A Comprehensive Analysis of the First Ten Editions of the WEBIST Conference

Giseli Rabello Lopes[1(✉)], Bernardo Pereira Nunes[2,3], Luiz André P. Paes Leme[4], Terhi Nurmikko-Fuller[5], and Marco A. Casanova[2]

[1] Computer Science Department, Federal University of Rio de Janeiro, Rio de Janeiro, RJ, Brazil
giseli@dcc.ufrj.br
[2] Department of Informatics, Pontifical Catholic University of Rio de Janeiro, Rio de Janeiro, RJ, Brazil
{bnunes,casanova}@inf.puc-rio.br
[3] Department of Applied Informatics, UNIRIO, Rio de Janeiro, RJ, Brazil
[4] Computer Science Institute, Fluminense Federal University, Niterói, RJ, Brazil
lapaesleme@ic.uff.br
[5] Oxford E-Research Centre, Oxford University, Oxford OX1 3QG, UK
terhi.nurmikko-fuller@oerc.ox.ac.uk

Abstract. An analysis of the proceedings of the first decade of the WEBIST conference, in terms of social networking and statistical analyses, as well as bibliometrics, unearthed information regarding existing patterns in the prevalent themes and topics of the conference, shedding light on the development of the event and its community as they grew and matured. In addition to the findings of this analysis we present a queriable Web-based application, which draws from a dataset of RDF triples, enabling the recreation of the examined patterns and the further exploration of the proceedings data.

Keywords: Conference analysis · Statistical analysis · Bibliometrics · Social network analysis · Webist analysis · Linked data

1 Introduction

Knowing about the past makes the present easier to understand, enables us to make predictions about the future, and helps guide us towards appropriate actions and correct decisions. However, with the perpetual flooding of newly available information, it became increasingly difficult to keep up to date, as well as to interpret and analyse data in meaningful ways. In response, there have been rapid technological advancements to support data analysts in both handling large amounts of data and in decision-making. In the commercial sector, companies and organisations relied on such analyses to overcome competitors, to improve customer relations and to identifying specific needs. In academia, data analysis has also been useful, helping to solve and uncover a number of

V. Monfort et al. (Eds.): WEBIST 2015, LNBIP 246, pp. 252–274, 2016.
DOI: 10.1007/978-3-319-30996-5_13

problems in domain as diverse as Health, Management, Marketing, Engineering and Computer Science [1].

Data analysis has been previously used to detect features such as related research groups, topics of interest, impact of authors and publications in a given field. Among others, an analysis of a group of four conferences in the Human-Computer Interaction (HCI) domain was conducted by Henry et al. [2]. Based on publication metadata (such as authors and keywords), it provided valuable insights into authors' behaviours and research topics investigated in HCI over the last two decades. Blanchard [3] presented a decade-long longitudinal study, which analysed the potential of cultural biases on the Intelligent Tutoring Systems (ITS) and Artificial Intelligence in Education (AIED) strands of the American Psychology Association (APA). Chen et al. [4] presented a visual analytic approach to identify co-citation clusters, classified and used to understand how astronomical research evolved between 1994 and 1998. Another example along the same lines was conducted by Gasparini et al. [5], who were able to identify central authors, institutions, important trends and topics in the HCI field. As for Information Systems (IS), Posada and Baranauskas [6] analysed a sister-event called International Conference on Enterprise Information Systems (ICEIS), and built a roadmap of the IS domain based on paper titles and authors from the last three years in ICEIS and the last eight years of selected papers published in a Springer series on IS. Chen et al. [7] performed a citation analysis of all papers published in the International Conference on Conceptual Modeling (ER) between 1979 and 2005. These analyses opened up a wide range of new research agendas and trends, as well as showing the value of a domain's introspective analysis.

Zervas et al. [8] presented a study on research collaboration patterns via co-authorship analysis in Technology-enhanced Learning fields. Similar analyses were conducted by Procopio Jr. et al. [9] for Databases fields and by Cheong and Corbitt [10] for IS (analysing the Pacific Asia Conference on IS). The analysis of co-authorships in research communities can reveal strong research groups in the area and also enable the creation of links between different groups.

We present an in-depth analysis of the first ten editions (2005–2014) of the WEBIST conference. So far, it attracted 2,867 researchers and professionals from several institutions, as well as published 1,449 papers, which in turn are being cited. The conference currently has five main tracks: *Internet Technology*, *Web Interfaces and Applications*, *Society, e-Business and e-Government*, *Web Intelligence* and *Mobile Information Systems*.

The analysis presented in this paper relies on techniques borrowed from social network analysis [11], bibliometrics and traditional statistical measures. In addition to presenting these analyses, we published the results in a format where they can be replicated and reused in further analysis. For this, we borrowed Batista and Loscio's approach [12] and used Linked Data (LD) principles. We also created a Web-based application that enables users to interactively explore data through a SPARQL endpoint.

In this paper, Sect. 2 overviews metrics and measures used in the analysis. Section 3 details the extraction, enrichment and publication process of raw WEBIST data into RDF data and presents a visualisation tool specifically created to manipulate and possibly assist users in finding new research groups, topics and insights. Section 4 presents several analysis conducted with the WEBIST tool. Finally, Sect. 5 concludes the work with remarks and future directions.

2 Background

This section provides the necessary background information required to understand the analysis conducted with the data. We review metrics and methods of statistical analysis, social network analysis and bibliometric indices.

2.1 Classical Statistical Measures

Standard deviation (σ) is a common measure of dispersion used to describe the central tendency of a distribution. Standard deviation [13] is defined as the square root of its variance, as shown in Eq. 1. Thus, considering a population X of N data points x_i, having average $\bar{X}$, σ is defined as:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{X})^2}, \quad \text{where} \quad \bar{X} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{1}$$

Note that a low σ value indicates that the data points has a high central tendency, i.e., tend to be very close to the average, whereas a high σ value indicates that the data points are dispersed over a large range of values.

The *Pearson's correlation coefficient* [14], often denoted by the letter r, measures the strength and direction of the linear correlation between two variables X and Y. Pearson's coefficient (see Eq. 2) is defined as the covariance of the variables divided by the product of their standard deviations to measure their dependence:

$$r = \frac{\sum_{i=1}^{N}(x_i - \bar{X})(y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{X})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{Y})^2}} \tag{2}$$

An r value between $+1$ and -1 indicates the degree of linear dependence between X and Y: r=1 indicates a total positive correlation between the two variables; and r=-1 indicates a total negative (inverse) correlation. For instance, as X values increase, Y values linearly decrease.

The *Lorenz curve* [15] represents the cumulative distribution of a probability density function. Such a function is built as a ranking of the members of the population disposed in ascending order of the amount being studied. The percentage of individuals is plotted on the x-axis and the percentage of the variable values on the y-axis. The distribution is perfectly equalitarian when every individual has the same variable value; a 45-degree line represents the perfect equality. On the other hand, the perfectly unequal distribution is that in which only one

individual has all the variable value, the curve is $y = 0$ for all $x < 100\,\%$, and $y = 100\,\%$ when $x = 100\,\%$, known as the perfect inequality line. This curve was initially created to study the social inequality of wealth and income distributions for a population, but it can be applied to analyse other distributions [16]. We used the Lorenz curve (Sect. 4) to study the distribution of papers by author.

The *Gini coefficient* [15] is a measure of statistical dispersion indicating the inequality among values of a frequency distribution. It is graphically represented as the area between the perfect equality line and the observed Lorenz curve.

The *Robin Hood index* [17], also called Hoover index, is used to measure the fraction of the total variable value that must be redistributed over the population to become a uniform distribution. It is graphically represented as the longest vertical distance between the Lorenz curve and the perfect equality line.

2.2 Social Network Analysis

Before introducing social network metrics and concepts [11,18–22], we recall that we may represent a social network as a graph $G = (N, E)$, where N is the set of nodes, where $n_i \in N$ represents an actor of the network, and E is the set of edges, where $e_i \in E$ represents a relational tie between a pair of actors.

The *Density* of a graph is defined as the number of the existing edges of the graph, divided by the maximum number of edges the graph can have. A density value equal to 1 indicates an entirely connected network, while 0 indicates a disconnected network. Considering an undirected graph, where the possible number of connections between each two nodes is 1, the density is defined as:

$$D = \frac{2|E|}{|N|\,(|N| - 1)} \tag{3}$$

where $|E|$ is the cardinality of the set of edges and $|N|$ is the cardinality of the set of nodes.

Modularity is a measure of the structure of networks and estimates the strength of division of a network into communities (groups). It is often used in optimisation methods for detecting community structure in networks. A high modularity value indicates a network having dense connections between the nodes within the communities, but sparse connections between nodes in different communities. Modularity is defined as [23]:

$$Q = \sum_i (e_{ii} - a_i^2) \tag{4}$$

where e_{ij} is the number of edges connecting nodes from the community i to nodes from the community j; $a_i = \sum_j e_{ij}$ is the number of edges with at least one node from the community i. Each edge contributes only once to the count (the contribution must be divided by half, one halve for e_{ij} and the other for e_{ji}).

A *Connected Component* of an undirected graph is a subgraph in which any two nodes are connected to each other by paths, and in which their nodes are not connected to any other nodes in the supergraph.

A *Giant Component* of a graph (also named *main component*) is the connected component which contains most of the nodes in the graph.

The *Giant Coefficient* of a graph is based on the size of the giant component G' of a graph G. It is defined as the number of nodes N' in the giant component divided by the total number of nodes N in the entire graph:

$$GC = \frac{|N'|}{|N|}, \quad \text{where} \quad N' \subseteq N \tag{5}$$

Diameter is associated with graph distance. It is defined as the maximum value among all shortest paths between two nodes of the graph (i.e., the longest distance between any pair of nodes belonging to the graph).

The *Average Clustering Coefficient* is a measure of the degree to which nodes in a graph tend to cluster together (connectivity of neighbours). It is defined as the average of the clustering coefficients of all the nodes in the graph:

$$\bar{C} = \frac{1}{|N|} \sum_{i=1}^{|N|} C_i \tag{6}$$

where C_i is the clustering coefficient of a node n_i and is calculated as the number of existing edges between the direct neighbours of n_i divided by the total number of possible edges directly connecting all neighbours of n_i.

2.3 Bibliometric Indices

This section introduces two common bibliometric indices often used to measure the impact, in terms of popularity, of researchers, scientific publications, conferences and journals.

The *h-index* was proposed to measure both the number of publications and the number of citations per publication of a scientist. According to Hirsch [24], a scientist has index h if h of his/her N_p papers have at least h citations each, and the other $(N_p - h)$ papers have no more than h citations each. This index is also applied to estimate the productivity and impact of conferences.

The *i10-index* indicates the number of publications of a scientist having at least ten citations[1].

3 WEBIST Workflow - from Raw to RDF Data

3.1 Overview of the Process

This section overviews the process of data acquisition, involving extraction, enrichment, preparation and consolidation, adopted to create the *WEBIST Dataset* and its use by the *WEBIST Analytics* tool. Figure 1 depicts the whole process.

[1] http://googlescholar.blogspot.com.br/2011/.

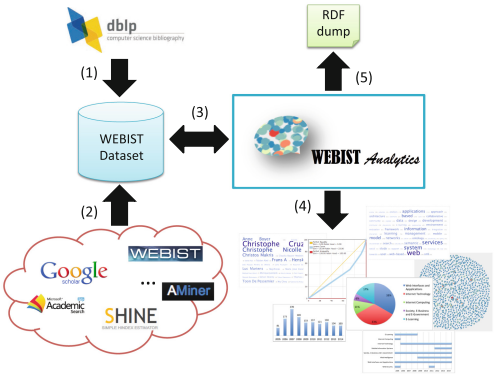

Fig. 1. WEBIST workflow.

Initially, we created an interlinked open dataset, called *WEBIST Dataset*, available in RDF, following the Linked Data principles [25], about the 10 editions of WEBIST conference. This dataset was created by aggregating data extracted from different data sources. The initial core of the data about WEBIST was extracted from DBLP (Digital Bibliography & Library Project)[2] (Step 1). Then, the data was enriched using data crawled from different Web sources such as Google Scholar Citations[3] (Step 2).

Based on the information loaded in the *WEBIST Dataset* (Step 3), the proposed Web application, called *WEBIST Analytics*, provides different functionalities such as exploratory search, and several analysis over the data, presented through different graphical visualisations (Step 4).

Moreover, using the *WEBIST Analytics* interface, the RDF dump of the *WEBIST Dataset* is available for download (Step 5). The *WEBIST Dataset* creation and *WEBIST Analytics* functionalities are detailed in the next subsections.

3.2 WEBIST Dataset

Data Acquisition. Over the last ten years, WEBIST conference data, such as paper acceptance or organisation committee, was published. Thus, to create a tool to seamlessly make sense of the data, we aggregated data extracted from different data sources, being aware of the possible necessity of initially submitting the data to deduplication [26] techniques.

The initial core of the data about WEBIST was extracted, in December 2014, from DBLP, a digital library about computer science publications. We were not able to find an updated source of DBLP data in RDF format (containing all editions of WEBIST conference). Thus, we had to extract the data directly from the XML version of DBLP available. This XML data also contained information

[2] http://www.informatik.uni-trier.de/~ley/db/.
[3] http://scholar.google.com/citations.

about the name disambiguation of the authors (different spellings of the name representing the same author in XML version of DBLP). Thus, the authors name disambiguation [27] was facilitated in this initial core. In summary, we collected information about the published papers and authors of WEBIST, reaching a total of 1,449 papers and 2,867 authors.

Data Enrichment. Data enrichment serves as a means to extending the initial data from additional data sources. For this, we developed a focused crawler to obtain this additional information. In this step, information from Google Scholar Citations and Google Scholar were used to obtain bibliometric indices of WEBIST authors. Specifically, the key of authors in Google Scholar Citations and the authors indices (h-index, i10-index and number of citations) were extracted from Google Scholar[4] and Google Scholar Citations, respectively. The crawling process used the name of the authors to perform the searches. Using this strategy, 748 authors profiles were found in Google Scholar Citations, representing 26.09 % of the total WEBIST authors. Other complementary information about some publications citations was crawled from Google Scholar. We collected the number of citations for the presumed most cited papers. The candidates to be most cited papers were obtained by the topmost ranked WEBIST papers presented in SHINE (Simple H-INdex Estimator)[5], Arnetminer[6] and Microsoft Academic Search[7]. Additional information about the main research areas and program committee (members and their affiliations) of each edition of WEBIST were extracted from each conference Web site[8] Moreover, other information about each conference edition, such as location, number of submissions, number of countries with submissions and paper acceptance rates (for full papers and oral presentations), were extracted from the forewords of the WEBIST proceedings available at SCITEPRESS digital library[9].

Data Transformation. Another crucial step is data transformation, carried out after data acquisition involving the preparation and enrichment steps, requiring a common format for the data. For this, we followed the Linked Data principles [25] that encourage data publishers to expose their data through HTTP mechanism and to use RDF as the data description language. According to these guidelines, the publishers should name things using HTTP URIs and provide appropriate clipping of data in RDF when users follow the URIs. All the data about WEBIST, obtained in the two previous steps, were first loaded in a relational database. After that, we used a relational-to-RDF framework (D2RQ) [28] that dynamically transforms relational data into RDF graphs. It provides an HTML browser for relational databases as well as a SPARQL interface to

[4] http://scholar.google.com.
[5] http://shine.icomp.ufam.edu.br.
[6] http://arnetminer.org/.
[7] http://academic.research.microsoft.com/.
[8] http://www.webist.org/ [*2005-2011:* WEBIST*$year$*; *2012-2014:*?y=*$year$*].
[9] http://www.scitepress.org.

query the database. This framework also provides a mapping language to define rules for transforming relational data and schema into RDF graphs.

Data Publication. The successful completion of these previous steps ensured that the dataset was available to others (both in terms of users and applications) that want to use it for different purposes. The RDF dump of the *WEBIST dataset* is available for download from the *WEBIST Analytics* interface.

3.3 WEBIST Analytics Application

WEBIST Analytics, a Web-based application, was created to provide multiple perspectives of the data produced by WEBIST conferences over the 10 editions. In addition to providing the *WEBIST dataset*, the proposed application is also composed of analytics tools, graphical visualisations and a simple search engine that assists users in finding, uncovering and making sense of the information available. *WEBIST Analytics* application can be accessed at: http://lab.ccead.puc-rio.br/webist_analytics/.

Based on the information loaded in the *WEBIST Dataset*, the proposed Web application provides different functionalities as both exploratory search and several analyses over the data, presented through different graphical visualisations. Free text search is available over two different WEBIST graphs, the co-authorships graph (among authors) and a more complete graph composed by co-authorships and authoring relations (among authors and publications). It allows users to search and retrieve related information about WEBIST conferences, including an interactive visualisation of networks. Other exploratory search is allowed via tag cloud visualisations. In this case, the terms in the tag cloud can be selected and the associated publications retrieved, which in turn assists users in finding papers related to each research topic.

4 Analysis and Results

This section presents and discusses the results of the analysis available in *WEBIST Analytics*. We observe that the results reported in this section were computed using the methods and metrics presented in Sect. 2.

4.1 WEBIST Overview

Table 1 overviews the last ten editions of the WEBIST conference with respect to the paper acceptance rate and the venue information. Since the first edition of the WEBIST conference, the full paper acceptance rate decreased and became stable under 15 % of all submitted papers. The low number of full papers accepted by WEBIST may suggest the level of rigorousness of the reviewers as well as the level of quality expected by the conference. On the other hand, the high acceptance rate for short papers (see oral presentations rates) may indicate an inclination

of WEBIST towards bringing together researchers with work in progress and researchers with consolidated work, possibly offering opportunities for knowledge transfer and discussion.

In addition to the paper acceptance rate, Table 1 provides information about the location of each WEBIST edition. Note that although WEBIST is an international conference, with the exception of its first edition that took place in USA, all editions were held in Europe, mostly Spain and Portugal. As the number of submitted papers from all over the world has roughly remained the same, independently of where the conference took place (USA, Germany, Netherlands, Spain or Portugal), the change of place could bring extra benefits such as new collaborations with local universities and researchers.

Table 1. Conference stats.

Year	Location	#submitted papers	#countries with submissions	% of accepted papers	
				Full papers	Oral pres.[a]
2005	Miami, USA	110	37	22 %	49 %
2006	Setubal, Portugal	218	more than 40	16 %	50 %
2007	Barcelona, Spain	367	more than 50	14 %	44 %
2008	Funchal, Madeira, Portugal	238	more than 40	13 %	40 %
2009	Lisbon, Portugal	203	47	13 %	36 %
2010	Valencia, Spain	205	46	12 %	36 %
2011	Noordwijkerhout, Netherlands	156	43	9 %	33 %
2012	Porto, Portugal	184	41	13.6 %	44.6 %
2013	Aachen, Germany	143	43	19 %	39.9 %
2014	Barcelona, Spain	153	49	15.03 %	41.83 %

[a]Oral presentation including full papers and short papers.

4.2 General Analysis

An initial analysis of all WEBIST conferences was conducted with regard to its authors and publications. In this analysis we gathered 1,449 publications, which included all full papers, short papers, posters and selected papers. Figure 2 depicts the distribution of the papers over the conference editions. The number of accepted papers reached its peak in 2007, where 270 papers were accepted to a single conference, a figure almost twice the average number of papers accepted to other editions. This peak number of publications may be an indication of the rapid increase in the popularity of WEBIST and its reaching a certain level of maturity over the years, settling on a stable conference-size and community.

A rough analysis of the community can be carried out based on the number of authors of a scientific publication. The number of authors of a paper gives us a hint of the average size of the community and research groups. Across the 10

editions of WEBIST, there have been contributions from 2,867 authors, which gives an average of 2.91 authors per publication (with a standard deviation (σ) of 1.35, the maximum number of authors being 14 per paper and the minimum 1). Figure 3 shows the distribution of the average number of authors per year.

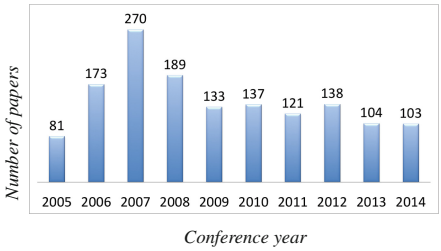

Fig. 2. Number of papers published per year.

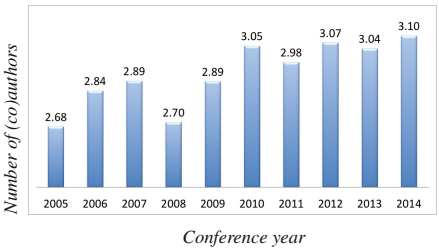

Fig. 3. Average number of (co)authors per paper over the conference years.

The list of topmost authors of WEBIST may reveal not only prolific authors, but possible experts and supporters for future editions of the conference. The engagement of researchers in a specific community could be initially measured by the number of papers they have had accepted in the earlier editions of the conference. The assumption is that, if they had over a specific number of papers, they might be eligible to make part of the program committee. After 10 editions, a total of 29 authors had more than 6 papers. The most active researcher had 15 published papers and the second had 12 papers. Figure 4 shows the top authors as a tag cloud[10]. The size of the names represents how active a research is in the WEBIST conference.

Figure 5 presents the Lorenz curve[11] along with an analysis based on the Gini coefficient and the Robin Hood Index (see Sect. 2). The Gini coefficient resulted in 25.99 % of inequality, while the Robin Hood Index was 23.06 %. The results show that the Lorenz Curve is closer to the equality than to the inequality line. This is an expected result for peer-reviewed conferences, where only high quality papers are accepted for publication. Although a few authors have more than 6 papers in WEBIST editions, the Lorenz Curve and the Robin Hood Index show that no redistribution is necessary, i.e., there is no bias in accepting papers from a research group or another, but simply merit. A high Robin Hood Index would indicate a possible need for further analysis in some publications.

4.3 Co-Authorships Network

Social Network Analysis (SNA) techniques were applied to the obtained information about the co-authorships in the WEBIST conference. The analysis was

[10] http://tagcrowd.com.

[11] http://www.peterrosenmai.com/lorenz-curve-graphing-tool-and-gini-coefficient-calculator.

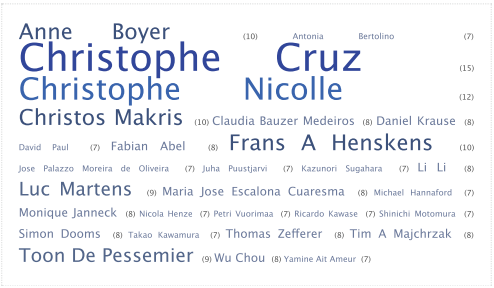

Fig. 4. Top authors with more than 6 papers.

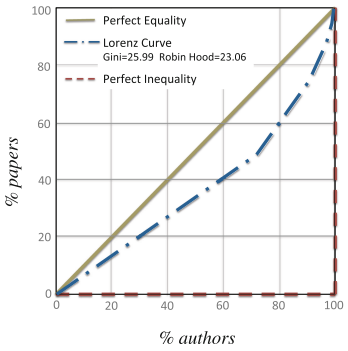

Fig. 5. Lorenz curve for the number of papers per author distribution.

conducted over an undirected graph G (defined in Sect. 2), where the nodes represent the authors and the edges represent a co-authorship between researchers. The WEBIST co-authorships network is comprised of 2,867 authors and 4,235 pairs of authors (edges) having at least one co-authored paper.

Table 2 shows an analysis of the co-authorship network using SNA measures. The analysis considers all WEBIST authors in the last 10 years. Briefly, we have:

- *Average Degree* shows that the authors, on the average, have co-authored papers with 2.9 other authors.
- *Density* shows a low proportion of co-authorships in the network relative to the total number possible (situation where all authors co-authored at least one paper with all others), only 0.1 %. It represents a weakly connected network. This shows an expected result in a conference network, where there are different groups of authors working in different papers. The measured modularity and the number of communities, as explained below, can reinforced this result.
- *Modularity* shows a high value representing the strength of division of the network into modules (also called groups, clusters or communities). Thus, WEBIST co-authorships network has co-authorships between the authors within the communities but none between authors in different communities.
- *Number of Communities* detected based on the modularity, was 803, being exactly the same as the **Number of Connected Components**. This shows that, in the analysed network, there are isolated communities that have not co-authorships in WEBIST with the authors of the other communities.

The following analysis takes into account only the giant component of the WEBIST network. Again, briefly, we have:

- *Giant Coefficient* represents the percentage of authors in the Giant Component of the WEBIST co-authorships network, being approximately 1.57 % (45 authors) of the total number of authors that published in all WEBIST conferences. These authors have 108 co-authorships between them (2.55 % of the

total possible co-authorships, i.e., if each of these authors co-authored on at least one paper with all others).
- *Diameter* represents the longest of all the shortest paths between two authors in the Giant Component, being estimated as 8. This shows that the farthest authors in the Giant Component have more than six degrees of separation, based on co-authorship in WEBIST papers. This reveals that the Giant Component probably results from a hierarchical structure, which is natural when research groups of different institutions are involved. The different research groups (subgroups) are connected by "hub" authors (probably research group leaders or professors) that collaborate in different research projects amongst the subgroups, while some researches (probably students) developed more specific tasks (sometimes related to only one paper).
- *Clustering Coefficient* measures the average degree to which authors in the network tend to cluster together, being approximately 93.4 %. This shows that many authors belonging the Giant Component worked with other authors that also worked together in at least one paper.

Table 2. Social networks analysis from the WEBIST co-authorships network.

Measure	Value
Average Degree	2.954
Density	0.001
Modularity	0.995
Number of Communities	803
Number of Connected Components	803
Giant Coefficient[a]	0.0157
Diameter[a]	8
Average Clustering Coefficient[a]	0.934

[a]Estimated considering the Giant Component.

4.4 Authors Indices

In this section, we consider different bibliometric indices to analyse the profiles of WEBIST authors. As previously stated (Sect. 3), we identified and extracted Google Scholar Citations profiles for 26.09 % of the WEBIST authors. Thus, the analysis presented in this section is related only to this subset of the authors.

The bibliometric indices from WEBIST authors were firstly analysed in terms of the Average and the Standard Deviation (σ) (see results in Table 3). The bibliometric indices, obtained from Google Scholar Citations data, were separated into global indices, estimated considering all the years of the citations, and the same indices estimated considering only the citations since 2009. On the average, the authors presented a considerable total number of citations and i10-index values greater than their h-index. However, the Standard Deviation was quite

Table 3. Average and standard deviation of number of citations and bibliometric indices from authors.

Measure	Average	σ
overall citations	1,634.49	4,087.46
citations *since 2009*	988.95	2,565.17
overall h-index	14.30	12.17
h-index *since 2009*	11.54	8.98
overall i10-index	28.16	54.32
i10-index *since 2009*	19.94	42.03

high, showing that the community, as expected in good conferences, is formed of both young and senior researchers, as further discussed in what follows.

To better understand the profile of the WEBIST authors, we performed further analyses by splitting the authors into two groups, named A and B. We assigned to Group A those authors who had an *overall* h-index greater than the h-index *since 2009* and assigned to Group B those authors who had a *overall* h-index equal to the h-index *since 2009*. This classification assumes that the authors whose *overall* h-index consisted solely of citations made after 2009 were researchers who had started their careers more recently than those whose *overall* h-index included citations from before 2009.

Table 4 presents the results using this classification. This table shows, for each conference year, the percentage of authors and the respective average of the h-index per class. The results evidence that, in all conference editions, the number of authors in Group A is greater than those in Group B. Also, the results show that, in all conference editions, the average h-index of authors in Group A is greater. Note that the average of h-index is 18.35 for authors in Group A considering all editions of WEBIST conference.

4.5 Program Committees Analysis and Indices

Program committee (PC) members of the first ten editions of the WEBIST conference were examined for potential information regarding discernible patterns or possible emerging social networks around particularly interconnected nodes. We looked at 569 individual researchers from 49 distinct countries. Figure 6 illustrates the dispersion of these PCs across a world map - the darker the color, the higher the number of participating institutions (countries which appear white had none). The topmost countries were found to be Italy, United States (USA), Germany, United Kingdom, Greece and Spain, representative of the international but not necessarily global reach of the WEBIST network of participating researchers. For all these countries, the number of participations of researchers as PC members (in the analysed period, each researcher could have participated in a maximum of 10 editions) was greater than 100. Cross-referencing these findings with those from Sect. 4.1 (USA, Germany, Netherlands, Spain and Portugal),

Table 4. Percentage and average of h-index of scholars in groups A and B.

Year	Percentage		Average of h-index	
	group A	group B	group A	group B
2005	84.62 %	15.38 %	19.77	9.50
2006	78.65 %	21.35 %	18.03	8.84
2007	80.59 %	19.41 %	18.58	7.24
2008	63.73 %	36.27 %	18.38	6.95
2009	67.86 %	32.14 %	20.39	8.15
2010	67.03 %	32.97 %	18.64	7.00
2011	67.06 %	32.94 %	20.16	5.54
2012	57.02 %	42.98 %	15.65	5.39
2013	53.03 %	46.97 %	20.74	6.52
2014	55.56 %	44.44 %	19.72	4.90
All	65.64 %	34.36 %	18.35	6.58

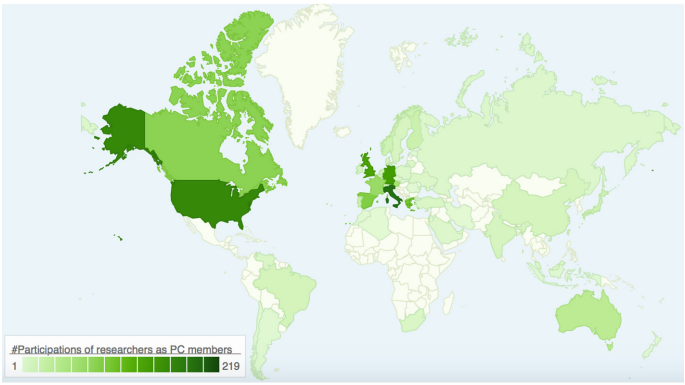

Fig. 6. Intensity of participations of PC members at institutions from the countries.

can provide some helpful suggestions in terms of potential future locations for conferences. These are in particular Italy (where WEBIST 2016 will be held), United Kingdom and Greece. Portugal was the most frequent location across previous conference sessions, but with 25 PC participants, it ranks as the 12th country overall.

The number of PC members is depicted in Table 5 (second column). On the average, the number of program committee members by conference year was approximately 175. To better illustrate the variation of researchers participating in the program committees, Fig. 7 shows a distribution based on the number of conference editions and how many researchers participated in that number of editions. Twelve researchers participated as a PC member in all of the ten editions of the WEBIST conference which form the dataset. Around a fifth,

(20.21 %) of the researchers participated as PC members for at least 50 % of the considered editions (at least five editions). Figure 8 depicts all the most active PC members (there are 34) who participated in at least 80 % of the conference editions as a tag cloud. This tag cloud represents the names of the researchers followed by their number of participations in the WEBIST PC in parentheses.

Table 5 also shows the percentage of new PC members (third column). This category consists of researchers who have not attended WEBIST in the capacity of a PC member before the corresponding conference edition; the percentage of variation in each program committee, as compared to the edition that immediately preceeds it is shown in the fourth column. On the average, the program committees had 27 % new members and 34 % of each committee had not participated as a PC in the previous year. This analysis shows that the WEBIST program committees have been composed of experienced researchers (the "core" of the PC) but that it is also constantly renewed and refreshed with the addition of new members.

Table 5. Number of program committee members over the conference edition.

year	#PC members	%new PC members	%different PC members related to the previous year
2005	129	-	-
2006	116	28.45 %	28.45 %
2007	200	49.00 %	49.00 %
2008	185	11.35 %	11.89 %
2009	151	15.23 %	23.84 %
2010	158	31.01 %	42.41 %
2011	116	26.72 %	43.10 %
2012	245	50.20 %	63.27 %
2013	221	12.22 %	19.46 %
2014	184	19.02 %	26.09 %
Avg	175.11	27.02 %	34.17 %

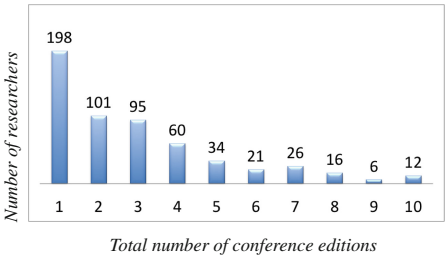

Fig. 7. Number of PC members participating in each total number of editions.

The following stage aimed to identify those PC members who also published in at least some WEBIST conference edition. In this analysis, the names of the authors (as extracted from DBLP) and the names of PC members (as extracted from WEBIST websites) were normalized (disregarding accents and not being case sensitive). A process of disambiguation was then carried out, by comparing the normalized versions of authors names to the normalized versions of PC members names. We were able to identify 114 equalities indicating that at least 20.03 % of the PC members are also authors in some WEBIST edition. Recall from Fig. 4 (Sect. 4.2) that 29 authors had published more than six papers in the first ten editions of WEBIST conference. Among them we identified 6 authors (20.69 % of the total) who were also PC members in some WEBIST edition (see Fig. 9). This reinforces our earlier conclusions regarding the most active authors. Moreover, none of the most active PC members (see Fig. 8) are amongst the topmost WEBIST authors (see Fig. 9) and all PC members published, on the average, only 2.31 papers in WEBIST. This reinforces the hypothesis of unbiased reviewing process (previously commented in Sect. 4.2) and one which is not favoring any group of authors, whether or not they are PC members.

Fig. 8. Top researchers with more than 7 participations in program committees.

Fig. 9. Top PC members with more than 6 papers published in WEBIST.

We estimated the number of citations and bibliometric indices (h-index and i10-index) from the PC members that published papers in some WEBIST conference in terms of the Average (see results in Table 6). To facilitate a comparison, we replicated the values previously presented in Table 3 (these can be seen in the second column of Table 6). On the average, PC members showed a considerably higher total number of citations, i10-index, and h-index than that obtained for all WEBIST authors across all editions. These results are coherent since it is to be expected that the program committees are composed of a selected group of experienced and qualified researchers.

Table 6. Average of number of citations and bibliometric indices from PC members.

Measure	Avg from Authors	Avg from PC Members	%Increase
overall citations	1,634.49	2,787.28	70.53 %
citations *since 2009*	988.95	1,528.32	54.54 %
overall *h*-index	14.30	20.48	43.22 %
h-index *since 2009*	11.54	15.68	35.88 %
overall i10-index	28.16	47.08	67.19 %
i10-index *since 2009*	19.94	30.33	52.11 %

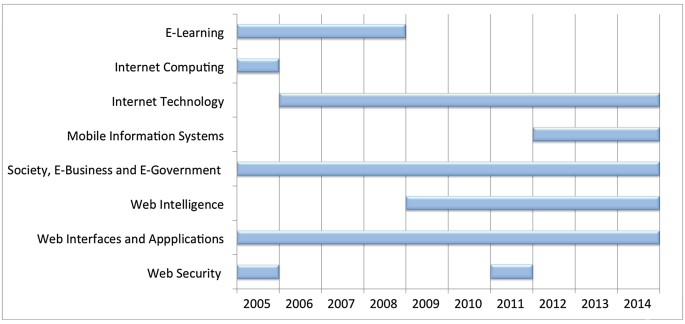

Fig. 10. Main conference areas per conference year.

4.6 Topics and Conference Areas

In this section, we analyse the topics of the papers published over the 10 years of WEBIST conference and their relation to the predefined main conference areas. Firstly, Fig. 10 presents, in alphabetical order, the main conference areas over the different conference editions. Some areas appear in all conference editions, such as *Society, E-Business and E-Government* and *Web Interfaces and Applications*. The third most frequent area is *Internet Technology*, which appeared from the second edition to the last one, probably as an expansion of *Internet Computing* (which appears only in the first conference edition). *Web Intelligence* and *Mobile Information Systems* appear more recently, in 2009 and 2012, respectively. *E-Learning* appears only in the first four editions of WEBIST conference. This phenomenon can be explained by the fact that the WEBIST conference, from 2009 to 2014, was held in conjunction with CSEDU (The International Conference on Computer Supported Education), a conference focused in innovative technology-based learning strategies and institutional policies on computer supported education (e-learning). *Web Security* appears only in specific editions (2005 and 2011).

Another analysis was performed over the topics covered by the papers published in WEBIST conferences. Figure 11 shows a tag cloud generated from the terms presented in the titles of the papers. This tag cloud represents the terms

Fig. 11. Top 50 terms of years 2005–2014.

followed by their total frequencies in parentheses. Moreover, the term size in the graphic is proportional to its frequency. Terms such as *web*, *systems*, *services*, *applications*, *model* and *information* are the most frequent. These terms are aligned with the research focuses of WEBIST conference that are technological advances and business applications of web-based information systems. Briefly, we have:

For a more detailed analysis, we considered the evolution of main conference areas and terms presented in titles of WEBIST papers per conference year (tag clouds from top 50 terms of each conference year are available at *WEBIST Analytics*). Specifically, we verified what happened to the frequency of particular terms that are directly related to updates in the main conference areas.

- *e-Learning* area was eliminated in 2009. *E-learning* term was a frequent top term in titles between 2005 and 2008, but this was not true in the following years (2009–2014).
- *Web Intelligence* area was included in 2009. Terms related to topics such as information filtering and retrieval, Web mining and classification appeared in different conference years (including years prior to 2009).
- *Web Security* area appears in editions from 2005 to 2011. The *security* term appears in the tag cloud of 2005 but not in 2011. We decided to investigate the quantity of papers published in 2011 that were directly associated with this main research area and discovered that only two short papers and one poster were published. This was probably the underlying reason which led to the deletion of this main research area in the following year.
- *Mobile Information Systems* area was included in 2012. The *mobile* term appears among the top 50 terms in 2012 (previously the term already appeared in the first conference editions, but became prominent only after the inclusion of the *Mobile Information Systems* area in 2012).

We also studied the evolution of the top 50 terms in the titles over a decade of WEBIST conferences. Table 7 presents the average and the standard deviation (σ) of the frequency of the top 50 terms. In the first editions of the conference, with the exception of 2005, both the average and σ were high, leading us to

conclude that there are likely to be terms that are related to major topics, as well as marginal topics in the accepted papers. In the most recent conference editions, the terms have a more equal distribution (greater equality frequency), showing that even whilst manifesting some peripheral change over the years, the conference found a core that is equally evolving. When analyzed in conjunction, the average and standard deviation demonstrate that the frequency of the top 50 terms (and consequently the relative frequency of the conference topics) is becoming more homogeneous. Moreover, a high diversity (dispersion) was observed, i.e., there were many terms (topics) covered by the conference over its 10 years.

The Pearson's correlation coefficient was estimated between the frequency of top 50 terms group from each conference edition (see results in Table 8). The sequence of the conference editions (underlined values in Table 8), except between 2006–2007, maintained a consistency within the group of top 50 terms: terms from one year correlated with the group of terms from the following year (Pearson's correlation coefficient is positive). Moreover, the correlation between the groups of top 50 terms from years 2008–2009 increased considerably compared with all the previous years (2005–2006; 2006–2007 and 2007–2008). This probably happened because, in this period, the main research areas were updated, with the removal of *E-learning* and the inclusion of *Web Intelligence*.

Finally, Table 8 shows an evolution on the research topics, considering the correlation between the top 50 terms of each conference edition and of all the others. The edition of 2010 presented, on the average, the highest Pearson's correlation coefficients between its top 50 terms and all others (being positive for all cases). Moreover, recall from Fig. 10 that WEBIST 2010 had as main research areas *Internet Technology*, *Society, E-Business and E-Government*,

Table 7. Average and standard deviation from frequency of top 50 terms per conference edition.

Year	Average	σ
2005	4.58	3.59
2006	9.08	6.90
2007	14.74	11.68
2008	10.50	9.12
2009	7.64	6.14
2010	7.18	5.37
2011	6.72	5.32
2012	7.12	4.53
2013	5.26	3.14
2014	5.08	3.02
All	70.30	55.66

Table 8. Pearson's correlation between the frequency of top 50 terms from each conference edition.

	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
2005		0.211	0.219	−0.012	0.128	0.178	0.105	−0.004	0.082	0.080
2006			−0.035	0.390	0.241	0.205	0.059	0.253	0.294	0.170
2007				0.174	0.178	0.259	0.084	0.189	0.178	0.140
2008					0.341	0.289	0.088	−0.007	0.118	0.005
2009						0.036	0.206	0.250	0.203	0.013
2010							0.325	0.316	0.404	0.122
2011								0.175	0.245	−0.103
2012									0.135	0.106
2013										0.395
2014										

Web Intelligence and *Web Interfaces and Applications*, which are the only areas that occur in the majority of conference editions (the "core" of research areas).

4.7 Paper Citation Analysis

In this section, we performed an analysis related to the WEBIST topmost cited papers (recall for Sect. 3 how these topmost papers were obtained) and estimated the h-index for the WEBIST conference series. The h-index obtained was 18, indicating that there are at least 18 papers with at least 18 citations. Thus, Fig. 12 presents the percentage of top 18 most cited papers per type of publication. The results show that the most cited papers are mostly full papers (more than 50 %, corresponding to 10 papers).

Figure 13 presents the top 18 most cited papers based on the percentage per main research areas. It can be seen that the *Web Interfaces and Applications* and *Internet Technology* areas had the highest number of most cited papers in the top 18 (around 33 % each). Surprisingly, *E-Learning*, which appeared only in the first four editions of WEBIST, had a higher percentage (around 17 %) of the most cited papers than *Society, E-Business and E-Government* (around 6 %) which appeared in all conference editions. As expected, the most recent main research areas do not have papers in the top 18 (2010 was the latest year with a paper in the top 18).

5 Discussion and Outlook

We described the *WEBIST Dataset* and the *WEBIST Analytics* Web application. The former aggregates data from different sources and follows the Linked Data principles, while the latter provides different functionalities for the searching, analysing, and visualising the dataset.

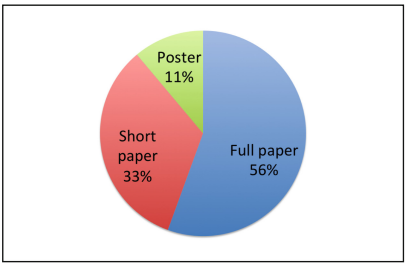

Fig. 12. Top 18 most cited papers per type of publication.

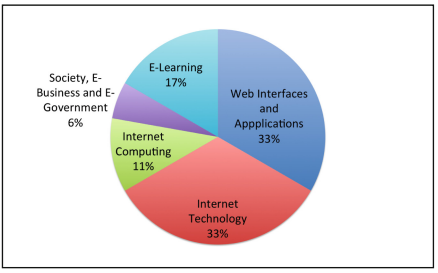

Fig. 13. Top 18 most cited papers per main research area.

A comprehensive analysis of the first ten editions of WEBIST illustrated the rapid growth in popularity achieved by WEBIST in 2007 and its maturation in subsequent years, reaching a stable conference-size, paper acceptance rate, community of IS experts, discernible research topics and supporters. The analysis highlighted the unbias of the reviewing process and how it contributed to the fast advancement of IS and the generation of knowledge: the WEBIST community plays a key role in knowledge transfer and impact in its domain (h-index = 18).

The *Web Interfaces and Applications* and *Internet Technology* tracks have been crucial to the development and popularity of WEBIST and they have accumulated the most cited papers. An important point to note is that the extinct *E-Learning* track, which appeared only four times as a main track, obtained a proportion of top cited papers which is higher than those of the *Society, E-Business and E-Government* track, although the latter appeared in all conference editions. Although the conference topics have became increasingly homogeneous, a higher diversity of topics and terms was observed. It is possible that a wider range of conference locations could bring about benefits, such as new collaborations with local universities and researchers.

The main contributions of this paper are the generated dataset and the Web application, which serve as a baseline for future analysis, including the extension of the proposed workflow to analyse multiple conferences and researchers from different fields.

Acknowledgements. This work was partly funded by CNPq under grants 444976/2014-0, 303332/2013-1, 442338/2014-7 and 248743/2013-9, by FAPERJ under grants E-26/101.382/2014 and E-26/201.337/2014 and by CAPES under grant 1410827.

References

1. Ott, R., Longnecker, M.: An Introduction to Statistical Methods and Data Analysis. Available 2010 Titles Enhanced Web Assign Series. Cengage Learning, Boston (2008)
2. Henry, N., Goodell, H., Elmqvist, N., Fekete, J.D.: 20 years of four HCI conferences: a visual exploration. Int. J. Hum. Comput. Interact. **23**, 239–285 (2007)

3. Blanchard, E.G.: On the weird nature of ITS/AIED conferences. In: Cerri, S.A., Clancey, W.J., Papadourakis, G., Panourgia, K. (eds.) ITS 2012. LNCS, vol. 7315, pp. 280–285. Springer, Heidelberg (2012)
4. Chen, C., Zhang, J., Vogeley, M.S.: Visual analysis of scientific discoveries and knowledge diffusion. In: Proceedings of 12th International Conference on Scientometrics and Informetrics (ISSI 2009), pp. 874–885 (2009)
5. Gasparini, I., Kimura, M.H., Pimenta, M.S.: Visualizando 15 anos de IHC. In: Proceedings of 12th Brazilian Symposium on Human Factors in Computing Systems, IHC 2013, SBC, pp. 238–247 (2013)
6. Posada, J.E.G., Baranauskas, M.C.C.: A study on the last 11 years of ICEIS conference - as revealed by its words. In: Proceedings of 16th International Conference on Enterprise Information Systems, vol. 3, pp. 100–111. SciTePress (2014)
7. Chen, C., Song, I.Y., Zhu, W.: Trends in conceptual modeling: citation analysis of the er conference papers (1975–2005). In: Proceedings of 11th International Conference on the International Society for Scientometrics and Informatrics, CSIC, pp. 189–200 (2007)
8. Zervas, P., Tsitmidelli, A., Sampson, D.G., Chen, N.S.: Kinshuk: studying research collaboration patterns via co-authorship analysis in the field of TeL: the case of educational technology & society journal. Educ. Technol. Soc. **17**(4), 1–16 (2014)
9. Procopio Jr., P.S., Laender, A.H.F., Moro, M.M.: Análise da rede de coautoria do Simpósio Brasileiro de Bancos de Dados. In: Brazilian Symposium on Databases - SBBD Posters (2011)
10. Cheong, F., Corbitt, B.J.: A social network analysis of the co-authorship network of the pacific asia conference on information systems from 1993 to 2008. In: PACIS, AISeL 23 (2009)
11. Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications. Cambridge University Press, Cambridge (1994)
12. Batista, M.G.R., Loscio, B.F.: OpenSBBD: Usando linked data para publicação de dados abertos sobre o SBBD. In: Brazilian Symposium on Databases - SBBD 2013, Short Papers (2013)
13. Bland, M.M., Altman, D.G.: Statistics notes: measurement error. BMJ **313**, 744 (1996)
14. Rodgers, J.L., Nicewander, A.W.: Thirteen ways to look at the correlation coefficient. Am. Stat. **42**, 59–66 (1988)
15. Gini, C.W.: Variability and mutability, contribution to the study of statistical distributions and relations. Studi Economico-Giuridici della R. Universita de Cagliari (1912)
16. Lopes, G.R., da Silva, R., Moro, M.M., de Oliveira, J.P.M.: Scientific collaboration in research networks: a quantification method by using gini coefficient. IJCSA **9**, 15–31 (2012)
17. Hoover, E.M.: Interstate redistribution of population, 1850–1940. J. Econ. Hist. **1**, 199–205 (1941)
18. Freeman, L.C.: Centrality in social networks: conceptual clarification. Soc. Netw. **1**, 215–239 (1979)
19. Hoser, B., Hotho, A., Jäschke, R., Schmitz, C., Stumme, G.: Semantic network analysis of ontologies. In: Sure, Y., Domingue, J. (eds.) ESWC 2006. LNCS, vol. 4011, pp. 514–529. Springer, Heidelberg (2006)
20. Marsden, P.V.: Egocentric and sociocentric measures of network centrality. Soc. Netw. **24**, 407–422 (2002)
21. Newman, M.E.J.: Scientific collaboration networks: I. network construction and fundamental results. Phys. Rev. E **64**, 016131 (2001)

22. Newman, M.E.J.: The structure and function of complex networks. SIAM Rev. **45**, 167–256 (2003)
23. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. Phys. Rev. E **69**, 026113 (2004)
24. Hirsch, J.E.: An index to quantify an individual's scientific research output. In: Proceedings of National Academy of Sciences of the United States of America, vol. 102, pp. 16569–16572 (2005)
25. Berners-Lee, T.: Linked Data. In: Design Issues. W3C (2006)
26. Elmagarmid, A.K., Ipeirotis, P.G., Verykios, V.S.: Duplicate record detection: a survey. IEEE Trans. Knowl. Data Eng. **19**, 1–16 (2007)
27. Borges, E.N., de Carvalho, M.G., Galante, R., Gonçalves, M.A., Laender, A.H.F.: An unsupervised heuristic-based approach for bibliographic metadata deduplication. Inf. Process. Manage. **47**, 706–718 (2011)
28. Bizer, C., Seaborne, A.: D2RQ - treating Non-RDF databases as virtual RDF graphs. In: Proceedings of 3rd International Semantic Web Conference (2004)