A Scientific Resource Space for Advanced Research Evaluation Scenarios

Cristhian Parra, Muhammad Imran, Daniil Mirylenka,
Florian Daniel, Fabio Casati, and Maurizio Marchese

Department of Information Engineering and Computer Science
University of Trento, Via Sommarive 5, 38123, Trento, Italy
{parra,imran,dmirylenka,daniel,casati,marchese}@disi.unitn.it

Abstract. In this paper, we summarize our experience and first results achieved in the context of advanced research evaluation. Striving for research metrics that effectively allow us to predict real opinions about researchers in a variety of scenarios, we conducted two experiments to understand the respective suitability of common indicators, such as the h-index. We concluded that realistic research evaluation is more complex than assumed by those indicators and, hence, may require the specification of even complex evaluation algorithms. While the reconstruction (or reverse engineering) of those algorithms from publicly available data is one of our research goals, in this paper we show how can we enable users to develop their own algorithms with *Reseval*, our mashup-based research evaluation platform, and how doing so requires dealing with a variety of data management issues that are specific to the domain of research evaluation. Therefore, we also present the main concepts and model of our data access and management solution, the *Scientific Resource Space (SRS)*.

Key words: research evaluation, scientific data access and management, resource space, reputation

1 Introduction

The evaluation of research, i.e the assessment of productivity or measuring and comparing impact, is an important process used to select and promote personnel, assign research grants, measure the results of research projects, and so on. The main instrument the community has been relying on so far is the computation of indicators that are based on bibliographical information, e.g., citations and publication counts, indexes like the h-index [10], etc. Yet, we all know that, for instance, in order to select a new researcher or a new professor to hire we don't just look at the h-indexes of the candidates, rank them in decreasing order, and select the first in the list. In fact, selection processes typically involve more complex decision logics and are, also, partially subjective (e.g., taking into account the opinion of the evaluators). As of today, however, such kind of complex evaluation logics are not supported by existing evaluation tools and, therefore, evaluation still requires significant effort in terms of manual work and interpretation.

Ideally, it should be possible to develop metrics also for those complex evaluation scenarios. In practice, however, doing so turns out to be far from easy, since each evaluation process is different from another and contains a lot of tacit knowledge the evaluators oftentimes are not even aware of themselves. Eliciting this kind of tacit knowledge is one of the first step toward an approach of metrics definition that we call "reverse-engineering". That is, given a set of concrete evaluations or rankings and a set of features describing the researchers involved in the evaluation process (e.g., their h-index, their participation in programme committees, the size of their social network, etc.), what we want to do in the long term is deriving which are the algorithms that allows us to reconstruct the same ranking and, hence, to better predict future assessments.

In the short term (in this paper) – recognizing that more data needs to be collected and analyzed to come up with good results – we however think that it can be already beneficial to allow people to define their own evaluation algorithm, to the best of their knowledge. The approach we follow in doing so is that of implementing a research-evaluation-specific mashup platform, that allows its users to source data from different online data sources, to process, aggregate or filter them, in short, to compose them into a complex evaluation logic.

The advent of the Web has placed much of the necessary resources online and today researchers have access to an overwhelming space of scientific publications thanks to instruments that range from traditional digital libraries (such as SpringerLink or Scopus) to specialized search engines (such as GoogleScholar) and metadata services (such as DBLP). All these sources are, however, heterogeneous and spread over the Web, making it complex to integrate them in a coherent and trustworthy way. There are already tools, such as Harzing's *Publish or Perish* (`http://www.harzing.com/pop.htm`) desktop application for the *Scholar H-Index calculator* (`https://addons.mozilla.org/en-us/firefox/addon/scholar-h-index-calculator/`), that use parts of these data sources and support the computation of simple metrics, which can also be reused and integrated if properly wrapped. Then, today's Web 2.0 enables the early sharing of knowledge through instruments like wikis, blogs, or personal web sites. These kinds of contributions are not peer-reviewed but might still have an impact on the scientific community, depending on the reputation of their authors (think, for instance, of the so-called technology evangelists).

In this paper (i) we report on our first step toward the reverse-engineering of evaluation metrics, i.e., the analysis of how well *existing metrics* may help us reconstruct the outcomes of complex evaluation processes (Section 2); (ii) we propose our idea of *Reseval mashup platform* for the drag-and-drop development of complex metrics (Section 3); and (iii) we describe our approach to deal with the above heterogeneity of sources and functionalities available online, i.e., we describe our *Scientific Resource Space* (Section 4), highlighting the peculiar data management issues that characterize the research evaluation domain. Then we discuss related works and provide an outlook over our future work.

2 On Scientific Reputation and Indicators

Reputation is regarded as the measure of our worth or credibility, according to what others think of us, based on past interactions [17]. Differences among reputation systems are only related to dimensions such as the conceptual model of reference and its granularity, or the sources of information in use [18]. The most basic definition is that reputation is in *the eye of the beholder*, a concept highly based on the experience of the e-commerce and economics domains and that could be different for science or other environments (i.e. social networks like Facebook).

How is this reputation built? Is the reputation of a scientist a good indicator of the excellence of his work?. These and many other questions are becoming relevant as we better realize that single indicators are no longer sufficient to represent scientific impact [2]. Answering them require us to *collect information* (both of reputation and of scientific output), to then analyse them in the search for patterns that shed light on the the nature of reputation phenomena within science. Reseval, powered by the SRS, helped us to cope the challenges of collecting data.

The final goal of this research thread, is to derive the logic of reputation in the mind of researchers, in a way that we can represent it as new metrics for research impact. As a first step towards that goal, we have conducted experiments on the relation between bibliometric indicators and perceived reputation. In the following, we describe these experiments and report on the preliminary results we have obtained from them.

2.1 Experiments

To study whether there is or not a relation between bibliometric indicators and perceived reputation, we needed to find sources of reputation information for a set of researchers and then compute bibliometric indicators for the same set of people. Reseval provide the indicators, but for the reputation information we followed two different approaches: (i) A **survey** asking about research impact and deployed in several conferences of Computer Science. (ii) Crawling results from **research position contests** in Italy and France, produced by selection committees.

Besides Reseval, the study also included some indicators obtained from ReaderMeter (`http://readermeter.org/`) and a parser for Google Scholar search results. Once all the information was available, correlation analysis was performed using Kendall-tau method comparing rankings resulting from reputation ratings and rankings resulting from bibliometric indicators. Only in the case of Italian research contests the analysis was different due to the fact that reputation rankings obtained from this source were only pairs of one selected candidate and one candidate put on a waiting list.

Reputation Survey. The *Liquidpub scientific reputation survey* was designed to be deployed in several conferences with a set of candidates relevant for that conference. Each survey consisted on **a sample of 40 candidates** taken from Jens Palsberg's top h-index researchers list (`http://www.cs.ucla.edu/~palsberg/h-number.html`). Half of the sample was computed according to a measure of affinity to the target conference, based on the distance within co-authorship networks of evaluated researchers with respect to others that published in the same conference. In total, 8 surveys were implemented and deployed in conferences such as *BPM (Business Process Management)*, *ICWE (Web Engineering)* and *VLDB (Very Large Databases)*, getting a total of 77 answers in a period of 3 months of being online (`http://reseval.org/survey/`)

Research Contests. The second approach for getting reputation information consisted of getting the results of contests for research position for Italy and France. In the case of Italy, available data at MIUR site was from 2008 and included, for each contest, the pair of selected candidates where one was the winner of the contests and the other was the second place. For **208 contests pairs** where both candidates had at least one recorded citation, we later calculated in what percentage of the times bibliometric indicators succeeded on predicting the first place of the contest. In the case of France, CNRS data included a list of more than **1000 researchers** participating in differences contests whose result were published in the form of a ranking of 2 or more people.

2.2 Preliminary Results

Reputation Survey. Analysis of reputation ratings in the survey compared to bibliometric indicators showed, for all conferences, a stable pattern of correlation coefficients below the threshold for considering them significant. For this to happen, the correlation coefficient has to be **greater than 0.5** (positive correlation) or **less than -0.5** (negative correlation). Figure 1 shows the value of these coefficients for each metric and source, based on the aggregated results from all the surveys.

Research Contests. Correlation analysis of CNRS rankings shows the same pattern of no-correlation we encountered in the surveys. Figure 2 shows a summary of theses coefficients, all near of zero. This being said, we still need to extend this dataset with the names of researchers that were eliminated on early phases of the selection process of CNRS, organized in three stages.

Table 1. Italian Contest results and bibliometric indicators' performance

	H-Index	Citation count	Cited publications
W < S	47.1% (98)	56.2% (117)	50.5% (105)
W > S	38.9% (81)	39.4% (98)	47.6% (99)
W = S	13.9% (29)	4.33% (9)	1.92% (4)

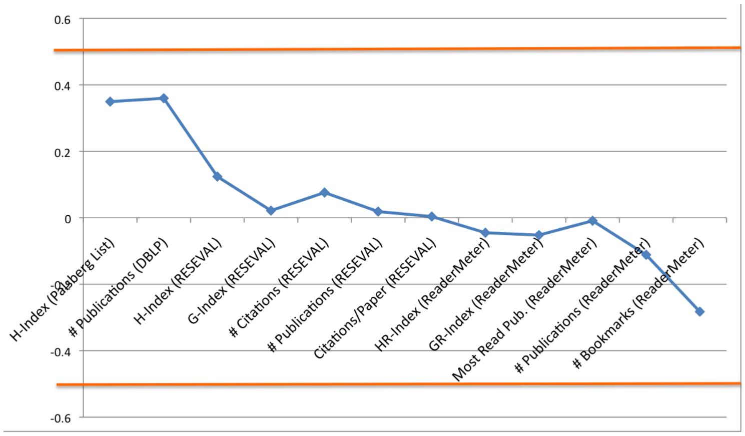

Fig. 1. Correlation coefficients between reputation and bibliometric indicators rankings

In the case of the Italian research contests, Table 2.2 shows percentage of cases in which the winner has a lower indicator than the second place (W < S), the winner has a better indicator than the second place (W > S) and finally where indicators are the same (W = S). We report only on those indicators that had the better performance, which are the h-index, the total citation count and the number of cited publications. As the table shows, no indicator have a performance better than **50%**.

3 Mashing Up Complex Evaluation Metrics

The experiments above show that research evaluation in practice is much more than just looking at one single indicator or metric. In fact, a good evaluation rather consists in a set of evaluation steps, the application of multiple metrics or data processing functions, comparisons, and similar, a scenario that naturally lends itself to be considered a composition problem. Yet, we are not in the presence of a traditional web service composition, but more in the presence of a *mashup* scenario, in that comparing evaluation results will also require the integration of user interfaces (UIs), e.g., charts and diagrams, in addition to services and data.

With the Reseval project, we aim at providing a mashup platform for research evaluation, taking advantage of both a service-oriented and a data-oriented approach. Reseval[1] is a preliminary research evaluation platform that is currently being developed. An example of a mashup is shown in Figure 3, which describes the evaluation algorithm adopted by the central administration of the University of Trento (UniTN), used for internally distributing resources to departments. In

[1] http://reseval.org/

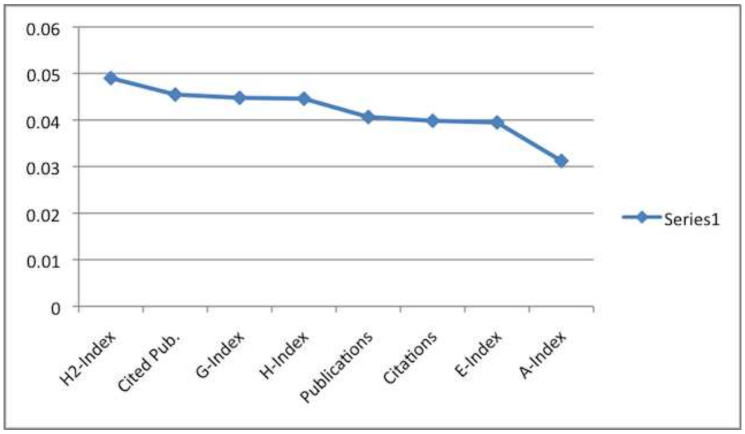

Fig. 2. Correlation coefficients between reputation and bibliometric indicators for Research Contests from CNRS

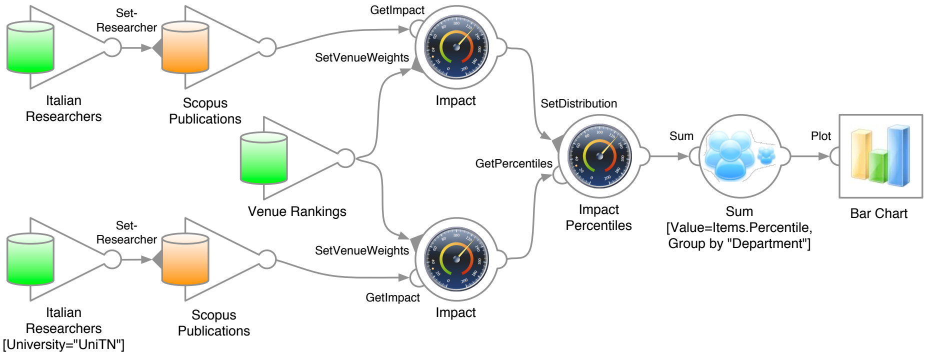

Fig. 3. A complex research evaluation logic mashed up in Reseval

essence, the algorithm computes how well each UniTN researcher performs (in terms of publications) in his disciplinary sector at Italian level, groups performances by department, and plots the final result. The comparison is based on bibliometric indicators.

The requirement we extract from this domain-specific scenario is that we need to empower people involved in the evaluation process (that is, the average faculty member in any academic discipline, as well as the staff member in charge of research evaluation) so that they are able to define and compare relatively complex evaluation processes, taking and processing data in various ways from different sources, visually analyze and understand the results.

Even though there are readily available applications for assessing research impact, all the currently available solutions/tools lack in our view some key features, namely: (1) completeness of data, (2) flexible and personalized metrics, (3) languages to support the user in defining sources, queries, and metrics, and (4) data processing options. Data completeness is indeed a main issue in the process of evaluating research people. In fact, some sources do not (completely) cover some disciplines; for instance, Web of Science is not good for Computer Science, while it is very important to compute citations received by all the documents published by a given author. As we will see, We tackle this issue by leveraging on an open, resource-oriented, Scientific Resource Space System (SRS) that is able to provide homogeneous programmatic access to heterogeneous resources and web services, regardless of how they are implemented, as long as they are web accessible.

We believe that the personalization of the evaluation processes is a key element for the correct use and practical success of the various evaluation indexes. Moreover, people involved in such evaluation process most of the time are not IT experts, capable of building proper software for crawling data sources, automatically parsing relevant information, merging data and computing the needed personalized metrics. Therefore, in order to empower the interested persons an appropriate and possibly easy-to-use IT platform need to be designed, implemented and tested.

Enabling users to develop their own applications or compose simple mashups or queries means simplifying current development practices. Some mashup approaches heavily rely on connections between components (this is the case of Yahoo! Pipes[2] and IBM Damia, for instance), and therefore are inherently imperative; other solutions completely disregard this aspect and only focus on the components and their pre- and post-conditions for automatically matching them, according to a declarative philosophy.

4 The Scientific Resource Space (SRS)

In order to support the composition paradigm proposed by Reseval and to compute quality indicators, we need to access the data we need. To this end, we have built upon the ideas of *dataspaces*, which extend concepts from traditional database management toward heterogeneous data sources [6][9]. The Scientific Resource Space (SRS) follows the intuition that every single piece of knowledge in the vast scientific arena (available online) can be treated as a resource uniquely identified by a URI. In this section, we present our design of such an SRS and report on the current status of its implementation.

4.1 Basic concepts

Managing a space of resources means bringing together inside one homogeneous environment a variety of heterogeneous kinds of resources and providing suitable means to access and use resources and to define and maintain all necessary relationships among the resources. In short, a *resource* can be any artifact we can refer to by a URI and that is accessible over the Web. This notion is very general and captures the requirement of supporting any arbitrary information

[2] http://pipes.yahoo.com/pipes/

such as simple web pages, online documents, web services, feeds, and so on. That is, resources might be simple sources of data or content, but they might also be as complex as SOAP or RESTful web services with their very own interaction logic.

A *resource space* can then be defined as a set of resources and relationships, where the set of resources limits the space to a manageable number of resources, and the relationships express how the resources in the space are interrelated. Theoretically, the biggest resource space with our definition of resource is the Web itself, but, of course, we do not aim at providing a new way of managing the Web. Instead, we think that only by setting suitable boundaries for the resources to be considered, i.e., by limiting the resource space, it is also possible to provide value-adding, novel functionalities that justify the development of a dedicated management system of a SRS.

So far, the concepts are general and prone to be applied on any domain. Our focus, however, is scoped to research evaluation in science. For this, we limit the concept of a resource to include only those artifacts that identify a *scientific resource* such as papers, researchers, journals, conferences, datasets, and so on.

Our implemented SRS has modeled both traditional scientific artifacts (papers, journals, conferences) in and other non-traditional (research blogs, datasets, experiments) in terms of their specific relationships (e.g. co-authorship, citation, less traditional but relevant nowadays' bookmarking, etc.) and possible attributes (e.g. title of a paper, name of an author, volume of a conference, etc.). Regarding the sources of metadata about scientific artefacts, we consider traditional digital libraries and scholarly search engines, as well as social networking services for scientists and even general-purpose social sites.

4.2 High Level Architecture and Implementation

The central component of SRS is the Metadata Warehouse (Figure 4), whose implementation largely follows the traditional ETL (Extract Transform Load) process. The Adapter Layer encapsulates the differences between the data sources. Each adapter is responsible for getting metadata according to the protocols and APIs provided by the source and transforming it into the model of Scientific Resource Space. This task is performed in a few steps. First, scientific metadata is gathered from a source and stored into preliminary tables. The metadata is then loaded into the staging and joined with metadata from other sources. At this stage, metadata elements from each source are preliminary merged based on the identifiers provided by source, ensuring that we introduce no duplicates at the source level. During the cleaning phase the staging area is analyzed to discover (entity matching) and merge entities duplicated across different sources. The algorithms of such matching may vary from simple and intuitive ones, such as comparing titles of the scientific papers, to potentially sophisticated ones like analyzing the co-authorship graphs of scientists. After being cleaned, the metadata is finally loaded into the target database, where it is made available for the applications. This loading is performed by computing and making the changes with respect to the current state of the target database.

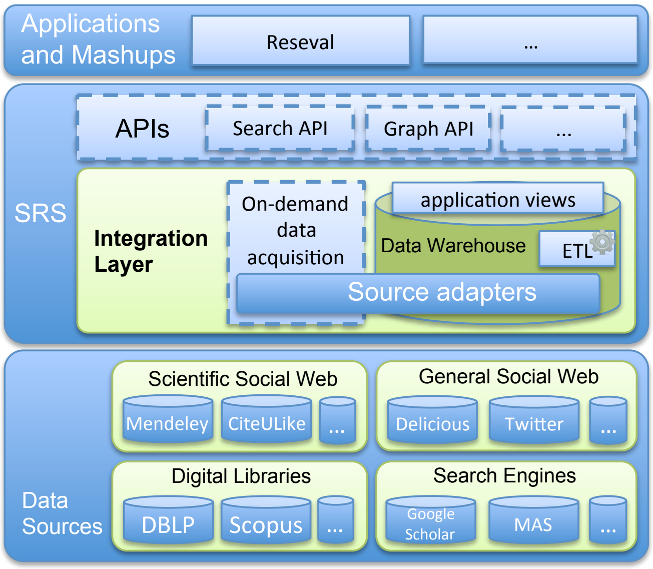

Fig. 4. High-level architecture of Scientific Resource Space (SRS)

The applications built on top of SRS focus on different aspects of scientific resource metadata. In order to provide useful functionality with reasonable performance, they require efficient access to their own presentation of the scientific resource space. For instance, Reseval works with different research metrics including publication and citation-based ones. The numbers of citations and self-citations for papers and authors are the primary units of data for Reseval, and are accessed frequently. For performance reasons, these numbers can not be calculated dynamically and have to be precomputed. SRS addresses this problem by creating the application-specific views which contain all the data needed by the application in a suitable format, and are updated at the final stage of the ETL process. These specific views, as well as general functionality of SRS can later be exposed through the APIs such as search API, or API for navigating the scientific resource space.

In order to enable source-dependent requests, SRS propagates the information about the sources of metadata elements through all the stages of the process to the target database and the application-specific views. At any time for any metadata element it is possible to learn which source or sources it originates from. This property of SRS allows Reseval to compute metrics with respect to any source or combination of them.

Apart from warehousing of metadata, SRS is also looking to support on-demand data acquisition from sources. This approach potentially allows our services to employ sources that expose only search interface, and also to provide personalization in cases when sources rely on user profiles.

5 Related Work

Bibliometrics. Bibliometric indicators have become a standard and popular way to assess research impact in the last few years. All significant indicators heavily rely on publication and citation statistics and other, more sophisticated bibliometric techniques. In particular, the concept of citation [8, 7] became a widely used measure of the impact for scientific publications, although problems with citation analysis as a reliable method of measurement and evaluation have been acknowledged throughout the literature [3]. Indeed, not always a paper is cited because of its merits, but also for some other reasons, as flaws, drawbacks or mistakes. A number of other indices have been proposed to balance the shortcomings of citations count and to "tune" it so that it could reflect the real impact of a research work in a more reliable way. Scientometrics was then introduced as a science for analyzing and measuring quantitatively science itself [4].

In the last decade a number of new metrics were introduced. Although these metrics are also based on citation analysis but they gained popularity over simple citation indexes. For instance h-index [10] was proposed by Jorge Hirsch, as a more comprehensive metric to access the scientific productivity and the scientific impact of an individual researcher. Focusing on a indicator which should indicate quality of a researcher, should consider the performance of top cited paper. Such indicator g-index is proposed by Egghe [5]. To overcome some limitations of both the h-index and the g-index, a new index has been proposed in [1] with the aim to combine the good properties of both indices and to minimize the disadvantages. An index is proposed in [13], is called AR-index, which not only takes into account citations of a researcher but also the publication age. The performance changes in researchers career which comes over the time were ignored previously, thus AR-index can increase or decrease over time.

Reputation and complex metrics. Michèle Lamont's book [16] holds a complete analysis on how evaluation is performed by professors. In the book, she analyses the hard details of peer reviews and 12 panels of experts in the humanities ans social science, extrapolating subjective criteria for decision-making in each different discipline, giving an interesting overview of possible **features** that influence reputation of researchers. The Altmetrics Initiative [11] goes one step further and aims at using social interactions for proposing new metrics of research impact better related to the reputation of researchers.

In the line of analyzing scientific promotion and its relationship with bibliometric indicators, [12][15][14] are some works that show results on how these indicators are related to scientific promotion or how they behave in some particular communities (e.g. Greek Departments of Computer Science). They are related to the experiments we have done (or plan to do) on the approach. However, none of them have tried to compare standard bibliometric indicators with direct reputation, which is the approach we want to use.

Information sources for research evaluation. Until recently researchers had essentially only one source for looking bibliometric type of information: the

Web of Science[3] an on-line commercial database from Thomson Scientific. Starting from the late 90's, many other competitors emerged like Citeseer[4], Scopus[5], Google Scholar[6] and Microsoft Academic[7], with the purpose of giving users a simple way to broadly search the scholarly literature.

Based on the existing sources, new tools are beginning to be available to support people in the research impact analysis. A useful tool is Publish or Perish[8], a desktop based software program that uses only Google Scholar to retrieves the citation data, and then analyzes it to generate the citations based metrics. A different approach is provided by Scholarometer: a social tool which is used in citation analysis and also for evaluation of the impact of an author's publications. It is a browser free add-on for Firefox that provides a smart interface for Google Scholar and requires users to tag their queries with one or more discipline names. Information sources and tools based on these sources are becoming available but they still have many shortcomings. For example they differ in data coverage, data quality. Moreover, these tools are data-source specific and can not be extended to use other data-source. Moreover personalization of metrics is still missing.

6 Conclusion and Future Work

In this work, we have summarized our experience on the research and development of new means for scientific research evaluation, highlighting its requirements in terms of domain-specific data management. As a first step, we approach the problem with *Reseval*, a mashup platform for the composition of complex evaluation metrics. We solve Reseval's data and metadata integration challenges by the means of a dedicated layer, our *Scientific Resource Space*. The resulting integrated infrastructure, has proven to be a powerful instrument also to drive our investigation on the nature of reputation and the reverse-engineering of evaluation metrics we presented in this work.

Yet, as this paper shows, or work is far from being done and our aim is to increase the coverage of our SRS and to improve its integration with other applications that go beyond Reseval. Our work on the reverse-engineering of reputation will benefit from this integration, and we will drive new research threads that are not only in the scope of research evaluation.

We plan to have a demo of the integrated tool ready for demonstration at the time of the conference.

[3] http://scientific.thomson.com/products/wos/
[4] http://citeseer.ist.psu.edu/
[5] http://www.scopus.com/home.url
[6] http://scholar.google.com/
[7] http://academic.research.microsoft.com/
[8] http://www.harzing.com/pop.htm

References

1. S. Alonso, F. Cabrerizo, E. Herrera-Viedma, and F. Herrera. hg-index: A new index to characterize the scientific output of researchers based on the h-and g-indices. *Scientometrics*, 82(2):391–400, 2010.
2. J. Bollen, H. Van de Sompel, A. Hagberg, and R. Chute. A principal component analysis of 39 scientific impact measures. *PloS one*, 4(6):e6022, 2009.
3. A. Chapman. Assessing research: citation count shortcomings. *The Psychologist*, 2:336–44, 1989.
4. D. de Solla Price. *Little science, big science–and beyond*. Columbia University Press New York, 1986.
5. L. Egghe. Theory and practice of the g-index. *Scientometrics*, 69(1):131–152, 2006.
6. M. Franklin, A. Halevy, and D. Maier. From databases to dataspaces: a new abstraction for information management. *ACM Sigmod Record*, 34(4):27–33, 2005.
7. E. Garfield and R. Merton. *Citation indexing: Its theory and application in science, technology, and humanities*, volume 8. Wiley New York, 1979.
8. E. Garfield and A. Welljams-Dorof. Of Nobel class: A citation perspective on high impact research authors. *Theoretical Medicine and Bioethics*, 13(2):117–135, 1992.
9. A. Halevy, M. Franklin, and D. Maier. Principles of dataspace systems. In *Proceedings of the twenty-fifth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 1–9. ACM, 2006.
10. J. E. Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, 102(46):16569–16572, 2005.
11. P. G. Jason Priem, Dario Taraborelli and C. Neylon. Alt-metrics: A manifesto. 2010.
12. P. Jensen, J. Rouquier, and Y. Croissant. Testing bibliometric indicators by their prediction of scientists promotions. *Scientometrics*, 78(3):467–479, 2009.
13. B. Jin. The AR-index: complementing the h-index. *ISSI Newsletter*, 3(1):6, 2007.
14. D. Katsaros, V. Matsoukas, and Y. Manolopoulos. Evaluating Greek Departments of Computer Science/Engineering using Bibliometric Indices.
15. J. Kulasegarah and J. Fenton. Comparison of the h index with standard bibliometric indicators to rank influential otolaryngologists in Europe and North America. *European Archives of Oto-Rhino-Laryngology*, 267(3):455–458, 2010.
16. M. Lamont. *How professors think: Inside the curious world of academic judgment*. Harvard Univ Pr, 2009.
17. G. Origgi and J. Simon. On the Epistemic Value of Reputation. In *Paradigms and conceptual systems in knowledge organization: Proceedings of the 11th International ISKO conference, 23–26 February 2010, Rome, Italy*, pages 293–300, 2010.
18. J. Sabater and C. Sierra. Review on computational trust and reputation models. *Artificial Intelligence Review*, 24(1):33–60, 2005.