Journal of Classification 35:5-28 (2018)
DOI: 10.1007/s00357-018-9247-0

Qualitative Judgement of Research Impact: Domain Taxonomy as a Fundamental Framework for Judgement of the Quality of Research

Fionn Murtagh

University of Derby, UK and Goldsmiths, University of London, UK

Michael Orlov

National Research University Higher School of Economics, Moscow, Russia

Boris Mirkin

National Research University Higher School of Economics, Moscow, Russia
Birkbeck, University of London, UK

Abstract: The appeal of metric evaluation of research impact has attracted considerable interest in recent times. Although the public at large and administrative bodies are much interested in the idea, scientists and other researchers are much more cautious, insisting that metrics are but an auxiliary instrument to the qualitative peer-based judgement. The goal of this article is to propose availing of such a well positioned construct as domain taxonomy as a tool for directly assessing the scope and quality of research. We first show how taxonomies can be used to analyze the scope and perspectives of a set of research projects or papers. Then we proceed to define a research team or researcher's rank by those nodes in the hierarchy that have been created or significantly transformed by the results of the researcher. An experimental test of the approach in the data analysis domain is described. Although the concept of taxonomy seems rather simplistic to describe all the richness of a research domain, its changes and use can be made transparent and subject to open discussions.

Keywords: Research impact; Scientometrics; Stratification; Rank aggregation; Multicriteria decision making; Semantic analysis; Taxonomy.

Corresponding Author's Address: Fionn Murtagh, Department of Computing, Goldsmiths, University of London, London SE14 6NW, UK, email: fmurtagh@acm.org.

Published online: 22 March 2018

1. Introduction: The Problem and Background

This article constructively supports the view expressed in the Leiden Manifesto (Hicks et al., 2015), as well as other recent documents such as DORA (Dora, 2013) and Metrics Tide Report (Metric Tide, 2016). All of these advance the principle that assessment of research impact should be made primarily according to qualitative judgment rather than by using citation and similar metrics. It may be maintained, due to the lack of comprehensive recording of process, that the traditional organization of qualitative judgment via closed committees is prone to bias, mismanagement and corruption. In this work, it is proposed to use domain taxonomies for development of open, transparent and unbiased frameworks for qualitative judgments.

In this article, the usefulness of this principled approach is illustrated by, first, the issue of context based mapping and, second, the issue of assessment of quality of research. We propose the direct evaluation of the quality of research, and this principled approach is innovative. We also demonstrate how it can be deployed by using that part of the hierarchy of the popular ACM Classification of Computer Subjects (ACM, 2016) that relates to data analysis, machine learning and data mining. We define a researcher's rank by those nodes in the hierarchy that have been created or significantly transformed by the results of the researcher. The approach is experimentally tested by using a sample of leading scientists in the data analysis domain. The approach is universal and can be applied by research communities in other domains.

In Part 1 of this work, starting with Section 3, there is the engendering and refining of taxonomy. We express it thus to indicate the strong contextual basis, and how one faces and addresses, policy and related requirements. In Part 2 of this work, staring with Section 5, ranking is at issue that accounts fully for both quantitative and qualitative performance outcomes.

2. Review of Research Impact Measurement and Critiques

The issue of measuring research impact is attracting intense attention of scientists because metrics of research impact are being widely used by various administrative bodies and by the public at large as easy-to-get shortcuts for assessment of comparative strengths among scientists, research centers, and universities. This is further boosted by the wide availability of digitalized data and, as well, by the fact that research nowadays becomes a widespread activity. The number of citations and such derivatives as the Hirsch index are produced by a number of organizations including the inventors, currently Thomson Reuters (Thomson Reuters, 2016), Scopus, and

Google. There is increasing pressure to use these or similar indexes in evaluation and management of research. There have been a number of proposals to amend the indexes, say, by using less extensive characteristics, such as centrality indexes in the inter-citation graphs or by following only citations in the work of "lead scientists" (Aragnón, 2013). Other proposals deny the usefulness of bibliometrics altogether; some propose even such alternative measures as the "careful socialization and selection of scholars, supplemented by periodic self-evaluations and awards" (Osterloh and Frey, 2014), that is, a social- and behavioral-based, administrative, exemplary model. Other, more practical systems, such as the UK Research Assessment Exercise (RAE), now the REF, Research Excellence Framework), intends to assess most significant contributions only, and in a most informal way, which seems a better option. However, there have been criticisms of the RAE-like systems as well: first, in the absence of a citation index, the peer reviews are not necessarily consistent in evaluations (Eisen, MacCallum, and Neylon, 2013), and, second, in the long run, the system itself seems somewhat shortsighted; it has cut off everything which is out of the mainstream (Lee, Pham, and Gu, 2013). There have been a number of recent initiatives undertaken by scientists themselves such as the San-Francisco Declaration DORA (Dora, 2013), Leiden Manifesto (Hicks et al., 2015), and The Metrics Tide Report (Metric Tide, 2016). DORA, for example, emphasizes that research impact should be scored over all scientific production elements including data sets, patents, and codes among others (Dora, 2013). Altogether, these declarations and manifestos claim that citation and other metrics should be used as an auxiliary instrument only; the assessment of research quality should be based on "qualitative judgement" of the research portfolio (Hicks et al., 2015). Yet there is no clarity on the practical implementation of these recommendations.

This article is a further step in this direction. Any unbiased consideration of metrics as well as of other systems for assessment of research impact (Eisen et al., 2013; Lee et al., 2013) leads to conclusions that "qualitative judgment" should be a preferred option (Dora, 2013; Hicks et al., 2015; Metric Tide, 2016). This article points out the concept of domain taxonomy which should be used as a main tool in actual organization of assessment of research impact in general and quality of research, specifically.

The remainder of this article is organized as follows. We begin by briefly reviewing direct and straightforward application of domain taxonomy, for supporting qualitative judgement. Relating to the policy-related work of a national research funding agency, and to the editorial work of a journal, these preliminary studies were pioneering.

The third section explains how a domain taxonomy can be used for assessing the quality of research. The fourth section provides an experiment

in testing the approach empirically. The fifth section compares the taxonomic ranking of our sample of scientists with rankings over citation and merit.

3. Qualitative, Content-Based Mapping, into which the Quantitative Indicators are Mapped

In this section and in the next section, we develop taxonomies using sets of keywords or selected actionable terms. It is sought to be, potentially, fully data-driven. Levels of resolution in our taxonomy can be easily formed through term aggregation. Mapping the taxonomy, as a tree endowed with an ultrametric, to a metric space, when using levels of aggregation, provides an approach to having focus (in a general sense, orientation and direction) in the analytics.

Here we give a first example, in which the taxonomies were generated with the goal to provide a tool for open and unbiased qualitative judgment in such contexts as research publishing and research funding. Concept hierarchies can be established by domain experts, and deployed in such contexts as research publishing and research funding.

A short review was carried out of thematic evolution of The Computer Journal, relating to 377 papers published between January 2000 and September 2007. The construction of a concept hierarchy, or ontology, was "bootstrapped" from the published articles. The top level terms, child nodes of the concept tree root, were "Systems – Physical", "Data and Information", and "Systems – Logical". Noted was that the category of "bioinformatics" did not require further concept child nodes. A limited set of sub-categories was used for "software engineering", these being "Design", "Education", and "Programming languages". Under the top level category of "Data and information", one of the eight child nodes was "Machine learning", and one of its child nodes was "Plagiarism". This was justified by the appropriateness of the contents of published work relating to plagiarism. Once the concept hierarchy was set up, the 377 published articles from the seven years under investigation were classified, with mostly two of the taxonomy terms being used for a given article. There was a maximum of four taxonomy terms, and a minimum of one. Table 1 displays the concept hierarchy that was used at that time.

A Correspondence Analysis of this data, here with a focus on the top level themes, presents an interesting and revealing view. A triangle pattern is to be seen, in Figure 1, where Inf is counterposed on the first factor to the two other, more traditional Computer Science themes. Factor 2 counterposes the physical and the logical in the set of published research work. The information displayed in Figure 1 comprises all information, that is the in-

Table 1. Concept hierarchy, incrementally constructed, representing a view of appropriate subject headings for articles published in the Computer Journal, 2000–2007 (continued on next page).

```
1. Systems -- Physical
   1.1. Architecture, Hardware
        1.1.1. Networks, Mobile
        1.1.2. Memory
   1.2. Distributed Systems
        1.2.1. System Modelling
        1.2.2. Networks, Mobile
        1.2.3. Grid, P2P
        1.2.4. DS Algorithms
        1.2.5. Semantic Web
        1.2.6. Sensor Networks
   1.3. Networks, Mobile
        1.3.1. Mobile Computing
        1.3.2. Networks
        1.3.3. Search, Retrieval
   1.4. Information Delivery
        1.4.1. Energy
               1.4.1.1. Photonics-based
               1.4.1.2. Nano-based
        1.4.2. Displays
        1.4.3. Bio-Engineering Applications
        1.4.4.  Miscellaneous Applications of Materials
2. Data and Information
   2.1. Storage
        2.1.1. Databases
        2.1.2. Graphics
        2.1.3. Imaging, Video
        2.1.4. Memory Algorithms
        2.1.5. Non-Memory Storage Algorithms
        2.1.6. Network Storage Algorithms
   2.2. Knowledge Engineering
        2.2.1. Data Mining
        2.2.2. Machine Learning
        2.2.3. Search, Retrieval
   2.3. Data Mining
        2.3.1. Imaging, Video
        2.3.2. Semantic Web
        2.3.3. Complexity
   2.4. Machine Learning
        2.4.1. Databases
        2.4.2. ML Algorithms
        2.4.3. Reasoning
        2.4.4. Representation
   2.5. Quantum Processing
   2.6. Algorithms
        2.6.1. Coding, Compression, Graphs, Strings, Trees
   2.7. Bioinformatics
   2.8. Computation Modelling
3. Systems -- Logical
   3.1. Information Security
        3.1.1. Networks, Mobile
```

Table 1. Concept hierarchy, incrementally constructed, representing a view of appropriate subject headings for articles published in the Computer Journal, 2000–2007 (continued).

```
3.2. Software Engineering
     3.2.1. Design
     3.2.2. Education
     3.2.3. Programming Languages
3.3. System Modelling
     3.3.1. Software Engineering
     3.3.2. Testing
     3.3.3. Ubiquitous Computing
     3.3.4. Workflow
     3.3.5. Games
     3.3.6. Human Factors
     3.3.7. Virtual Materials Science
```

ertia of the cloud of publications, and of the cloud of these top level themes. The year of publication, as a supplementary attribute of the publications, is inactive in the factor space definition, and each is projected into the factor space. We see the movement from year to year, in terms of the top level themes. There is further general discussion in Murtagh (2008).

The perspective described, for archival, scholarly journal publishing, relates to the narrative or thematic evolution of research outcomes.

4. Application of Narrative Analysis to Science and Engineering Policy

This same perspective as described in the previous section was prototyped for the narrative ensuing from national science research funding. The aim here was thematic balance and evolution. Therefore it was complementary to the operational measures of performance—numbers of publications, patents, PhDs, company start-ups, etc. In Murtagh (2010), the full set of large research centers (8 of these, with up to 20 million euro funding) and a class of less large research centers (12 of these, each with 7.5 million euro funding) were mapped into a Euclidean metric endowed, Correspondence Analysis, factor space. In this space there is displayed the centers, their themes, and, as a prototyping study, just one attribute of the research centers, their research budget. The first factor clearly counterposed centers for biosciences to centers for telecoms, computing and nanotechnology. The second factor clearly counterposed centers for computing and telecoms to nanotechnology. This is further elaborated in Section 4.1.

All in all, there is enormous scope for insight and understanding, that starts from subject matter and content. Quantitative indicators are well accommodated, with their additional or complementary information. It may well be hoped that in the future, qualitative, content-based analytics,

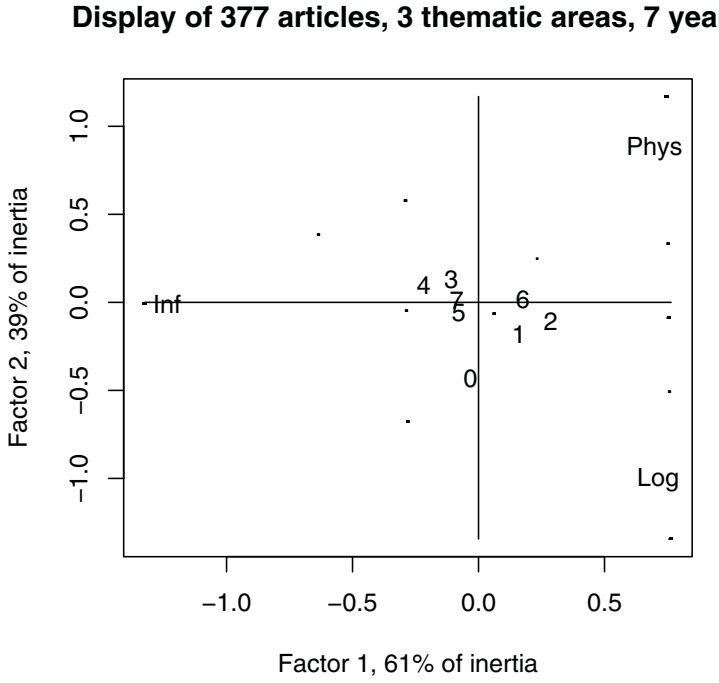

Figure 1. Principal factor plane of Correspondence Analysis of 377 published articles (positions shown with dots, not all in this central region), crossed by three primary thematic areas. These are: Information and Data (Inf), Systems–Physical (Phys), and Systems–Logical (Log). The years of publication shown (0 = 2000, 1 = 2001, etc.), used as supplementary elements in the analysis.

coupled with quantitative indicators, will be extended. For this purpose, it may well be very useful to consider not just published research, but all written, and subsequently submitted, research results and/or plans. Similarly for research funding, the content-based mapping and assessment of rejected work is relevant too, not least in order to contextualize the content of all domains and disciplines.

The role of taxonomy is central to the information focusing that is under discussion in this section. Information focusing is carried out through mapping the ontology, or concept hierarchy, as a level of aggregation, corresponding therefore to non-terminal, i.e. non-singleton, nodes. Our interest in this data is to have implications of this for data mining with decision policy support in view.

Consider a fairly typical funded research project, and its phases up to and beyond the funding decision. A narrative can always be obtained, in one form or another, and is likely to be a requirement. All stages of the proposal

and successful project life cycle, including external evaluation and internal decision making, are highly documented—and as a consequence narrative–based.

As a first step, let us look at the very general role of narrative in national research development. The following comprise our motivation: overall view, i.e. overall synthesis of information; orientation of strands of development; their tempo and rhythm.

Through such an analysis of narrative, among the issues to be addressed are the following: strategy and its implementation in terms of themes and subthemes represented; thematic focus and coverage; organizational clustering; evaluation of outputs in a global context; all the above over time.

The aim here is to view the "big picture". It is also to incorporate contextual attributes. These may be the varied performance measures of success that are applied, such as publications, patents, licences, numbers of PhDs completed, company start-ups, and so on. It is instead to appreciate the broader configuration and orientation, and to determine the most salient aspects underlying the data.

4.1 Assessing Coverage and Completeness

SFI Centers for Science, Engineering and Technology (CSETs) are campus-industry partnerships typically funded at up to €20 million over 5 years. Strategic Research Clusters (SRCs) are also research consortia, with industrial partners and over 5 years are typically funded at up to €7.5 million.

We cross-tabulated 8 CSETs and 12 SRCs by a range of 65 terms derived from title and summary information; together with budget, numbers of PIs (Principal Investigators), Co-Is (Co-Investigators), and PhDs. We can display any or all of this information on a common map, for visual convenience a planar display, using Correspondence Analysis.

In mapping SFI CSETs and SRCs, we will now show how Correspondence Analysis is based on the upper (near root) part of an ontology or concept hierarchy. This we view as *information focusing*. Correspondence Analysis provides simultaneous representation of observations and attributes. Retrospectively, we can project other observations or attributes into the factor space: these are supplementary observations or attributes. A 2-dimensional or planar view is likely to be a gross approximation of the full cloud of observations or of attributes. We may accept such an approximation as rewarding and informative. Another way to address this same issue is as follows. We define a small number of aggregates of either observations or attributes, and carry out the analysis on them. We then project the full set of observations and attributes into the factor space. For mapping of SFI

CSETs and SRCs a simple algebra of themes as set out in the next paragraph achieves this goal. The upshot is that the 2-dimensional or planar view is a better fit to the full cloud of observations or of attributes.

From CSET or SRC characterization as: Physical Systems (Phys), Logical Systems (Log), Body/Individual, Health/Collective, and Data & Information (Data), the following thematic areas were defined.

1. eSciences = Logical Systems, Data & Information
2. Biosciences = Body/Individual, Health/Collective
3. Medical = Body/Individual, Health/Collective, Physical Systems
4. ICT = Physical Systems, Logical Systems, Data & Information
5. eMedical = Body/Individual, Health/Collective, Logical Systems
6. eBiosciences = Body/Individual, Health/Collective, Data & Information

This categorization scheme can be viewed as the upper level of a concept hierarchy. It can be contrasted with the somewhat more detailed scheme that we used for analysis of articles in the Computer Journal, (Murtagh, 2008).

CSETs labeled in the Figures are: APC, Alimentary Pharmabiotic Center; BDI, Biomedical Diagnostics Institute; CRANN, Center for Research on Adaptive Nanostructures and Nanodevices; CTVR, Center for Telecommunications Value-Chain Research; DERI, Digital Enterprise Research Institute; LERO, Irish Software Engineering Research Center; NGL, Center for Next Generation Localization; and REMEDI, Regenerative Medicine Institute.

In Figure 2, eight CSETs and major themes are shown. Factor 1 counterposes computer engineering (left) to biosciences (right). Factor 2 counterposes software on the positive end to hardware on the negative end. This 2-dimensional map encapsulates 64% (for factor 1) + 29% (for factor 2) = 93% of all information, i.e. inertia, in the dual clouds of points. CSETs are positioned relative to the thematic areas used. In Figure 3, sub-themes are additionally projected into the display. This is done by taking the sub-themes as *supplementary elements* following the analysis as such. From Figure 3 we might wish to label additionally factor 2 as a polarity of data and physics, associated with the extremes of software and hardware.

4.2 Change Over Time

We take another funding program, the Research Frontiers Programme, to show how changes over time can be mapped.

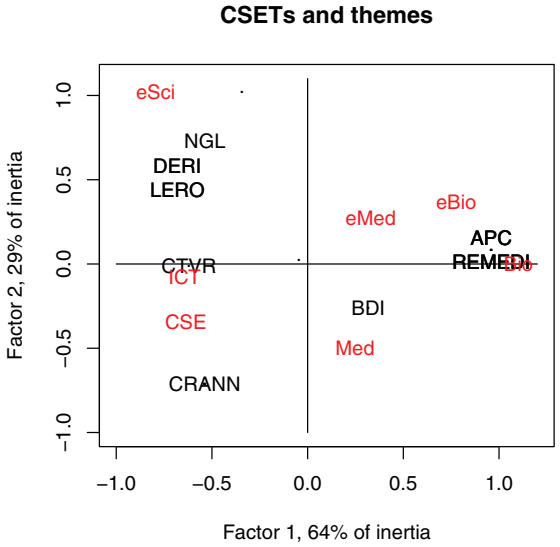

Figure 2. CSETs, labeled, with themes located on a planar display, which is nearly complete in terms of information content.

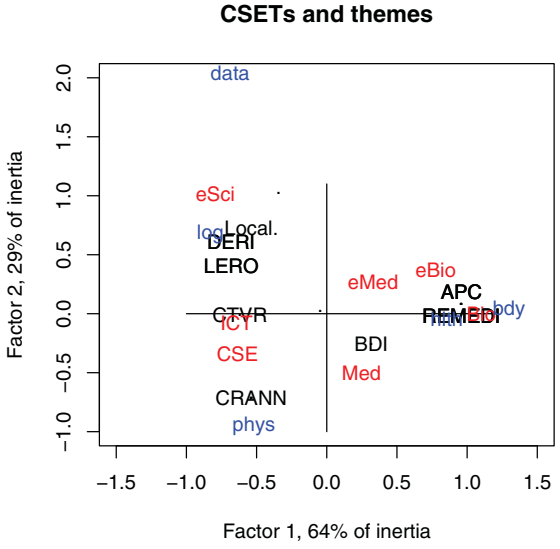

Figure 3. As Figure 2 but with sub-themes projected into the display. Note that, through use of supplementary elements, the axes and scales are identical to those on Figure 2. Axes and scales are just displayed differently in this figure so that sub-themes appear in our field of view.

This programme follows an annual call, and includes all fields of science, mathematics and engineering. There are approximately 750 submissions annually. There was a 24% success rate (168 awards) in 2007, and 19% (143 awards) in 2008. The average award was €155k in 2007, and €161k in 2008. An award runs for three years of funding, and this is moving to four years in 2009 to accommodate a 4-year PhD duration. We will look at the Computer Science panel results only, over 2005, 2006, 2007 and 2008.

Grants awarded in these years, respectively, were: 14, 11, 15, 17. The breakdown by institutes concerned was: UCD – 13; TCD – 10; DCU – 14; UCC – 6; UL – 3; DIT – 3; NUIM – 3; WIT – 1. These institutes are as follows: UCD, University College Dublin; DCU, Dublin City University; UCC, University College Cork; UL, University of Limerick; NUIM, National University of Ireland, Maynooth; DIT, Dublin Institute of Technology; and WIT, Waterford Institute of Technology.

One theme was used to characterize each proposal from among the following: bioinformatics, imaging/video, software, networks, data processing & information retrieval, speech & language processing, virtual spaces, language & text, information security, and e-learning. Again this categorization of computer science can be contrasted with one derived for articles in recent years in the Computer Journal (Murtagh, 2008).

Figures 4, 5 and 6 show different facets of the Computer Science outcomes. By keeping the displays separate, we focus on one aspect at a time. All displays however are based on the same list of themes, and so allow mutual comparisons. Note that the principal plane shown accounts for just 9.5% + 8.9% of the inertia. Although accounting for 18.4% of the inertia, this plane, comprising factors, or principal axes, 1 and 2, accounts for the highest amount of inertia (among all possible planar projections). Ten themes were used, and what the 18.4% information content tells us is that there is importance attached to most if not all of the ten.

4.3 Conclusion on the Policy Case Studies

The aims and objectives in our use of the Correspondence Analysis and clustering platform is to drive strategy and its implementation in policy. What we are targeting is to study highly multivariate, evolving data flows. This is in terms of the semantics of the data—principally, complex webs of interrelationships and evolution of relationships over time. This is the *narrative of process* that lies behind raw statistics and funding decisions. We have been concerned especially with *information focusing* in Section 4.1, and this over time in Section 4.2.

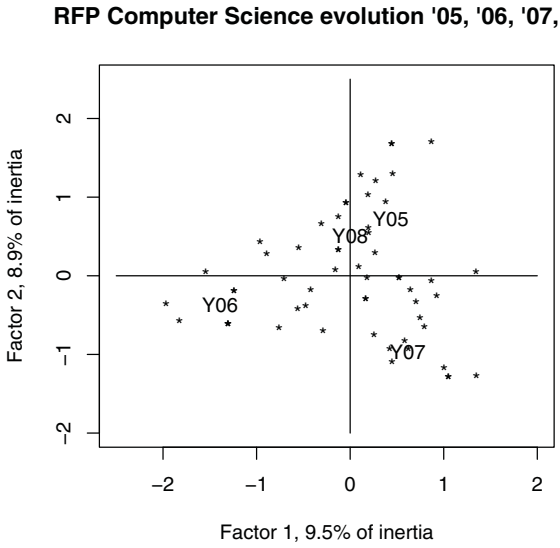

Figure 4. Research Frontiers Programme over four years. Successful proposals are shown as asterisks. The years are located as the average of successful projects.

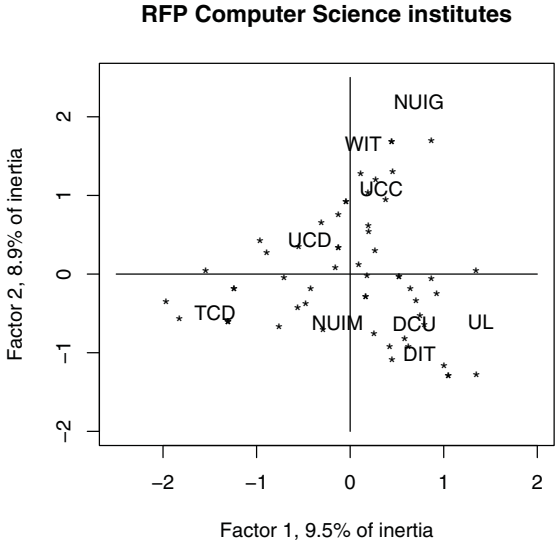

Figure 5. As Figure 4, displaying host institutes of the awardees.

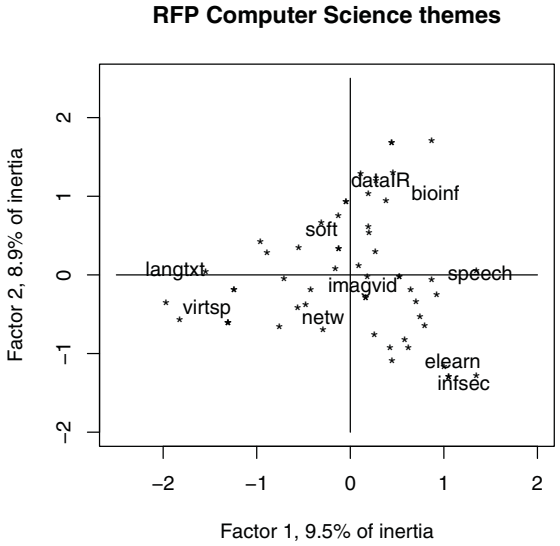

Figure 6. As Figures 4 and 5, displaying themes.

5. Domain Taxonomy and Researcher's Rank for Data Analysis

Here we turn to a domain taxonomy, that is the Computing Classification System maintained and updated by the Association of Computing Machinery (ACM-CCS); the latest release, of 2012, is publicly available at ACM (2012). Parts of ACM-CCS 2012 related to the loosely defined subject of "data analysis" including "Machine learning" and "Data mining", up to a rather coarse granularity, are presented in Table 2.

It should be noted that a taxonomy is a hierarchical structure for shaping knowledge. The hierarchy involves just one relation "A is part of B" so that it leaves aside many other aspects of knowledge including, for example, the differences between theoretical interrelations, computational issues and application matters of the same set of concepts. These, however, may sneak in, even if unintentionally, in practice. For example, topics representing "Cluster analysis" occur in the following six branches within the ACM-CCS taxonomy: (i) Theory and algorithms for application domains, (ii) Probability and statistics, (iii) Machine learning, (iv) Design and analysis of algorithms, (v) Information systems applications, (vi) Information retrieval. Among them, (i) and (ii) refer to theoretical work, (iv) to algorithms, (v) and (vi) to applications. Item (iii), Machine learning, probably embraces all of them.

Table 2. ACM CCS 2012 high rank items covering data analysis, machine learning and data mining

Subject index	Subject name
1.	Theory of computation
1.1.	Theory and algorithms for application domains
2.	Mathematics of computing
2.1.	Probability and statistics
3.	Information systems
3.1.	Data management systems
3.2.	Information systems applications
3.3.	World Wide Web
3.4.	Information retrieval
4.	Human-centered computing
4.1.	Visualization
5.	Computing methodologies
5.1.	Artificial intelligence
5.2.	Machine learning

Unlike in biology, the taxonomies of specific research domains cannot be specified exactly because of the changing structure of the domain and, therefore, are subject to much change. For example, if one compares the current ACM Computing Classification System 2012 (ACM, 2012) with its previous version, the ACM Classification of Computing Subjects 1998 which is available at the same site, one cannot help but notice great differences in both the list of sub-domains and the structure of their mutual arrangement.

We consider the set of branches in Table 2 as a taxonomy of its own, referred to below as the Data Analysis Taxonomy (DAT). An extended version of the taxonomy, along with three to four more layers of higher granularity, presented in Mirkin and Orlov (2015, pp. 241-249), will be used throughout for illustration of our approach.

Out of various uses of a domain taxonomy, we pick up here its use for determining a scientist rank according to the rank of that node in the taxonomy which has been created or significantly transformed because of the results by the scientist (Mirkin, 2013).

The concept of taxonomic rank is not uncommon in the sciences. It is quite popular, for example, in biology: "A Taxonomic Rank is the level that an organism is placed within the hierarchical level arrangement of life forms" (see http://carm.org/dictionary-taxonomic-rank). As mentioned in Mirkin and Orlov (2015), *Eucaryota* is a domain (rank 1) containing *Animals* kingdom (rank 2). The latter contains *Cordata* phylum (rank 3) which contains *Mammals* class (rank 4) which contains *Primates* order (rank 5)

which contains *Hominidae* family (rank 6) which contains *Homo* genus (rank 7) which contains, finally, *Homo sapiens* species (rank 8). Similarly, the rank of the scientist who created the "World wide web" (Berners-Lee, 2010), (the item 3.3 in Table 2 at layer 2 of the DAT taxonomy, is 2; and the rank of the scientist who developed a sound theory for "Boosting" (Schapire, 1990), (the item 1.1.1.5 in DAT (Mirkin and Orlov, 2015)), is 4, whereas the rank of the scientists who proposed a sound approach to "Topic modeling" (Blei et al., 2003) (the item 5.2.1.2.4 in DAT (Mirkin and Orlov, 2015)) is 5. This specification of taxonomic rank, TR, is associated with qualitative innovation, whereas the dominant current approach is to only reward or take account of low rank, and particular, topic items.

Using taxonomic ranks (TRs) based on domain taxonomies for evaluating the quality of research differs from the other methods currently available, through the following features:

- The TR method directly measures the quality of results themselves rather than any related feature such as popularity;
- The TR evaluation is well subject-focused; a scientist with good results in optimization may get rather modest evaluation in data analysis because a taxonomy for data analysis would not include high-level nodes on optimization;
- The TR rank can get reversed if the taxonomy is modified so that the rank-giving taxon gets a less favorable location in the hierarchical tree;
- The granularity of evaluation can be changed by increasing the granularity of the underlying taxonomy;
- The TR evaluations in different domains can be made comparable by using taxonomies of the same depth;
- The maintenance of a domain taxonomy can be effectively organized by a research community as a special activity subject to regular checking and scrutinizing;
- Assigning the TR to a scientist or their result(s) is derived from mapping them to a sub-domain that has been significantly affected by them, and this is not a simple issue. The persons who do the mapping must be impartial and have deep knowledge of the domain and the results.

The last two items in the list above refer to the core of the proposal in this paper. They can be considered a clarification of the main claim over evaluation of the research impact made by the scientists: qualitative considerations should prevail over metrics (Dora, 2013; Hicks et al., 2015; Metric Tide, 2016). Here the wide meaning of "qualitative" is reduced to two

points: (a) developing and maintaining of a taxonomy, and (b) mapping results to the taxonomy. Both taxonomy developing any mapping decisions involve explicitly stated judgements which can be discussed openly and corrected if needed. This differs greatly from the currently employed procedures of peer-reviewing which can be highly subjective and dependent on various external considerations (Eisen et al., 2013; Engels et al., 2013; Van Raan, 2006). The activity of developing and maintaining taxonomies can be left to the governmental agencies and funding bodies, or to scholarly academies, or to discipline and sub-discipline expert organizational bodies, whereas the mapping activity should be left, in a transparent way, to scientific discussions involving all relevant individuals. Of course, there is potential for further developments of the formats: taxonomies could be extended to include various aspects characterizing research developments, and mapping can be softened up to include spontaneous and uncertain judgements.

6. A Prototype of Empirical Testing

We focus on the field of Computer Science related to data analysis, machine learning, cluster analysis and data mining along with its taxonomy derived from the ACM Computing Classification System 2012 (ACM, 2012), as explained above. We pick up a sample of 30 leading scientists in the field (about half from the USA, and other, mostly European, countries are represented by 2–3 representatives), such that the information of their research results is publicly available. Although we tried to predict the leaders, their Google-based citation indexes are highly different, from a few thousand to a hundred thousand. We picked up 4–6 most important papers by each of the sampled scientists and manually mapped each of the papers to taxons significantly affected by that. Since some of the relevant subjects, such as "Consensus clustering" and "Anomaly detection", have not been presented in the ACM-CCS, we added them to DAT (Data Analysis Taxonomy) as leaves, implying that a previous terminal node becomes a non-terminal node. The results of the mapping are presented in Table 3. The table also presents the derived taxonomic ranks and the same ranks, 0–100 normalized. To derive the taxonomic rank of a scientist, we first take the minimum of their ranks as the base rank. Then we subtract from it as many one tenths as there are subdomains of that rank in their list and as many one hundredths as there are subdomains of greater ranks in the list. For example, the list of S23 comprises ranks 4, 5, 4 leading to 4 as the base rank. Subtraction of two tenths and one hundredth from 4 gives the derived rank 3.79. The normalization is such that the minimum rank, 3.50, gets a 100 mark, and the maximum rank, 4.89, gets a 0. The last column, the stratum, is assigned according to the distance of the mark to either 70 or 30 or 0.

Table 3. Mapping main research results to the taxonomy; layers of the nodes affected; Tr – taxonomic ranks derived from them; Trn – taxonomic ranks normalized to the range 0 to 100; and three strata obtained by k-means partitioning of the ranks.

Scientist	Mapping to taxonomy	Layers	Tr	Trn	Stratum
S1	4.1.2.7, 5.2.1.2.7, 5.2.3.7.7	4,5,5	3.88	73	1
S2	2.1.1.2, 2.1.1.2, 5.2.2.7, 5.2.3.5, 5.2.3.5	4,4,4,4,4	3.50	100	1
S3	3.2.1.4.2, 5.2.1.2.3, 5.2.1.2.7, 5.2.3.5.4 , 5.2.3.7.6	5,5,5,5,5	4.50	29	2
S4	1.1.1.4.3, 3.4.4.5, 5.2.1.1.1,5.2.1.2.7, 5.2.3.2.1,5.2.3.7.8	5,4,5,5,5,5	3.90	71	1
S5	3.2.1.4.4, 3.2.1.4.4, 3.2.1.4.5, 3.2.1.4.6, 3.2.1.11.1	5,5,5,5,5	4.50	29	2
S6	3.1.1.5.2, 3.1.2.1.1, 3.1.2.1.1 , 3.2.1.6., 3.2.1.7	5,5,5,4,4	3.77	81	1
S7	5.2.3.5.6, 5.2.3.5.7	5,5	4.80	7	3
S8	3.2.1.3.1, 3.2.1.4.1, 5.2.3.3.1, 5.1.3.2.1, 5.1.3.2.4	5,5,5,5,5	4.50	29	2
S9	5.2.1.2.3, 5.2.3.3.2, 5.2.3.5.1, 5.2.3.5.4, 5.2.3.6.2	5,5,5,5,5	4.50	29	2
S10	5.2.3.3.2, 5.2.3.13.1	5,5	4.80	7	3
S11	3.2.1.2, 3.2.1.2.1,3.2.1.3.3,3.2.1.4.1, 3.2.1.7.2	4,5,5,5,5	3.86	74	1
S12	3.2.1.9.1.1,3.2.1.10,3.2.1.11.2,5.1.1.7.1, 5.2.3.1.3,5.2.3.4.1	6,4,5,5,5,5	3.86	74	1
S13	1.1.1.3, 5.2.1.2.1,5.2.1.2.1,5.2.2.7.1, 5.2.3.7.1	4,5,5,5,5	3.86	74	1
S14	3.2.1.3.1	5	4.90	0	3
S15	5.2.4.3.1	5	4.90	0	3
S16	5.2.4.2.3	5	4.90	0	3
S17	2.1.3.7.1, 5.2.4.3.1, 5.2.3.7.5., 5.2.1.2.4, 5.2.3.2.4, 5.2.3.7.3.2, 5.2.3.5.4., 5.2.4.3.1	5,5,5,5,6,5,5	4.39	36	2
S18	3.2.1.9.1,3.2.1.9.2,5.2.3.3.3.1	5,5,6	4.79	8	3
S19	3.2.1.7.5,3.2.1.9.3,5.2.3.2.1.1,5.2.4.5.1	5,5,6,5	4.69	15	3
S20	3.2.1.4.3,5.2.3.7.7,5.2.3.7.8.1	5,5,6	4.79	8	3
S21	1.1.1.6,2.1.1.2,2.1.1.8.3,3.2.1.6, 3.4.1.6,5.1.2.4,5.2.1.1.3	4,4,5,4,4,4,5	3.57	95	1
S22	3.2.1.2.2,5.2.1.2.7.1,5.2.3.1.2,5.2.3.6.2.1	5,6,5,6	4.78	9	3
S23	3.2.1.3,3.2.1.3.1,3.4.4.1	4,5,4	3.79	79	1
S24	2.1.5.3.1	5	4.90	0	3
S25	5.2.3.3.3.2, 5.2.3.8.1	6,5	4.89	1	3
S26	3.2.1.11.1,3.2.1.11.1,3.3.1.6,5.2.2.7, 5.2.3.5.6	5,5,4,4,5	3.77	81	1
S27	3.2.1.3.2,3.2.1.4.1,5.2.1.2.1,5.2.3.1.1	5,5,5,5	4.60	21	2
S28	3.2.1.8	4	3.90	71	1
S29	5.2.3.3.2.1,5.2.3.3.3.3,5.2.3.3.4	6,6,5	4.88	1	3
S30	5.1.3.2.1.1,5.2.1.2.7.2,5.2.3.3.5	6,6,5	4.88	1	3

7. Comparing Taxonomic Rank with Citation and Merit

We compared our taxonomic ranks with more conventional criteria: (a) Citation and (b) Merit. The Citation criterion was derived from Google-based indexes of the total number of citations, the number of works receiving 10 or more citations, and Hirsch index h, the number of papers receiving h citations or more. The merit criterion was computed from data on the following three indices: the number of successful PhDs (co)-supervised, the number of conferences co-organized, and the number of journals for which the researcher-scientist is a member of the Editorial Board.

To aggregate the indexes into a convex combination, that is, a weighted sum, automatically, a principled approach which works for correlated or inconsistent criteria has been developed. According to this approach, given the number of strata (in our case 3), the aggregate criterion is to be found so that its direction in the index space is such that all the observations are projected into compact well-separated clusters along this direction (Mirkin and Orlov, 2013, 2015).

To be more specific, consider a data matrix scientist-to-criteria $\mathbf{X} = (x_{ij})$ where $i = 1,...,N$ are indices of scientists, $j = 1,...,M$ are indices of M criteria, and x_{ij} is the score of jth criterion for the ith scientist. Let us consider a weight vector $\mathbf{w} = (w_1, w_2, ..., w_M)$ such that $w_j \geq 0$ for every j and $\sum_j w_j = 1$, for the set of criteria. Then the combined criterion is $\mathbf{f} = \sum_{j=1}^{M} w_j x_j$ where x_j is jth column of matrix $\mathbf{X}$. The problem is to find K disjoint subsets $S = \{S_1, ..S_k, ..., S_K\}, k = 1, ..., K$ of the set of indices i, referred to as strata, according to values of the combined criterion $\mathbf{f}$. Each stratum k is characterized by a value of the combined criterion c_k, the stratum center. Geometrically, strata are formed by layers between parallel planes in the space of criteria. At any stratum S_k, we want the value of the combined criterion $f_i = \sum_{j=1}^{M} w_j x_{ij}$ at any $i \in S_k$ to approximate the stratum center c_k. In other words, in the equations $x_{i1}w_1 + x_{i2}w_2 + ... + x_{iM}w_M = c_k + e_i$, e_i are errors to be minimized over vector $\mathbf{w}$. A least-squares formulation of the linear stratification (LS) problem: find a vector $\mathbf{w}$, a set of centers $\{c\}$ and a partition S to solve the problem in (1), as follows.

$$\begin{aligned} \min_{w,c,S} \quad & \sum_{k=1}^{K} \sum_{i \in S_k} (\sum_{j=1}^{M} x_{ij} w_j - c_k)^2 \\ \text{such that} \quad & \sum_{j=1}^{M} w_j = 1 \\ & w_j \geq 0, j \in 1...M. \end{aligned} \tag{1}$$

This problem can be tackled using the alternating minimzation approach, conventional in cluster analysis. For any given weight vector $\mathbf{w}$, the crite-

Table 4. Scores of two criteria, x and y, over 8 scientists labeled, for convenience, by using an uppercase notation of the corresponding strata (see Figure 7).

Label	Criterion x	Criterion y
C1	2	0
C2	0	1
B1	6	0
B2	5	0.5
B3	3	1.5
B4	1	2.5
A1	4	2
A2	2	3

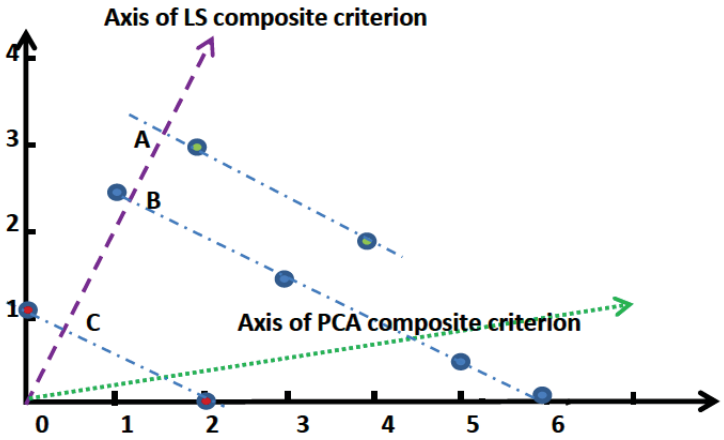

Figure 7. Eight scientists on the plane of criteria x and y. The LS and PCA combined criteria are represented with broken lines.

rion in (1) is just the conventional square-error clustering criterion of the K-means clustering algorithm over a single feature, the combined criterion $\mathbf{f} = \sum_{j=1}^{M} w_j x_j$. Finding an appropriate $\mathbf{w}$ at a given stratification S can be reached by using standard quadratic optimization software.

To illustrate the approach as it is and, also, its difference from the widely used Principal Component Analysis (PCA) approach to linearly combining criteria, let us consider the following example. In Table 4, scores of two criteria over 8 scientists are presented.

Although usually criteria values are normalized into a 0–100% scale, we do not do that here to keep things simple. It appears, the data ideally, with zero error, fall into three strata, $K = 3$, as shown in Figure 7, according to combined criterion $\mathbf{f} = \frac{1}{3}x + \frac{2}{3}y$. In contrast, the PCA based linear combi-

Table 5. Scores of two combined criteria, the LS based and PCA based.

Label	LS	PCA
C1	0.67	1.54
C2	0.67	0.23
B1	2.00	4.63
B2	2.00	3.97
B3	2.00	2.66
B4	2.00	1.34
A1	2.67	3.54
A2	2.67	2.23

Table 6. Pairwise correlations between criteria (only the part above the diagonal is shown).

Criterion	Citation	Merit
TR	-0.12	-0.04
Citation		0.31

nation, $\mathbf{z} = 0.7712\mathbf{x} + 0.2288\mathbf{y}$, admits a residual of 13.4% of the total data scatter, and leads to a somewhat different ordering, at which two top stratum scientists get lesser aggregate scores than two scientists of the B stratum.

For convenience, the combined criteria scores are presented in Table 5. In the aggregated Citation criterion, the Hirsch index received a zero coefficient, while the other two were one half each. The zeroing of the Hirsch index weight is in line with the overwhelming critiques this index has been exposed to in recent times, (Albert, 2013; Osterloh and Frey, 2014; Dora, 2013; Van Raan, 2006). A similarly aggregated Merit criterion is formed with weights 0.22 for the number of PhD students, 0.10 for the number of conferences, and 0.69 for the number of journals, which is consistent with the prevailing practice of maintaining a heavy and just submission reviewing process in leading journals.

To compare these scales, let us compute Pearson correlation coefficients between them, see Table 6. As expected, the Citation and Merit criteria do not correlate with the Taxonomic rank of the scientists. On the other hand, the traditional Citation and Merit criteria are somewhat positively correlated, probably because they both relate to the popularity of a scientist.

8. Conclusions

Assessments can be carried out at different levels, a region, an organization, a team or an individual researcher; within a domain or inter domains. What we can metaphorically express as wider horizons, are brought to our attention, through analysis of quality. Among the recommendations aris-

ing from this work, on the regional level, there are three on the particular subjects of our concern:

- Set out a more structured and strategic process for proposing projects.
- Conduct a systematic analysis of the existing infrastructure.
- Take a more systematic approach to evaluating the impact of operational projects.

With these recommendations, we are emphasizing the importance of these underpinning themes. These themes, and their underpinnings, should be pursued assertively for journals and other scholarly publishing, and also for research funding programmes.

We both observe and demonstrate that evaluation of research, especially at the level of teams or individuals can be organized by, firstly, developing and maintaining a taxonomy of the relevant subdomains and, secondly, a system for mapping research results to those subdomains that have been created or significantly transformed because of these research results. This would bring a well-defined meaning to the widely-held opinion that research impact should be evaluated, first of all, based on qualitative considerations. Further steps can be, and should be, undertaken in the directions of developing and maintaining a system for assessment of the quality of research across all areas of knowledge. Of course, developing and/or incorporating systems for other elements of research impact, viz., knowledge transfer, industrial applications, social interactions, etc., are to be taken into account also. In comprehensively covering quality and quantitative research outcomes, there can be distinguished at least five aspects of an individual researcher's research impact:

- Research and presentation of results (number, quality)
- Research functioning (journal/volume editing, running research meetings, reviewing)
- Teaching (knowledge transfer, knowledge discovery)
- Technology innovations (programs, patents, consulting)
- Societal interactions (popularization, getting feedback)

Many, if not all, of the items in this list can be maintained by developing and using corresponding taxonomies. The development of a system of taxonomies for the health system in the USA, IHTSDO SNOMED CT (SNOMED CT, 2016), extended now to many other countries, and languages, should be considered an instructive example of such a major undertaking.

This suggests directions for future work. Among them are the following.

In methods: (i) Enhancing the concept of taxonomy by including theoretical, computational, and industrial facets, as well as dynamic aspects to it; (ii) Developing methods for relating paper's texts, viz. content, and taxonomies; (iii) Developing methods for taxonomy building using such research paper texts, i.e. content; (iv) Developing methods for mapping research results to taxonomy units affected by them; (v) Using our prototyping here, developing comprehensive methods for ranking the impact of results to include expert-driven components; (vi) Also based on our prototyping here, developing accessible and widely used methods for aggregate rankings.

In substance: (i) Developing and maintaining a permanent system for assessment of the scope and quality of research at different levels; (ii) Developing a system of domains in research subjects and their taxonomies; (iii) Cataloguing researchers, research and funding bodies, and research results; (iv) Creating a platform and forums for discussing taxonomies, results and assessments.

A spin-off of our very major motivation for qualitative analytics is to propose using a full potential of the research efforts on a regional level. In our journal editorial roles, we realize very well that sometimes quite predictable rejection of article submissions can raise such questions as the following: is there no qualitative interest at all in such work? How can, or how should, improvement be recommended? At least as important, and far more so in terms of wasteful energy and effort, is the qualitative analysis of rejected research funding proposals. (As is well known, a relatively small proportion of the research projects gets a "go ahead" nod. For example, The European Horizon 2020 FET-Open, Future Emerging Technologies, September 2015 proposal submission resulted in less than a 1.4% rate (Hallantie, 2016): 11 successful research proposals out of 800 proposal submissions.) Given the workload at issue, on various levels and from various vantage points, there is potential for data mining and knowledge discovery in the vast numbers of rejected research funding proposals. Ultimately, and given the workload undertaken, it is both potentially of benefit, and justified, to carry out such analytics.

References

ABRAMO, G., CICERO, T., ANGELO, C.A. (2013), "National Peer-Review Research Assessment Exercises for the Hard Sciences Can Be a Complete Waste Of Money: The Italian Case", *Scientometrics, 95(1)*, 311–324.

ACM (2012), The 2012 ACM Computing Classification System, https://www.acm.org/publications/class-2012.

ALBERT, B. (2013), "Impact Factor Distortions", *Science, 340(6134)*, 787.

ARAGNÓN, A.M. (2013), "A Measure for the Impact of Research", *Scientific Reports, 3, Article number: 1649.*

BERNERS-LEE, T. (2010), "Long Live the Web", *Scientific American, 303(6)*, 80–85.

BLEI, D.M., NG, A.Y., JORDAN, M.I., and LAFFERTY, J. (2003), "Latent Dirichlet Allocation", *Journal of Machine Learning Research, 3*, 993–1022.

CANAVAN, J., GILLEN, A., and SHAW, A. (2009), "Measuring Research Impact: Developing Practical and Cost-Effective Approaches", *Evidence and Policy: A Journal of Research, Debate and Practice, 5.2*, 167–177.

DORA (2013). San Francisco Declaration on Research Assessment (DORA), http://www.ascb.org/files/SFDeclarationFINAL.pdf.

EISEN, J.A., MACCALLUM, C.J., and NEYLON, C. (2013), "Expert Failure: Re-Evaluating Research Assessment", *PLoS Biology, 11(10): e1001677.*

ENGELS, T.C., GOOS, P., DEXTERS, N., and SPRUYT, E.H. (2013), "Group Size, h-Index, and Efficiency in Publishing in Top Journals Explain Expert Panel Assessments of Research Group Quality and Productivity", *Research Evaluation, 22(4)*, 224–236.

HALLANTIE, T. (2016), "What It Takes to Succeed in FET-Open", https://ec.europa.eu/digital-single-market/en/blog/what-it-takes-succeed-fet-open.

HICKS, D., WOUTERS, P., WALTMAN, L., DE RIJCKE, S., and RAFULS, I. (2015), "The Leiden Manifesto for Research Metrics". *Nature, 520*, 429–431.

LEE, F.S., PHAM, X., and GU, G. (2013), "The UK Research Assessment Exercise and the Narrowing of UK Economics", *Cambridge Journal of Economics, 37(4)*, 693–717.

METRIC TIDE (2016), "The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management", http://www.hefce.ac.uk/pubs/rereports/Year/2015/metrictide/Title,104463,en.html.

MIRKIN, B. (2013), "On the Notion of Research Impact and Its Measurement", Institute of Control Problems, Moscow (in Russian), *Control in Large Systems, Special Issue: Scientometry and Experts in Managing Science, 44*, 292–307.

MIRKIN, B., and ORLOV, M. (2013), "Methods for Multicriteria Stratification and Experimental Comparison of Them", Preprint (in Russian) WP7/2013/06, Higher School of Economics, Moscow, 31 pp.

MIRKIN, B., and ORLOV, M. (2015). "Three Aspects of the Research Impact by a Scientist: Measurement Methods and an Empirical Evaluation", in *Optimization, Control, and Applications in the Information Age*, eds. A. Migdalas, and A. Karakitsiou, Springer Proceedings in Mathematics and Statistics, 130, pp. 233–260.

MURTAGH, F. (2008), "Editorial", *The Computer Journal, 51(6)*, 612–614.

MURTAGH, F. (2010), "The Correspondence Analysis Platform for Uncovering Deep Structure in Data and Information", *The Computer Journal, 53(3)*, 304–315.

NG, W.L. (2007), "A Simple Classifier for Multiple Criteria ABC Analysis", *European Journal of Operational Research, 177*, 344–353.

ORLOV, M., and MIRKIN, B. (2014), "A Concept of Multicriteria Stratification: A Definition and Solution", *Procedia Computer Science, 31*, 273–280.

OSTERLOH, M., and FREY, B.S. (2014), "Ranking Games", *Evaluation Review*, Sage, pp. 1–28.

RAMANATHAN, R. (2006), "Inventory Classification with Multiple Criteria Using Weighted Linear Optimization", *Computers and Operations Research, 33*, 695–700.

SCHAPIRE, R.E. (1990), "The Strength of Weak Learnability", *Machine Learning, 5(2)*, 197–227.

SIDIROPOULOS, A., KATSAROS, D., and MANOLOPOULOS, Y. (2014), "Identification of Influential Scientists vs. Mass Producers by the Perfectionism Index", Preprint, ArXiv:1409.6099v1, 27 pp.

SNOMED CT (2016), IHTSDO, International Health Terminology Standards Development Organization, SNOMED CT, Systematized Nomenclature of Medicine, Clinical Terms, http://www.ihtsdo.org/snomed-ct.

SUN, Y., HAN, J., ZHAO, P., YIN, Z., CHENG, H., and WU, T. (2009), "RankClus: Integrating Clustering with Ranking for Heterogeneous Information Network Analysis", *EDBT '09 Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology*, New York: ACM, pp. 565–576.

THOMSON REUTERS (2016), "Thomson Reuters Intellectual Property and Science", (Acquisition of the Thomson Reuters Intellectual Property and Science Business by Onex and Baring Asia Completed, Independent business becomes Clarivate Analytics), http://ip.thomsonreuters.com.

UNIVERSITY GUIDE (2016), "The Complete University League Guide", http://www.thecompleteuniversityguide.co.uk/league-tables/methodology.

VAN RAAN, A.F. (2006). "Comparison of the Hirsch-index with Standard Bibliometric Indicators and with Peer Judgment for 147 Chemistry Research Groups". *Scientometrics, 67(3)*, 491–502.