The value of public sector R&D. CIHE-UK~IRC task force on enhancing value: Getting the most out of UK research. Available at: https://www.cbr.cam.ac.uk/fileadmin/user_upload/centre-for-business-research/downloads/specialreports/specialreport-enhancingimpact.pdf. Accessed Oct 2017.

Hughes, A., Kitson, M., Probert, J., Turner, R., Bullock, A., & Milner, I. (2011). *Hidden connections knowledge exchange between the arts and humanities and the private, public and third sectors*. Cambridge: Center for Business Research.

Hughes, A., Kitson, M., Probert, J., Turner, R., Bullock, A., & Milner, I. (2014). *Cultural connections: The role of the arts and humanities in competitiveness and local development*. Cambridge: Center for Business Research.

King's College London and Digital Science. (2015, March). *The nature, scale and beneficiaries of research impact: An initial analysis of Research Excellence Framework (REF) 2014 impact case studies*, Research Report 2015/01.

Lawson, C., Hughes, A., Salter, A., Kitson, M., Bullock, A., & Hughes, R. B. (2016). *Knowledge exchange in UK universities: Results from a panel of academics 2005–2015*. London: NCUB.

League of European Research Universities. (2012). *Social siences and humanities: Essential fields for European research in Horizon 2020*. LERU Advice Paper, June 2012. Available at http://www.leru.org/index.php/public/publications/category/advice-papers/

League of European Research Universities (2013). Essential role of SSH research for societal challenges. LERU Notes (Series of 7 separate notes), April–June 2013. Available at http://www.leru.org/index.php/public/publications/category/leru-notes/

League of European Research Universities. (2013b). *Roadmap to improve the position of SSH research in the European Research Area and in Horizon 2020*. LERU Notes. Available at http://www.leru.org/index.php/public/publications/category/leru-notes/

Martin, B. R. (2011). The research excellence framework and the 'impact agenda': are we creating a Frankenstein monster? *Research Evaluation, 20*(3), 247–254.

Martin, B. R. (2012). Are universities and university research under threat? Towards an evolutionary model of university speciation. *Cambridge Journal of Economics (Special Issue), 36*(3), 543–565.

Molas-Gallart, J. (2015). Research evaluation and the assessment of public value. *Arts & Humanities in Higher Education, 14*(1), 111–126.

Molas-Gallart, J., & Castro-Martinez, E. (2007). Ambiguity and conflict in the development of "Third Mission" indicators. *Research Evaluation, 16*(4), 321–330.

Molas-Gallart, J., & Tang, P. (2011). Tracing 'productive interactions' to identify social impacts: An example from the social sciences. *Research Evaluation, 20*(3), 219–226.

Molas-Gallart, J., Tang, P., & Morrow, S. (2000). Assessing the non-academic impact of grant-funded socio-economic research: Results from a pilot study. *Research Evaluation, 9*(3), 171–182.

Molas-Gallart, J., Salter, A., Patel, P., Scott, A., & Duran, X. (2002). *Measuring third stream activities. Final report to the Russell Group of Universities*. Brighton: SPRU University of Sussex.

Morton, S. (2015). Progressing research impact assessment: A 'contributions' approach. *Research Evaluation, 24*, 405–419.

Nussbaum, M. C. (2012). *Not for profit: Why democracy needs the humanities*. Princeton: Princeton University Press.

Olmos-Penuela, J., Molas-Gallart, J., & Castro-Martınez, E. (2013). Informal collaborations between social sciences and humanities researchers and non-academic partners. *Science and Public Policy.* https://doi.org/10.1093/scipol/sct075.

Olmos-Penuela, J., Benneworth, P., & Castro-Martinez, E. (2015). Are sciences essential and humanities elective? Disentangling competing claims for humanities' research public value. *Arts & Humanities in Higher Education, 14*(1), 61–78.

Hughes, A., Moore, B. and Ulrichsen, T. (2010). Synergies and trade-offs between Research, Teaching and Knowledge Exchange, A Report to HEFCE by PACEC and the Centre for Business Research, University of Cambridge, April. Available at https://www.cbr.cam.ac.uk/publications/research-reports/. Accessed Oct 2017.

Penfield, T., Baker, M. J., Scoble, R., & Wykes, M. C. (2014). Assessment, evaluations, and definitions of research impact: A review. *Research Evaluation, 23*, 21–32.

Perkmann, M., Tartari, V., McKelvey, M., Autio, E., Brostrom, A., D'Este, P., Fini, R., Geuna, A., Grimaldi, R., Hughes, A., Kitson, M., Krabel, S., Llerena, P., Lissoni, F., Salter, A., & Sobrero, M. (2013). Academic engagement and commercialisation: A review of the literature on university-industry relations. *Research Policy, 42*(2), 423–442.

Perkmann, M., Fini, R., Ross, J. M., Salter, A., Silvestri, C., & Tartari, V. (2015). Accounting for universities' impact: using augmented data to measure academic engagement and commercialisation by academic scientists. *Research Evaluation, 24*, 380–391.

Reeves, M. (2007). *Measuring the economic and social impact of the arts: A review*. London: Arts Council of England.

Rosli, A., & Rossi, F. (2016). Third-mission policy goals and incentives from performance-based funding. Are they aligned? *Research Evaluation, 25*(4), 427–441.

Samuel, G. N., & Derrick, G. E. (2015). Societal impact evaluation: Exploring evaluator perceptions of the characterization of impact under the REF2014. *Research Evaluation, 24*, 229–241.

Science Europe. (2013). *Humanities in the societal challenges. 12 compelling cases for policy makers*. Bruxelles: Scientific Committee for the Humanities.

Science Europe. (2015). *Radical innovation: Humanities research crossing knowledge boundaries and fostering deep change*. Bruxelles: Scientific Committee for the Humanities.

Smith, S., Ward, V., & House, A. (2011). Impact in the proposals for the UK's Research Excellence Framework. Shifting the boundaries of academic autonomy. *Research Policy, 40*(10), 1369–1379.

Spaapen, J., Dijstelbloem, H., & Wamelink, F. (2007). *Evaluating research in context. A method for comprehensive research assessment*. The Hague: COS.

Spaapen J, van Drooge L, Propp T, et al. (2011) Social impact assessment methods for research and funding instruments through the study of productive interactions between science and society. SIAMPI final report, The Netherlands. Available at: www.siampi.eu/Content/SIAMPI_Final%20report.pdf

Supiot, A. (2015). *La gouvernance par les nombres. Cours au Collège de France (2012–2014)*. Paris: Fayard.

UNDP-UNCTAD (2010). *Creative economy. A feasible development option*, Report 2010. Geneva, Switzerland: United Nations, UNCTAD/DITC/TAB/2010/3. Available at http://unctad.org/en/Pages/DITC/CreativeEconomy/Creative-Economy.aspx

UNESCO. (2016). *Creative cities network*. Paris: UNESCO. Available at http://en.unesco.org/creative-cities/content/about-us.

Wilsdon, J., et al. (2015). The metric tide: Report of the independent review of the role of metrics in research assessment and management. Available at http://www.hefce.ac.uk/pubs/rereports/year/2015/metrictide/. Accessed Oct 2017.

Wolf, B., Lindenthal, T., Szerencsits, T., Britt Holbrook, J., & Heß, J. (2013). Evaluating research beyond scientific impact how to include criteria for productive interactions and impact on practice and society. *GAIA, 22*(2), 104–114.

Part V
What Do We Learn from the Italian Experience?

Language, Culture and Traversing the Scholarly Evaluation Landscape

Alesia Zuccala

1 Introduction

Months before I started to write this chapter, I read an article titled: *11 Untranslatable Words from Other Cultures*. One of the words that appeared in this 'untranslatable' list was the Italian term: '*culaccino*'. This term may not be used often in Italy, but in an everyday context it simply means: "the mark left on a table by a cold glass" (Sanders 2013). *Culaccino*, amongst other internationally 'untranslatable' words (i.e., from Germany, Japan, Spain, Russia etc.) first appeared in a blog (Maptia.com), then later in Sanders (2013) news article, which was written to highlight a book: *Through the Language Looking Glass: Why the World Looks Different in Other Languages* (Deutscher 2010). The thesis of this book is captured by the following question: "If we hold language up as a mirror to the mind, what do we see reflected there: human nature or the cultural conventions of our society?" (p. 14).

According to Deutscher (2010), contemporary linguists tend to agree that "language is primarily an instinct" and "that the fundamentals of language are coded in our genes and are the same across the human race" (p. 6). Every international culture can impart its own labels onto things yet "the concepts behind these labels have been formed by the dictates of nature" (p. 13). In other words, humans generally observe things across nature in the same way, but speak or write about them differently. Because of this, "common sense would suggest that all cultures and all languages should be exactly the same" (p.7). Deutscher (2010) uses the example of words like "cat" and "dog", which are so "clearly delineated by nature [that even though] children always need to be taught the labels for such concepts in a particular language of their society, they don't need to be told how to distinguish between the concepts themselves" (p. 12).

A. Zuccala (✉)
Department of Information Studies, University of Copenhagen, Copenhagen, Denmark
e-mail: a.zuccala@hum.ku.dk

© Springer International Publishing AG 2018

A. Bonaccorsi (ed.), *The Evaluation of Research in Social Sciences and Humanities*, https://doi.org/10.1007/978-3-319-68554-0_17

On the other hand, certain parts of a culture require further reflection, especially when a label is attributed to the realm of abstraction. Deutscher's (2010) second example is that "neither French nor German has a single concept, with a single label, that covers exactly the rank of meanings of the English concept of '*mind*'. If you ask a bilingual dictionary to translate '*mind*' into French, the dictionary will explain patiently that it depends on the context." (p. 14). A reverse situation would be the following: "English does not have a single concept that covers exactly the range of meanings of the French *esprit*. Again, a "dictionary would give a long list of different English words as possible translations" (p. 14).

In light of Deutscher's (2010) text, it can be useful to also think about scholarly research evaluation as a type of 'culture'. The evaluation culture, as we know it today, has evolved into a 'language' and this language includes many labels. Sometimes labels are given to fields of study (e.g., humanities, social sciences, natural sciences), or analytic methods (e.g., bibliometrics), and at other times they adhere to abstract concepts (e.g., quality; impact). Labels and concepts can also be influenced to a certain degree by the *national* context and language in which they are situated. In this sense the evaluation culture is a *plural* culture. In a country like Italy, this means that the way in which scholarly products are evaluated may not necessarily be transferable to all nations, but, the system itself can at least be observed at an international level (note: my discussions with a few authors from this volume occurred during the heat of the Italian summer; a natural time for everyone to observe *culaccinos*). The Italian culture is so greatly reflected by its language, that it will definitely influence how scholars from this country perceive and value academic research. This is especially true for research produced within the social sciences and humanities (SSH). Yet, to the Italian research community's credit they have provided us with an English translation of their work. My task, therefore, is to take this evaluative imprint, this metaphorical 'mark left by a cold glass on a table', and show how it relates to similar 'marks' left by other international scholars.

2 The Language of Scholarly Research Evaluation

In English, the language of scholarly research includes the following items for assessment: journals, articles, books, book chapters, reviews, etc. The Italian research community, of course, uses its own labels for these items, but each can be translated directly into English (as well as many other languages). In fact, most cultures that engage in scholarly research recognize similar, if not exactly the same objects related to specific labels. The technologies available to instantiate, preserve, and determine counts for these labeled items are also now practically universal (note: perhaps not always accessible, but at least universal). So, how can the Italian research system appeal to a general expectation of 'sameness', yet reveal something different?

Three themes pertaining to research evaluation appear in this volume and each one has potential to unify or distinguish the Italian evaluation community from the international one. Notwithstanding Italy's economic and political structure, scholars

from this country, like others across Europe, have become pre-occupied with the following: (1) labeling and establishing classifications for scholarly outputs /academics /communities to be evaluated; (2) discovering or establishing useful datasets/peer communities to support the evaluation of these labeled constructs; and (3) defining and observing quality in scholarship and recognizing broad notions of impact.

I will examine specific chapters from this book under the microscope of the first two themes, and with the third theme, I will take the liberty of reflecting on some culturally-informed ideas surrounding quality and impact.

2.1 Labels, Definitions and Classifications

During a meeting with the editor of this volume, I was provided with a thorough introduction to the development of the Agenzia Nazionale di Valutazione del Sistema Universitario e della Ricerca (ANVUR). I will not describe this organization in detail, or compare it extensively to other national structures/organizations, unless it fits within one or more of my thematic discussions. Some of the individual chapters already provide an in-depth and comparative coverage of the Italian system vis-à-vis others. Here, the aim is to show how this volume is successful at examining and defining labels, in particular, those that have become a critical part of the Italian debate for the Social Sciences and Humanities (SSH).

2.1.1 Books

Sometimes when we identify what an item is *not*, we move closer to determining what it actually *is*. For example, in a previous text concerning SSH research evaluation, Hicks (2004) stated that the "danger of ignoring books is illustrated by exploring the differences between the worlds of book and journal publishing. Books are not just large journal articles" (p. 482). This is a logical assertion, yet Hicks (2004) does not elaborate further upon the publication industry or on the notion of what makes a book precisely more unique. Instead, she focuses on the metrics: "evidence is found in the lack of correlation between cites to books and journal articles" (p. 482).

Williams et al. (2018) have taken a more detailed approach to exploring what a book is, first by appealing to the reader's common sense (e.g., *How many pages are counted before a work is declared a book rather than an article?),* then, by examining a definition of 'book' (e-book included), and noting that books are different from long reports and periodicals because they "do not qualify for an ISBN". Even with a useful definition, the authors show that there can still be anomalies and specifics. For instance, in some African countries, a book published by a scholar does not necessarily "adhere to the ISBN". Also, in Lithuania, where books do adhere to an ISBN, Williams et al. (2018) have found that for

evaluation purposes, the item still has to be "4,000 characters multiplied by a field based coefficient 8".

When different languages are used to label and define a "book", is both the object and concept behind it basically the same? Yes and no. The basic label attributed to the object may be easy to translate, but the concept of *book* becomes more complicated when a decision has to be made about what to actually *do* with that book, and whether or not there is enough detailed record keeping (i.e., metadata for the book). The evaluation culture dictates why, when, where, and how any given label is to be valued. Williams et al. (2018) use parts of their chapter to examine different national evaluation systems (e.g., the Dutch, British, Polish, Spanish, etc.) and conclude that qualitative and quantitative approaches each have unique drawbacks. Here, we see that the heart of the problem is not necessarily the approach, but how labels and definitions are established for a chosen approach. After "having summed up the situation in a number of European countries" the authors proclaim that "it is clear that no definition [of a book] can be applied everywhere".

Similar to other national systems, the Italians also have to rely on labels and definitions. In their chapter, Basili and Lanzillo (2018) define a book, or more specifically a monograph, in the following terms: a "broad, unified and articulated close examination" of a topic, which "cannot be decomposed into separate partitions". And, in distinct Italian style, they explain that "each element [of the monograph] helps to form the complex of the opera". Even if two authors are responsible for this composition, a monograph "cannot and should not be compared to a collection of chapters written by more than one person". At the ANVUR, this definition of *book* is broadened to include "book chapters and conference proceedings", all of which must have an ISBN. A book's scholarly purpose may also be characterized in terms of a "critical edition, translation, or scholarly commentary".

The unique part of Basili and Lanzillo's (2018) chapter, however, is not that they present us with yet another definition for a "book" but that they have also chosen to examine how the "same object can have a different role depending on the type of research it was meant to convey". This is highly pertinent, for example, to the field of legal studies, where many books fit the basic criteria for what a *book* is (e.g., it possesses an ISBN), but cannot be taken at face value on the basis of genre. In law, certain books that are published as manuals might contain original work/research that is similar to what other research disciplines would expect from a traditional scholarly monograph (see Peruginelli and Faro 2018).

2.1.2 Fields/Disciplines

Before the first evaluation took place in Rome at the ANVUR (i.e., the VTR – *Valutazione Triennale della Ricerca*) and prior to the development of other national evaluation systems, classifications were established for journals. Glänzel and Schubert (2003) note that journal classifications were originally created for retrieval purposes. Now, the scientific community relies on categories/classes for broad

aspects of evaluation. In Denmark, for example, the classification system used to evaluate research does not follow the categories that we see for journals indexed in commercial databases (i.e., Clarivate Analytics' Web of Science). Instead, the Danish use an ad hoc approach, whereby scholars selected by the Danish Research Agency are invited to form peer groups (68 in total) representing an academic specialty. Each group works together to identify specialty-area journals that fit a level 1 (normal) or level 2 (prestigious) category. Level 1 journals are linked to a 1.0 reward point and level 2 are linked to 3.0 reward points (Giménez-Toledo et al. 2016). Approaches differ everywhere, and few classification schemes are alike, but many tend to be hierarchical, and each can be created using a *cognitive* approach, a *pragmatic* approach, or a *scientometric* approach when needed (see Glänzel and Schubert 2003).

The Italian evaluation system relies on a *cognitive* approach to classifying fields, and this includes 14 broad areas followed by sub-categories labeled with alphanumeric codes. For example, area 11, designated as "*Scienze storiche, filosofiche, pedagogiche e psicologich* (History, philosophy, pedagogy and psychology), includes "M-STO/01 STORIA MEDIEVAL (Medieval History) as well as "M-STO/04 M-STO/04 STORIA CONTEMPORANEA (Contemporary History). This cognitive classification has been useful for delineating between research outputs that can best be evaluated using bibliometrics (areas 1–9) versus those that benefit from assessments involving peer review (areas 10–14) (see Faggiolani and Solimine 2018). International scholars interested in the full Italian schema, known as the *Settori Scientifico-Disciplinari* (SSD) can find an English translation available online (see https://www.cun.it/uploads/storico/settori_scientifico_disciplinari_english.pdf).

Field labels are necessary for scholarly evaluation procedures, primarily for setting boundaries, and establishing areas for intensive study, but cultural conceptions behind them are open for debate. International scholars might examine, for instance, the degree to which History, Politics and Law constitute a natural cognitive grouping within the social sciences. Peruginelli and Faro (2018) suggest why it may not be a good idea: "law shares many of the peculiarities of social sciences since law is a social phenomenon, but when normativity and legal certainty are concerned, legal scholarship is probably closer to humanities". Also, an examination of the history of music might influence our perception of how well the field of musicology is suited to a quantitative-oriented evaluation program: Is music an art or is it a science? (Cohen 1984, 2010). The Arts and Humanities, or the Humanities and Social Sciences are frequently grouped, but some parts of social science are focused on human development and other areas are based on economic and political issues (Glänzel and Schubert 2003).

In this volume, the approach to field categorizations does not simply settle upon comparing the SSD to what has been introduced in other nations or at the European level (e.g., the ERC model). Faggiolani and Solimine (2018) emphasize why problems underlying evaluation systems are complex: "the correct classification of the branches of knowledge constitutes one of the key elements capable of fostering [all

fields] with major cultural repercussions'. While the authors are critical of the Italian SSD for being based primarily on an "administrative logic" as oppose to the "logic of scholarly communication" they show how this classification scheme can be transformed using a "specificity analysis". This objective of this analysis was to resolve conflicting points of administrative logic and communication logic, by aligning each product from a classified field, subfield, or level with a specific set of valuation criteria. Faggiolani and Solimine (2018) accomplished this using an automatic text analysis of terms used in a document written for the purpose of "identifying evaluation criteria and parameters for the SSH (areas 10–14)". In this case the text was the *Criteri, parametri e indicatori per l'Abilitazione scientifica nazionale*. With the results for the social science-like disciplines of history versus law, they have been able to identify key cultural-communication similarities and differences: (1) History (*monograph, continuity, law*) and (2) Law (*excellence, premise, monograph*).

2.1.3 Peer Review

The terms *peer* and *review* together propagate a host of challenges for the culture of scholarly research evaluation. European-based guidelines have been introduced to assist evaluators with peer selection procedures and setting up effective reviews. Capaccioni and Spina (2018) refer to these guidelines, but outline more specifically how the Italian community has initiated their own peer review procedure. In the 2011–2014 VQR (*Valutazione Della Qualita della Ricerca*) individual reviewers and committee reviewers were used to assess outputs from the SSH (Areas 10–14 in the SSD), with peer review defined in terms of "anonymous qualitative judgement". The review procedure focused on monographs, book chapters, research articles and other relevant outputs characterized by "innovation", "methodological rigor" and "proven or potential impact in the respective international scientific community" (see Capaccioni and Spina 2018).

Normally peer review is concerned exclusively with quality (i.e., often multiple notions of quality), but the alternative process of informed peer review gives individuals or committees the option of enlightening their judgements by observing quantitative data. When metric measures are introduced, peer review has potential to differ from its traditional counterpart. Scholars who are apprehensive about bibliometrics may fear that informed reviews will eventually disappear and that statistical indicators will take over. This fear is rooted in part by a mistrust in indicators in general, but is also reflects the built-in workloads, ethical issues and cost issues associated with recruiting and assigning scholars to reviewer roles. There are benefits, however, to keeping the evaluation culture open to informed peer review. For instance, peer experts/panel members comprised of both national and foreign scholars in Italy recently found that it can be problematic to expand the notion of quality to include internationalisation. Capaccioni and Spina (2018) argue that internationalisation is feasible as long as it "is intended as the ability of our disciplinary sectors to have a dialogue with the world of international

research". But, they recognize also that not all scholarly outputs from the SSH will fit this ideal.

What we learn from the Italian experience is how critical it is to identify national products that fit within the realm of international research, "regardless of the language in which they are written" (see Capaccioni and Spina 2018). This is an open challenge to all national evaluation systems. In cases where internationalization is unclear, a procedure can be initiated to match each product from the SSH disciplines with pre-determined levels for review. Conceptual levels help reviewers decide when quantitative data might serve as appropriate co-source of information for quality judgments. Capaccioni and Spina (2018) explain how the levels graduate, starting with whether or not a SSH product is published in an international outlet, or translated to English or other languages, whether or not it is present in commercial databases (i.e., ISI-WoS/Scopus), if it has been co-authored by both a national and foreign author, produced as a result of funding by and international or European grant, and whether or not the product is included in alternative databases, like library catalogs.

3 Databases and Data Quality for the SSH

The international debate surrounding databases (or data types) suitable for evaluating SSH disciplines began a few years ago (Archambault and Vignola Gagné 2004; Gorraiz et al. 2013; Hicks and Wang 2009; Moed et al. 2009; Sivertsen and Larsen 2012); however, the Italian research community is not contributing too late to this discussion. This volume enters into a time period marked by a growing shift towards national data infrastructures for the SSH (see Giménez-Toledo et al. 2016), as well as an interest in Google Scholar, Google Books, and other under-exploited databases, like international library catalogs (Kousha et al. 2011; Torres-Salinas and Moed 2009; White et al. 2009). Here, I wish to highlight some of the techniques that the Italian researchers in this volume have used to retrieve, refine and utilize data for their most recent evaluation procedure, but first I want to start this section by examining what data means for the social sciences and humanities, and why it has to be approached with an open mind.

Data for the humanities scholar or the social scientist differs from data for disciplines across the sciences. Scientists collect data as they experiment with or observe natural phenomena. For the humanist, data is that which has been generated from the human mind (Bod 2013). Likewise, the social scientist works with data that is generated from human activity or human behavior. Statistical measures are applied to data quite often by social scientists, including scientific methods, but scholars from many humanities disciplines (e.g., philosophy; history; theology) prefer to write and reflect on data that is primarily textual and symbolic. This does not mean, however, that the products of humanistic research cannot be quantified. What it means is that when we work with textual and symbolic material *quantitatively* versus *qualitatively* there is potential to obtain different types and levels of insight (see

Zuccala et al. 2014). In fact, the history of 'citation-ology' (i.e., the study of referencing and citation behavior) has already demonstrated that what we learn from approaching the highly *textual*, *contextual*, and *symbolic* citation, using a qualitative method of investigation can differ greatly than when we approach the same citation using a quantitative method of analysis (Bornmann and Daniel 2008; Brooks 1985; McCain 2006; Small 1978). When we grapple with measuring products and citations from humanistic research, it is less important; therefore, to distinguish data for specific forms of 'treatment' (e.g., *this data is qualitative/quantitative and that is not, so this should only be examined quantitatively/qualitatively*), and more important to focus on how the data needs to be curated to effectively support a chosen method.

In Great Britain, for instance, Thelwall and Delgado (2015) have already advocated for the humanities and humanities-based evaluations with a call out for "no metrics please, just data" (p. 817). The main argument put forth by these scholars is that data for evaluation procedures should not just be more plentiful, but enriched with contextual information. Consider now how a contextual effort might be made with respect to evaluating a scholarly monograph. A peer committee responsible for determining the monograph's quality (i.e., as an original, groundbreaking piece of scholarship) might choose to read excerpts from different chapters and/or search for some of the written reviews it has received. The inclusion of reviews then leads to the following question: was at least one review written about the book, or were none written at all? If the response to the first part is yes, the evaluation team has to decide if the monograph's quality should be linked symbolically to the binary *presence* of public scholarly praise/criticism, or its *absence*. The general context then requires that the evaluators need to determine what this means at both a national or international level. But, the mere presence or absence of a review signifies only one thing, because when several reviews have been written, the number tallied together might produce yet another proxy of quality, and this in turn can bring to the assessment a point of comparison: should books with many reviews be valued or weighted more than those with few or no reviews? Now, even if there is a possibility to count reviews, the process does not have to stop there. Everything that is stated or written in the counted reviews can lead to yet another level of quality assessment because scholarly reviewers are invited to use their text to express, to some extent, both negative and positive opinions. Up to this point the evaluation procedure only gives attention to the content of the book and relevant forms of written recognition. Another proxy of quality could be the symbolic value (and weight?) that evaluators give to the number of citations received, or the publisher of the monograph. When the publisher serves as an indicator of quality this opens the debate up further to establishing notions of production quality, editorial quality, etc.

Since evaluation systems are not designed to focus on just one monograph, a procedure like this is, in reality, compounded when thousands of monographs are included. Context is important, but there is little room to incorporate all that is contextual when it is more critical to reduce the tension between what we expect data to look like and which databases currently record the most reliable data. In Italy, following the VQR 2004–2010, the recognition of and attempts at reducing this ten-

sion for the SSH have been linked to the following databases: Google Scholar and Library Catalogs.

3.1 Google Scholar

Google Scholar is receiving more and more attention as a tool for metric-based research evaluations but international researchers have been reluctant to use this resource without making comparisons to commercial indexes, mainly Scopus and Web of Science (e.g. Bar-Ilan 2008; Jacsó 2005; Meho 2007; Prins et al. 2016). The general consensus, following such comparisons, is that Google Scholar does not have a transparent data curation process, and that it tends to be "particularly poor in information" (Biolcati-Rinaldi et al. 2018). Metadata tags used for information such as type of document and language of contribution are absent from Google Scholar; hence Scopus and the Web of Science which both include more tags, support better opportunities for data retrieval (Biolcati-Rinaldi et al. 2018; Jacsó 2010). International scholars recognize, however, that one of the chief benefits of Google Scholar is its comprehensive coverage (Harzing and van der Wal 2008; Harzing 2014). With disciplines across the SSH, it can be particularly useful for examining books. For example, Kousha et al. (2011) extracted a sample of 1000 books submitted to the 2008 Research Assessment Exercise (RAE) in the United Kingdom, and found that Google Scholar produced "3.2 times more citations" than Scopus, including "medians [that were] three times as high" (p. 2147).

In Italy, the motivation for experimenting with Google Scholar stems from a Ministerial decree surrounding the methods that can be used at ANVUR to evaluate scientific outputs (from areas 1 to 9 in the SSD) versus outputs from the SSH (areas 10–14 in the SSD). Ferrara et al. (2018) note that a "distinction between bibliometric and non-bibliometric scientific areas [was originally] defined by the Italian Ministry of University and Research (MIUR)". Thus far, the Italian evaluation system still recognizes the 'gold standard' of peer review, especially for the Social Sciences and Humanities, but this has not stopped a team of researchers from investigating whether or not data from Google Scholar might be reliable and valid for a real assessment procedure for certain subjects, such as Political Science, History, and General Sociology (i.e., subjects from the research area 14 of the SSD). Again, the work presented in this volume confirms that Google Scholar covers more data relevant to authors and outputs than Scopus or Web of Science (see Biolcati-Rinaldi et al. 2018). However, up to this point, this rate of coverage has never been assessed as thoroughly before without considering key differences between optimizing and cleaning data concerning authors from scientific research fields versus authors from the SSH. Because scientists co-author papers more frequently than scholars from the SSH, the potential of using co-author names in a disambiguation process is limited in these latter fields. This means that disambiguation techniques for author homonyms in the SSH need "solutions based on keyword and linguistic analysis" (see Ferrara et al. 2018).

In the follow-up study, which uses the disambiguated Google Scholar dataset, Biolcati-Rinaldi et al. (2018) present statistically significant correlation values resulting from a comparison of average peer review scores (for three scientific products submitted during the VQR 2004–2010) and two types of indicators extracted from Google Scholar, Scopus and Web of Science (i.e., h-index values and citations per contribution values). Here we have some insight as to the degree to which a relationship exists between the *peer recognized quality* of a research product and its *indicators of scientific recognition*. The Italian system could be leaning toward operationalizing informed peer review for certain disciplines in the future, but Biocalti-Rinaldi et al., indicate that it is still important to "keep distinct the bibliometric level from the evaluative one" because they see the "two disciplines as relatively autonomous".

3.2 Library Catalogs

In 2009, two teams of researchers from different parts of the globe simultaneously suggested that library catalogs might be a good source of data for evaluating book-oriented fields across the SSH (Torres-Salinas and Moed 2009; White et al. 2009). Prior to 2011, which was the year when Thomson Reuters (now Clarivate Analytics) Book Citation Index™ was finally introduced, data resources for books were especially poor and bibliometricians simply could not rely on journal citation indexes alone for evaluating SSH outputs. Torres-Salinas and Moed (2009) decided, therefore, that one could try to "use the same tools and methods as those applied by bibliometricians" but focus on data from library catalogs. Their underlying rationale was that the "inclusion of a book in academic libraries [could be seen] as an expression of its utility for the academic community" (p. 2). An analogy could be drawn "between traditional citation analysis of journal articles and a Library Catalog Analysis of book titles" (p. 2). At the same time, White et al. (2009) advanced this analogy by coining the term "*libcitation*". According to this research group, the libcitation possesses a symbolic value that is quite similar to the citation: "when librarians commit scarce resources to acquiring and cataloging a book, they are in their own fashion citing it, just as scholars do when they refer to it in new works of their own; both are engaged in bibliographic speech acts" (p. 1084).

To date, a handful of studies have grown from these two contributions, and for researchers with a growing interest in library holding counts, the OCLC-WorldCat seems to be the international catalog of choice (Linmans 2010; Kousha et al. 2017; Zuccala and White 2015). OCLC-WorldCat is a union catalog, and because it currently covers libraries worldwide – i.e., academic libraries, national libraries and public libraries – much can be said about a book's perceived *cultural visibility* and *utility*. The libcitation symbolizes the cultural and perceptual element; while the citation can corroborate perception by serving as a symbol of utility. For this reason, it can be interesting to compare the number of academic/institutional libraries that hold a book worldwide, to the number of citations it has received (or did not receive)

after it was published and cataloged (see Zuccala and White 2015). According to Zuccala and White (2015), "both citations and libcitations exhibit the highly skewed distributions typical of bibliometrics" (p.308). Yet, in a test of the relationship between citations and libcitations, Spearman correlation results show that none "are strong enough to indicate that libcitations can substitute for citations as a measure" (p.309).

Biagetti et al. (2018) have taken an altogether different approach to studying library catalogs, and in chapter x, their focus is on the question of data reliability. Here, a suggestion is made that the libcitation may not be the most reliable indicator of a book's esteem [as per White et al. 2009's argument], primarily because the underlying process – i.e., the book's *acquisition* process – cannot always be verified. Biagetti et al. (2018) remind us that many libraries have pre-approved acquisition plans set up with top publishers. Also, many books are purchased by librarians with special attention and intention; while others appear in the same catalog because they have been donated, or sent to the library as a gift. One could survey a sample of libraries to ascertain these underlying factors, as Biagetti et al. (2018) suggest. Or, one could simply accept White et al.'s (2009) approach, which is to trust the role of librarians: "libcitations reflect librarians' knowledge of audiences…what librarians know about the prestige of publishers, the opinions of reviewers, and the reputations of authors" (p. 1084). There is no 'correct' answer to this problem; thus, a refusal to take the libcitation at face value is just similar to questioning the validity of counting citations – i.e., it is not particularly easy to ascertain the reason for a citation either (note: think citer motivation studies and citation context studies).

4 Some Reflections on Scholarly Quality and Impact

4.1 Culture and Perceptions of Quality

In the introductory section of this chapter I referred to a book written by Guy Deutscher (2010), titled *Through the Language Looking Glass: Why the World Looks Different in Other Languages* and suggested that scholarly research evaluation is a type of culture (a pluralistic culture, in fact). I would like to come back to this point, and focus on some of the definitions that we can give to abstract terms like *quality* and *impact*. Within Europe it might seem ideal if every country could agree upon definitions for these interrelated terms, especially with respect to scholarly research, but this may be quite problematic. Not only would this defy Deutscher (2010)'s thesis concerning *why the world looks different in other languages*, it also negates the fact that much of what is produced by international scholars from the social sciences and humanities is *supposed* to be cultural, or at least situated within a cultural context.

The word *culture* is frequently used in the field of anthropology, and for most anthropologists it is defined in terms of "the shared set of (implicit and explicit) values, ideas, concepts, and rules of behaviour that allow a social group to func-

tion." It is also understood to be "dynamic and evolving" because it is the "constructed reality that exists in the minds of social group members" (Hudelson 2004). The Italian culture, for example, has many different phrases or terms that can be used to evoke a shared idea of *quality*. One such term, which is not easy to translate directly into English, is *sprezzatura*. Baldassare Castiglione, the Italian courtier, diplomat, soldier and Renaissance author, was the first person to coin this term because he believed that it captured best what it meant to be a courtier. The ideal Rennaisance courtier was expected to be skilled at many things (i.e., athletics, music, dancing etc.) but somehow he was also expected to be quite modest. Rebhorn (1978) defines this performance-oriented quality as the ability to display "an easy facility in accomplishing difficult actions", yet "hide the conscious effort that went into them" (p. 33).

Another modern phrase for recognizing *quality* in Italy is *regola d'arte*. In English it means that something has been produced "in accordance with the state of the art". Normally, an Italian is more likely to use this word to show appreciation for an object crafted by a workman (e.g., a lamp), not a piece produced by a scholar (e.g., a book), but we can still see from this country's language that in the minds of the Italians, something of high quality is that which has been approached "artfully"; thus *quality* is analogous to being artful.

Since culture is expressed through language, the Italian terms *sprezzatura* and *regola d'arte* indicate how *quality* is a term that has potential to be quite contextually and culturally specific. These examples also demonstrate the degree to which they are time dependent. During the Renaissance period the word *sprezzatura* had a much more positive connotation than it does today. Now it would be taken to mean that a person possesses "an ostentatiously nonchalant attitude of studied indifference" (English translation derived from http://www.treccani.it/vocabolario/sprezzatura/). Modern Italians are thus more likely to recognize a quality performance using other terms. How then can we expect to create a unified notion of scholarly quality, when the international culture of research evaluation is not really a singular culture? Scholarly quality has to reflect a cultural orientation, and naturally different notions of *quality* will evolve over time.

4.2 The Problem of 'International' Impact

In the *Evaluation Society*, Peter Dahler Larsen's (2012) approach to evaluation was to "illustrate what [it] looks like when it is culturally compatible with reflexive modernization. For Larsen (2012), evaluation has become "the manifest sign" that an organization [and even a nation] "is capable of adapting itself to changing conditions" and that it is falling "in line with present cultural expectations" (p. 144).

While the Italian research community has its own language for perceiving and recognizing *quality*, scholars today are under pressure to adapt to changing conditions produced by the international evaluation culture; conditions that require them to produce texts that have (or can potentially have) an international impact. One of

the key questions under debate is whether or not a scholarly product has to be produced in or translated to English in order for this to occur (see Capaccioni and Spina 2018). Thus, similar to the concept of *quality*, the word *impact* needs to be examined quite carefully. If the language of a given culture can play a significant role in the recognition of *scholarly quality*, translation theory can help us understand the problem of international impact.

According to Pellizzi (2012), there are two main approaches to translation: one that is *literary* and another that is *scholarly*. A *literary* translation is focused on the target-language culture and aims to produce the best analog of the source language text at the time of translation. For a scholarly text, this would entail that the translator belongs to the source-language and target-language culture, as well as the research domain for which the translation is needed. In this sense, both language and culture not only refer to a country or region and its societal and linguistic distinctiveness, but the scholarly community or tribe within that country/region, which also possesses values, traditions, terminologies, etc. in its own right. The pitfall related to this form of translation is that the translator might to a greater or lesser degree subjectively re-interpret or alter the original meaning of a text (either knowingly or unknowingly) according to his or her own view of the research domain. Pellizzi (2012) points out that especially for non-fiction texts literary translators run the risk of "going native", or not actually producing an exact translation of the original, but instead making an adapted free translation of the original text.

The *scholarly* translation is focused on the source language text and the cultural-intellectual and historical context in which it was produced (Pellizzi 2012). Because it is different from the *literary* approach to translation, being fluent in a language and having an insider's understanding of a culture is of lesser importance and might even be an advantage. A translator with an outside view has potential to remain true to the original text and thus create a translation that is considered more objective. However, this does not mean that this approach is complete free from complications or consequences. The implied objectiveness of a scholarly translation is that the translator has to make certain that the reader is always aware of the gap between the intellectual world of the original author and the new translation. If a translator is disconnected from a full understanding of both the cultural and linguistic subtleties behind the text, meaning may get lost or distorted somehow in the translation process.

The message that we get from from Pellizzi (2012) is that when we translate an original language text into another language there is the potential risk of altering its original inherent *quality*. Bruno Bettleheim's (1983) book, titled *Freud and Man's Soul* provides the perfect example:

> When in middle age, I was fortunate enough to be permitted to start a new life in the United States, and began to read and discuss psychoanalytic writings in English. I discovered that reading Freud in English translation leads to quite different impressions that I had formed when I read him in German. It became apparent to me that the English renditions of Freud's writings distort much of the essential humanism that permeates the originals" (p. 3).

Bettleheim continues to explain that

> [in] his work and in his writings, Freud often spoke of the soul- of its nature and structure, its development, its attributes, how it reveals itself in all we do and dream. Unfortunately nobody who reads him in English could guess this because nearly all his many references to the soul, and to matters pertaining to the soul have been excised in translation… [The] most important and original concepts of psychoanalysis, makes Freud's direct and always deeply personal appeals to common humanity appear to readers of English as abstract, depersonalized highly theoretical, erudite, and mechanized—in short, scientific statements about the strange and very complex workings of our mind (p. 5).

Since the humanistic efforts of Freud have at one time been altered to suit the intellectual palate of American scholars, and since history has a tendency to repeat itself, evaluators might try to establish qualitative and quantitative assessment procedures that are much more responsive to the problem of 'internationalised' impact. Social scientists and humanities scholars should not be pressured towards producing outputs that have 'potential' for international impact, if it does not reflect their cultural sensibilities. Instead, the system should be designed to guide them with critical information about when it is worthwhile to focus on translating their own works, or have them translated by others for international readers. The translation process itself does not necessarily have to be problematic, as long as it takes a 'clean' *literary* or a 'clean' *scholarly* approach (note: a new approach to clean data!). Moreover, texts that hold a certain meaning and quality within a particular cultural context might perhaps be evaluated separately from their translated versions, as well as documents intended to be of international interest at the outset. This is one (suggested) way to establish a pluralistic culture of evaluation; it may in fact be the only way to respect the value of humanistic research.

5 Conclusion and Acknowledgments

After the Agenzia Nazionale di Valutazione del Sistema Universitario e della Ricerca (ANVUR) was first established in 2006, many Italian scholars have made painstaking efforts to improve upon various methods of evaluation related to scholarly communication practices and outputs across in Italy. Since I have taken a fairly broad approach to traversing the current evaluation landscape, it is important to note that I did not cover every issue that needs to be addressed, or every issue that was highlighted within this particular volume. I would therefore like to urge readers to give special attention to each individual chapter, and treat the present chapter as a useful summary, based on a specific thematic intention.

I would also like to mention that I am not a policy-maker and do not wish to make any official recommendations to that effect. I do believe; however that future evaluation practices should not be hindered due to a lack of resources, fear, ethical issues (which can be monitored), or cultural comparisons that might leave one country believing that its approach to evaluation is better than another. The plural evaluation culture that I describe in this chapter – a European-Anglo culture at this point – possesses greater opportunities for uniting countries/nationalities over a

respect for differences, rather than forced ideals. Each country has its own ‘mark’ to leave and each can have some influence, but no mark should be so permanent that we are locked into a system that cannot evolve.

I wish to thank Andrea Bonaccorsi for giving me the opportunity to learn a great deal about the Italian research evaluation system and for trusting me with the preparation of this chapter. My gratitude also goes to Ginevra Peruginelli for hosting me at her home and at the ITTIG in Florence, at a time when it was crucial to attend a meeting concerning this volume. Alfio Ferrara, Ferruccio Biolcati-Rinaldi, and Luca Lanzillo also attended this meeting, and I am grateful to each of them for sharing their research experiences with me and for engaging me in a lively discussion.

References

Archambault, E., & Vignola Gagné, E. (2004). The use of bibliometrics in the social sciences and humanities. Science-Metrix, Final Report. Prepared for the Social Sciences and Humanities Research Council of Canada (SSHRCC).

Bar-Ilan, J. (2008). Which h-index? – A comparison of WoS, Scopus and Google Scholar. *Scientometrics, 74*(2), 257–271.

Basili, C., & Lanzillo, L. (2018). Research quality criteria in the evaluation of books. In this volume.

Bettleheim, B. (1983). *Freud and man’s soul. An important re-interpretation of Freudian theory*. New York: Vintage Books.

Biagetti, M. T., Iacono, A., & Trombone, A. (2018). Is the diffusion of books in library holdings a reliable indicator in research assessment? In this volume.

Biolcati-Rinaldi, F., Molteni, F., & Salini, S. (2018). Assessing the reliability and validity of Google Scholar indicators. The case of social sciences in Italy. In this volume.

Bod, R. (2013). *A new history of the humanities: The search for principles and patterns from antiquity to the present*. Oxford: Oxford University Press.

Bornmann, L., & Daniel, H.-D. (2008). What do citations measure? A review of studies on citing behaviour. *Journal of Documentation, 64*(1), 45–80.

Brooks, T. A. (1985). Private acts and public objects. An investigation of citer motivations. *Journal of the American Society for Information Science and Technology, 36*(4), 223–229.

Cohen, H. F. (1984). *Quantifying music. The science of music at the first stage of the scientific revolution, 1580 to 1650*. Dordrecht, NL: Reidel.

Cohen, H. F. (2010). Music as science and as art. In R. Bod, J. Maat, & T. Weststeijn (Eds.), *The Making of the Humanities. Volume I. Early Modern Europe*. Amsterdam, NL: Amsterdam University Press.

Deutscher, G. (2010). *Through the language looking glass: Why the world looks different in other languages*. New York: Metropolitan Books.

Faggiolani, C., & Solimine, G. (2018). Mapping the role of the book in evaluation at the individual and department level in Italian SSH: A multisource analysis. In this volume.

Ferrara, A., Montanelli, S., & Verzillo, S. (2018). Google Scholar as a citation database for non-bibliometric areas: The EVA project results. In this volume.

Giménez-Toledo, E., Manana-Rodrıguez, J., Engels, T. C. E., Ingwersen, P., Polonen, J., Sivertsen, G., Verleysen, F. T., & Zuccala, A. A. (2016). Taking scholarly books into account. Current developments in five European countries. *Scientometrics, 107*(2), 685–699. https://doi.org/10.1007/s11192-016-1886-5.

Glänzel, W., & Schubert, A. (2003). A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. *Scientometrics, 56*(3), 357–367.

Gorraiz, J., Purnell, P., & Glänzel, W. (2013). Opportunities and limitations of the book citation index. *Journal of the American Society for Information Science and Technology, 64*(7), 1388–1398.

Harzing, A.-W. (2014). A longitudinal study of Google Scholar coverage between 2012 and 2013. *Scientometrics, 98*(1), 565–575.

Harzing, A.-W., & van der Wal, R. (2008). Google Scholar as a new source for citation analysis. *Ethics in Science and Environmental Politics, 8*, 61–73.

Hicks, D. (2004). The four literatures of social science. In H. F. Moed et al. (Eds.), *Hand book of quantitative science and technology research* (pp. 473–496). The Netherlands: Kluwer Academic Publishers.

Hicks, D., & Wang, J. (2009). Towards a bibliometric database for the social sciences and humanities. Retrieved from http://works.bepress.com/cgi/viewcontent.cgi?article=1017&context=diana_hicks

Hudelson, P. M. (2004). Culture and quality: An anthropological perspective. *International Journal for Quality in Health Care, 16*(5), 345–346.

Jacsó, P. (2005). As we may search – Comparison of major features of the Web of Science, Scopus, and Google Scholar citation-based and citation-enhanced databases. *Current Science, 89*(9), 1537–1547.

Jacsó, P. (2010). Metadata mega mess in Google Scholar. *Online Information Review, 34*(1), 175–191.

Kousha, K., Thelwall, M., & Rezaie, S. (2011). Assessing the citation impact of books: The role of Google Books, Google Scholar, and Scopus. *Journal of the American Society for Information Science and Technology, 62*(11), 2147–2164.

Kousha, K., Thelwall, M., & Abdoli, S. (2017). Goodreads to assess the wider impact of books. *Journal of the Association for Information Science and Technology, 68*(8), 2004–2016.

Linmans, A. J. M. (2010). Why with bibliometrics the Humanities does not need to be the weakest link. Indicators for research evaluation based on citations, library holdings, and productivity measures. *Scientometrics, 83*(2), 337–354.

McCain, K. W. (2006). How influential is Brooks' Law? A longitudinal citation context analysis of Frederick Brooks' The Mythical Man-Month. *Journal of Information Science, 32*(3), 277–295.

Meho, L. (2007). Impact of data sources on citation counts and rankings of LIS faculty: Web of Science versus Scopus and Google Scholar. *Journal of the American Society for Information Science and Technology, 58*(13), 2105–2125.

Moed, H.F., Linmans, J., Nederhof, A, Zuccala, A., Lopez Illescas, C., & de Moya Anegon, F. (2009). Options for a comprehensive database of research outputs in the Social Sciences and Humanities. Research report to the Project Board of the Scoping Study "Towards a Bibliometric Database for the Social Sciences and Humanities". Standing Committees for the Social Sciences and the Humanities of the European Science Foundation (ESF).

Peruginelli, G., & Faro, S. (2018). Research quality evaluation: The case of legal studies. In this volume.

Prins, A. A. M., Costas, R., van Leeuwen, T., & Wouters, P. F. (2016). Using Google Scholar in research evaluation of humanities and social science programs: A comparison with Web of Science data. *Research Evaluation, 25*(3), 264–270.

Rebhorn, W. A. (1978). *Courtly performances: Masking and festivity in Castiglione's book of the courtier*. Detroit: Wayne State University Press.

Sanders, E. F. (2013, October 26). 11 untranslatable words from other cultures. *Huffington Post*. Retrieved from http://www.huffingtonpost.com/ella-frances-sanders/11-untranslatable-words-f_b_3817711.html.

Sivertsen, G., & Larsen, B. (2012). Comprehensive bibliographic coverage of the social sciences and humanities in a citation index: An empirical analysis of the potential. *Scientometrics, 91*(2), 567–575.

Small, H. G. (1978). Cited documents as concept symbols. *Social Studies of Science, 8*, 327–340.

Thelwall, M., & Delgado, M. M. (2015). Arts and humanities research evaluation: No metrics please, just data. *Journal of Documentation, 71*(4), 817–833. https://doi.org/10.1108/JD-02-2015-0028.

Torres-Salinas, D., & Moed, H. F. (2009). Library catalog analysis as a tool in studies of social sciences and humanities: An exploratory study of published book titles in economics. *Journal of Informetrics, 3*(1), 9–26.

White, H., Boell, S. K., Yu, H., Davis, M., Wilson, C. S., & Cole, F. T. H. (2009). Libcitations: A measure for comparative assessment of book publications in the humanities and social sciences. *Journal of the American Society for Information Science and Technology, 60*(6), 1083–1096.

Williams, G., Basso, A., Galleron, I., & Lippiello, T. (2018). More, less or better: The problem of evaluating books in SSH research. In this volume.

Zuccala, A., & White, H. D. (2015). Correlating libcitations and citations in the humanities with WorldCat.org and Scopus Data. In A. A. Salah, Y. Tonta, A. A. Akdag Salah, C. Sugimoto, & U. Al (Eds.), Proceedings of the 15th International Society for Scientometrics and Informetrics (ISSI), Istanbul, Turkey, 29th June to 4th July, 2015. (pp. 305–316). Bogazici University.

Zuccala, A., van Someren, M., & van Bellen, M. (2014). A machine-learning approach to coding book reviews as quality indicators: Towards a theory of mega-citation. *Journal of the American Society for Information Science and Technology, 65*(11), 2248–2260. https://doi.org/10.1002/asi.23104.

Index

© Springer International Publishing AG 2018

A. Bonaccorsi (ed.), *The Evaluation of Research in Social Sciences and Humanities*, https://doi.org/10.1007/978-3-319-68554-0