ISSI 2013

SCIENTOMETRICS

14th International Society of Scientometrics and Informetrics Conference

15th–19th July 2013

Vienna, Austria

PROCEEDINGS Volume I

PROCEEDINGS OF
ISSI 2013
Vienna

VOLUME 1

14th International Society of
Scientometrics and Informetrics Conference

Vienna, Austria
15th to 20th July 2013

Editors

Juan Gorraiz, Edgar Schiebel, Christian Gumpenberger, Marianne Hörlesberger, Henk Moed

Sponsors

ASIS&T, USA
Elsevier B.V.
EBSCO Information Services, USA
Federal Ministry for Science and Research, Austria
Federal Ministry for Transport, Innovation and Technology, Austria
Information Assistant, Verein für Informationsmanagement, Vienna
ORCID, Inc.
Science-Metrix/R&D Reports
Swets Information Services
Thomson Reuters
ZSI - Centre for Social Innovation, Vienna

Printed by Facultas Verlags- und Buchhandels AG,
Stolbergasse 26, A-1050 Wien

ISBN: 978-3-200-03135-7
ISSN: 2175-1935

ARE CITATIONS A COMPLETE MEASURE FOR THE IMPACT OF E-RESEARCH INFRASTRUCTURES?

Jonkers Koen[1], Derrick Gemma Elizabeth[1], Lopez Illescas Camen[2], Van den Besselaar, Peter[3]

koen.jonkers@csic.es
[1] CSIC Institute for Public Goods and Policies, Department of Science and Innovation dynamics, C/Albasanz 26-28, 28033 Madrid, Spain
[2] SCImago group. Department of Information Science. University of Extremadura, Badajoz, Spain.
[3] Vrije University van Amsterdam, Department of Organization, Science and network institute, Amsterdam, The Netherlands

Abstract
This micro-level study explores the extent citation analysis provides an accurate and representative assessment of the use and impact of bioinformatics databases. The case study suggest that there is a relation between number of visits and number of citations. The second finding is that citation analysis underestimates acknowledged use by between 5 and 30% for most of the databases and applications studied. The paper discusses the implications of the findings for various aspects of impact measurement.

Conference Topic
Topic 2: Old and New Data Sources for Scientometric Studies: Coverage, Accuracy and Reliability

Introduction
This paper explores to what extent citation analysis provides an accurate and complete assessment of the usage of e-research infrastructures in the research underlying published scientific articles. One of the reasons that measuring impact is generally based on citations, may be the mere existence of large, accessible databases such as WoS and Scopus. This is in addition to the preference evaluators have for measures that are "countable". The extent to which citations fully reflect the usage of knowledge claims by other scientists, however, is disputed. A number of alternative metrics, including citations in patents and social media statistics, have been promoted as ways to assess the broader impact of research, among many others e.g. De Jong et al (2011). However, for measuring scholarly impact of research, citation based indicators are still the dominant approach.
Recently, measuring impact of research infrastructures has been put on the agenda. The scholarly use and impact of research technologies, as of scientific knowledge claims, could be assessed through citation analysis. For many

scientific innovations, especially in the case of research infrastructures citations may no longer be a sufficient way with which to represent 'impact', as the user community may be very diverse. Where citations can help to measure scholarly use as a component of an infrastructure's impact, there are a number of alternatives that complement the measurement of its visibility and influence, such as the log-files that measure the visits to the website of the infrastructure. Considering the importance of research instruments in biotechnological innovation processes (e.g.: Senker, 1995) a full assessment of the impact of e-research infrastructures should also include an analysis of the references in patents. Nevertheless, citations may be a relevant representation of the use and impact of research infrastructures.

This article aims to investigate firstly to what extent that is the case: to what extent do citations to the original articles that introduce a research infrastructure provide an accurate representation of use and impact? If so, the intensity of use (measured in number of visits to the URLs of the infrastructures' domains) is systematically related to the citations to the articles in which these research infrastructures were introduced. Citations would therefore be a strong indicator of usage.

Apart from citations, papers may include in-text references to the research infrastructure. Therefore, the second aim of the paper is to investigate whether citations are an adequate representation of these in-text references to used e-research technologies. In other words, we investigate how much of the acknowledged use of research technologies is neglected when using only citation counts, while not considering the in-text references. Both questions will be explored, using research databases with biological info hosted by ExPASy.

Theoretical background: Why citations?

Two main bodies of theory underlie the use of citation analysis for the assessment of research output. The normative theory of citations states that researchers cite documents that are relevant to their topic, and that provide useful background for their research. By citing they acknowledge an intellectual debt (Bornmann & Daniel, 2006). Cronin (1984) argues that citations perform a scholarly communication function between texts in line with the normative theory of citations, and according Martin and Irvin, citations can indicate a measure of reward for past work or scientific status (Martin & Irvine, 1983).

The second theory, whilst not mutually exclusive to the first, emphasises that citations to documents are not free from personal bias or social pressures. Therefore the "social-constructive theory of citations" states that citing is a social process, and as such citations are used as an aid for persuasion(Gilbert, 1977; Cozzens, 1989).

The social constructivist theories provide some explanations for why people would add additional citations, beyond those that could be expected on the basis of the normative view of citations. In an age in which citation analysis is becoming an increasingly prominent feature of research evaluation, authors are

inclined to cite in an attempt to raise the visibility of their own work or that of their colleagues, with or without the implicit expectation that this favour will be returned. Unlike previous contributions, this paper is not concerned with these additional citations but with the phenomenon that authors may *not cite* certain knowledge claims even if they explicitly state their usage.

One potential explanation for this is that the origin of knowledge claims can be lost over time as new (arguably improved) claims emerge. The original knowledge claims may be absorbed into the common knowledge of a research discipline or even of the general public (Martin & Irvine, 1983). Researchers who use the knowledge claim may either not be aware of the existence of a citable item or consider it superfluous. Forgetting is another obvious motivation for not including a citation, as is the consideration that the knowledge claim in question does not merit a citation. Finally the possibility exists that alternative forms of acknowledgements besides citations are being used.

The motivation to include a reference can differ from author to author and from reference to reference. It is therefore probably too simplistic to think within just the two theories discussed in this section. In fact, it may be impossible to develop a convincing 'theory of citations' (Weingart, 2005), as citing behaviour and citations as indicators for impact and quality may actually be two unrelated issues. The more aggregated, the more citation counts may be detached from citing behaviour and the more useful they may be for investigating impact. Despite the highlighted limitations, there are several characteristics of citations that contribute to our understanding of what they actually represent, and these can be used to determine when it is appropriate to apply citation analysis and when a suitable alternative or complement is required.

Data and Methodology

Not all types of knowledge claims receive, on average, an equal amount of citations (Martin & Irvine, 1983). Reviews, for example, tend to receive more citations than articles (Asknes, 2005; Moed et al, 1995). Peritz (1983) showed that methodological papers in sociology were more frequently cited when compared to non-methodology papers. There are grounds to expect this is the case in the life sciences as well. A famous example is one of the most cited articles of all times (*Protein measurement with the folin phenol reagent*). Published in 1951 and with 299,133 "WoS citations" in Dec 2012, the article outlines a commonly used method in biochemistry to determine protein concentrations (The Lowry method) (Lowry et al, 1951; Garfield, 1998). The databases and applications on which this study focuses, are research tools which are used by many life scientists. The papers introducing them therefore have the potential to receive a high number of citations as well.

The databases and applications analysed in this project are hosted by the Expert Protein Analysis Server, ExPASy, developed and maintained by the Swiss Institute of Bioinformatics. They are used by life scientists to analyze and interpret among other the genetic and protein sequence information they

citation is a strong indicator of usage. In other words, we expect that the ratio of use (measured as visits to the site) and citations is about the same for the four infrastructures.
Secondly we aim to explore the extent to which citations are an adequate representation of the in-text references to e-research technologies - in this case, databases with biological information hosted by ExPASy. In other words, we want to explore if and how much of the acknowledged use of these research technologies is neglected when measuring citations alone, and whether this differs between the four infrastructures. We expect that the number of references to the articles introducing these databases in general is roughly similar to the references to these technologies made in the text.

Table 1 Source publications

PROSITE	SWISS-2Dpage	HAMAP	ENZYME
Sigrist CJA_2010_Nucleic Acids Res		limaetal_2009_nucleic acid res	Bairoch_2000_nucleic_acid_res
Falquet L_2002_Nucleic Acids Res	Hooglandetal_2004_proteomics	Gattiker A_2003_computa biol chem	Bairoch_1999_nucleic_acid_res
Sigristetal_2002_briefings bioinformatics_Scopus	Hooglandetal_2000_NAR		Bairoch_1996_nucleic_acid_res
De Castro E_2006_Nucleic Acids Res	Hooglandetal_1999_NAR		Bairoch_1994_nucleic_acid_res
Hulo N_2006_Nucleic Acids Res	Hooglandetal_1999_electrophoresis		Bairoch_1993_nucleic_acid_res
Hoffman K_1999_Nucleic Acids Rest	Tonellaetal_1998_electrophoresis		
Sigrist CJA_2005_Bioinformatics	Hooglandetal_1998_NAR		
Hulo N_2008_Nucleic Acids Res	Appeletal_1996_NAR		
Hulo N_2004_Nucleic Acids Res	Sanchezetal_1996_electrophoresis		
Bairoch A_1997_Nucleic Acids Res_1 AND Bairoch A_1997_Nucleic Acids Res_2	Pasqualietal_1996_electrophoresis		
Bairoch A_1996_Nucleic Acids Res	Appeletal_1996_electrophoresis		
Bairoch A_1994_Nucleic Acids Res	Sanchezetal_1995_electrophoresis		
Bairoch A_1993_Nucleic Acids Res	Appeletal_1994_NAR		
Bairoch A_1992_Nucleic Acids Res	Appeletal_1993_electrophoresis		
Bairoch A_1991_Nucleic Acids Res			

To answer both research questions, measures are needed of the frequency with which researchers use a database and the frequency with which they cite it. The first type of data consists of usage data of the databases, which is based on the

encounter in their research. These databases form an interesting example with which to consider how the knowledge claims which are entailed in research technologies are transmitted within the scientific community. The databases under study are PROSITE, Swiss-2dPAGE, HAMAP and ENZYME. We have selected these databases because they are only accessible through the ExPASy server, in contrast to some of the other (ExPASy) databases which can be accessed through multiple servers[11]. This makes counting of visits feasible when one has access to the original log files.

PROSITE is a protein database (Sigrist et al, 2012). It consists of entries describing protein families, domains and functional sites as well as amino acid patterns, signatures, and profiles in them. The SWISS-2DPAGE database assembles data on proteins identified on various 2-D and 1-D PAGE maps. Each SWISS-2DPAGE entry contains textual and image data on one protein, including mapping procedures, physiological and pathological information, experimental data and bibliographical references (Hoogland et al, 2004). HAMAP is a system, based on manual protein annotation that identifies and semi-automatically annotates proteins that are part of well-conserved families or subfamilies: the HAMAP families. HAMAP is based on manually created family rules and is applied to bacterial, archaeal and plastid-encoded proteins, which are contained in the database under study (Lima et al, 2009). ENZYME is a repository of information relative to the nomenclature of enzymes. It is primarily based on the recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (IUBMB) and it describes each type of characterized enzyme for which an EC (Enzyme Commission) number has been provided (Bairoch, 2000).

The four databases differ somewhat from each other. Two (PROSITE and SWISS-2dPAGE) contain a great amount of data, generated by researchers worldwide, and collected and maintained by researchers from (a.o.) the Swiss Institute of Bioinformatics. The other two (HAMAP and ENZYME) contain a set of rules which are used to classify information in other protein sequence databases.

This paper aims to analyse firstly the extent to which citations to original articles provide an accurate representation of the usage of the databases with biological information hosted by ExPASy. We expect that the usage intensity (measured in number of visits to URL domains) is systematically related to the frequency of citations to the articles in which these research technologies are introduced: i.e.

[11] There may be some exceptions to this in the form of ExPASy mirror servers at some universities in several European countries, China, Australia, and Japan. The size of the weblogs of these mirror servers, however, is dwarfed by the size of the main server of ExPASy. These mirror servers were especially important in the times before quick internet facilitated easy access to the server based in Switzerland. In any case it is unlikely that the inclusion of the weblog data from these mirror servers would have made a difference in the distribution of the number of visits to the four databases. In contrast to the study by Jonkers et al (2012) the weblog data for the different directories used in this study was not cleaned by removal of visits from robots, web-crawlers etc. This may account for a substantial share of the reported web-traffic.

number of visitors which each of the directories that gives access to these databases receive. For the analysis of the ExPASy server weblog (Jonkers et al, 2012) use is made of the free software Funnel Web Analyzer developed by QUEST (2010). This data allows for the construction of an indicator of the number of visitors of these databases in the time period 2003-2008, which is used as a proxy for usage intensity.

The researchers responsible for establishing the biological databases under study request users to refer in their publications to one of a number of references mentioned on their website. Over the years, the responsible researchers have published articles with updates of and extensions to the databases. We use all the articles in order to cover all relevant references. For HAMAP we found two core references, for SWISS-2DPAGE thirteen, for PROSITE fifteen and for ENZYME five core references (see table 1).

Using the bibliometric databases[12] Scopus[13] and SCI[14], we retrieved all papers citing these articles in the period 2000-2011 (time of download June 2012). Both databases provide powerful analytical tools for citation analysis and although "*Scopus* is a database with criteria similar to those of *Thomson Reuters*, not only in the development of the collection but also in its coverage on the world level" (Moya-Anegón et al., 2007, p. 76), each database still shows differences in terms of collection policy. The *WoS* list of indexed journals is shorter than that of *Scopus*, while the time period covered by *WoS* is longer. Cited references in a large number of sources indexed in *Scopus* do not go back further than 1996. The implications of these two apparently different policies (depth versus breadth) are analysed by several information scientists (Fingerman, 2006; Ball & Tunger, 2006). This paper is mainly based on Scopus, because of its better coverage of Science Direct journals. This is relevant for our analysis, as we want to use a specific tool for full text analysis, which will be discussed below.

The number of in-text references to the infrastructures was analysed using the software "section search" of NEXTBIO (2012) offered through the SCIVERSE platform. This program analyses full texts of articles contained in the Science direct database (mainly journals owned by Elsevier) for the sections: *Title*, *Abstract*, *Introduction*, *Methods*, *Results*, *Discussion*, *Summary* and *Captions*. It

[12] Since both databases are available on the market, the number of papers comparing them from a scientometric perspective has been growing (e.g. López-Illescas et al., 2008; Gorraiz & Schlögl, 2007; Jacso, 2006).

[13] Scopus covers over 19,500 titles from more than 5,000 publishers worldwide. It includes coverage of 18,500 peer-reviewed journals and over 4.9 million conference papers, 400 trade publications and 350 book series. It provides 100 % coverage of Medline. On May 1, 2012, it contained about 47 million records, 70% with abstracts, of which 26 million records going back to 1996. [Scopus, 2012. http://www.scopus.com]

[14] *Thomson Reuters' Web of Science* covers over 12,000 research journals worldwide and provides access to "the *Science Citation Index* (1900-present), *Social Sciences Citation Index* (1956-present), *Arts & Humanities Citation Index* (1975-present), Index Chemicus (1993-present), andwww.thomsonscientific.com/products/ccr (1986-present), plus archives 1840 - 1985 from INPI." [Thomson Reuters, 2012. http://thomsonreuters.com].

does not cover the bibliography.[15] This search yields the list of articles and reviews in which one (or more) of the databases was mentioned in the text by the authors. As will be clear to the reader a search for the keyword "enzyme" will yield a large number of false positives as this word is not only used to refer to this database but also to a specific, and often researched, type of protein. Also a search for "enzyme database" yields false positives, as several other enzyme databases exist that are found through such a search.

Since NEXTBIO only analyses Science Direct journals, we refined our citation analysis. To do so we collected the smaller set of references made in Science Direct life science journals. We controlled whether all Science Direct[16] journals identified were covered in Scopus, and this proved to be the case, confirming the expectation that Scopus includes all Science Direct journals. This implies that the citation counting in Scopus covers all journals included in the NEXTBIO analysis in addition to potential references in journals not included in the Science Direct database. The next step was to compare the number of publications in which the authors refer to one of the databases in the full text with the citations of the source articles found in Scopus.

By comparing the citations made in Science Direct journals to the articles found through NEXTBIO's "section search" disregarding those that are also found through the citation analysis (M), an assessment of the extent that citation analysis leads to an underestimation of acknowledged use was made, using the following formula:

$$U\,(\%) = \left(1 - \frac{C}{C+M}\right) * 100\% \qquad (1)$$

U refers to underestimation (%); *C* refers to the number of citing Science Direct articles; and *M* refers to the number of articles mentioning the database in Science Direct journals (minus the publications also appearing in C). As the citation behaviour of authors publishing in Science Direct journals was expected to be similar to those of authors publishing in other journals, the expected total number of citations if all acknowledged reports of usage would have been reflected in citations, can be inferred.

The databases that will be presented in table 2 and 3 were selected because they are accessed only through the ExPASy server and could therefore serve to show the potential use of weblog analyses. To explore the usefulness of the proposed methodology further an additional 36 bioinformatic applications hosted on the ExPASy server were studied (Annex 1 provides a short description of each of the applications). Some of the applications to which the ExPASy server provides access (e.g. MARCOIL, pROC, PRATT, TMPred, TCS, T-Coffee, TagIdent, Swiss-PdbViewer, SwissParam, RAxML, PepPepSearch, PaxDb, OpenStructure,

[15] Reviews are included in addition to articles and for this reasons they were also included in our citation analysis.

[16] The Science Direct database contains over 2500 journals (primarily owned by Elsevier). Links on the following webpage provide information on coverage.

neXtProt, MyHits, MassSearch) were developed by other organisations, but they have also been analysed because they are hosted on the ExPASy server as well. Only thirteen of these 36 applications can be studied because of the limitations of the proposed approach. These thirteen are, apart from HAMAP and Swiss-2DPage: Msight; MIAPEGelDB; MALDIPepQuant; Make2D-DB II; HCD/CID Spectra merger; GlycosuiteDB; OpenStructure; MyHits; tagident; SwissParam; MARCOIL.

Results

We introduced an alternative measure for database use (see also Jonkers et al, 2012; Duin et al 2012,), which is independent of the academic literature. Table two shows that as expected the database which shows the highest usage intensity (in terms of the number of visits in the period 2003-2008) is also the database which is cited most frequently (PROSITE). Due to the small sample size we cannot do correlation analysis. But the data fit in the expected pattern, and the number of unique visitors is ten (HAMAP and Swiss 2DPAGE) to around thirty (PROSITE and Enzyme) times higher than the number of citations. More details about the existence and nature (linear or not) of the relationship cannot be derived from the available data.

Table 2 Citations (2003-2009) and visits (2003-2008)

	PROSITE	HAMAP	SWISS-2DPAGE	ENZYME
Citations in Scopus	2225	79	239	248
Visits	71890	914	3081	9194
Visits / citations	32	12	12.9	37
Log10 visits / log10 citations	1.45	1.56	1.1.49	1.66

Table 3 results data collection: citations and text mentions of the databases (2000-2011)

	PROSITE	HAMAP	SWISS-2DPAGE
Citations by articles/reviews all Scopus	4634	102	575
Citations in SD journals in Scopus	1000	16	52
Mentions in full text (minus references) of SD articles	1730	7	29
Mentions in full text without formal reference in Scopus	X	2	20
Total mentions + cites in SD journals in Scopus	X	18	72
Underrepresentation	X	11.1%	27.8%
Expected number of cites and mentions in entire Scopus	X	113	735

X: data not available

Table 3 presents a) the number of citations which were made to the source articles in which the four databases were introduced in Scopus between 2000 and 2011, b) in Science Direct Journals in Scopus in the same period. The table also includes the number of publications (articles and reviews from Science Direct journals) found through the full text section searches. It was expected that most of these mentions of acknowledged use would be found in the methods section, but this is certainly not exclusively so.

The second part of the analysis shows that the rate of underestimation found in the case of two of the four databases was 11.1% and 27.8% respectively. This indicates a) a substantial under-estimation of acknowledged use of e-research technologies through citation analyses and b) a considerable variation in the extent to which this underestimation occurs.

We find that 11 articles/reviews in Science Direct journals mention the HAMAP database in their full text. One of these is one of the original source articles, which leaves 10 after its exclusion. 7 of these have been published before 2012 and we decided to exclude this last year. The reason for doing so is that the online versions of the bibliometric databases used did not provide stable results for this year when measurements were made in the summer of 2012. Another motivation was that records for 2012 would not be complete as measurements were made before the end of this year. The total number of articles/reviews found in Scopus which cite one of the two source articles of HAMAP is 110, 102 of which were made in the years before 2012. Sixteen of these citations are made in Science Direct journals. Five of the ten articles which refer to the HAMAP database in the full text, do not cite either of the two HAMAP source articles. When excluding 2012, this is two out of seven. Some eighteen articles in Science Direct journals either cite one of the source articles of the HAMAP database, or mention it in the text. The total number of citations to the source articles in Science Direct journals is sixteen. Hence only a small underestimation of around 11% is found. As it is expected that citing behavior in other journals included in Scopus is similar to Elsevier journals, it is expected that there are around 113 articles/reviews which either cite HAMAP or refer to it in the text in the Scopus database.

A similar approach is followed to analyze the results from the citation and full text search for acknowledged use of the database SWISS-2DPAGE. 575 articles/reviews are found in Scopus which refer to one of the thirteen source articles. NEXTBIO finds 52 results in which Swiss-2DPage is found in the text (+ two false positives). 20 of these NEXTBIO results do not include a formal reference included in Scopus. The estimate for underestimation here is thus substantially higher at around 27.8%. Since authors publishing in Science Direct journals are expected to cite in a similar way as authors publishing in non-Science Direct Scopus journals, a total of 735 articles/reviews is expected to be present in the Scopus database that either cite the source articles of SWISS-2DPAGE, or mention the use of it in the text.

Considering the relatively large rate of underestimation of “acknowledged use” through formal citations, a manual analysis was performed of the articles that

mentioned Swiss 2DPAGE but did not cite any of the thirteen source articles. One expectation was that - as this database collects, stores and provides access to the empirical results of other studies - these 'non-citing' articles would refer to the underlying source articles instead. This, however, was not the case. Instead of including a formal reference, thirteen of these articles provided a URL to the Swiss-2DPAGE site. Two articles could not be accessed. Only five mentioned Swiss 2DPage in the text, while not presenting any acknowledgement (citation or URL) to their readers. [17]

Table 4 Underestimation of acknowledge usage by citation analysis for other ExPASy applications (2000-2011)

	Scopus cites	C	NEXTBIO	M	C+M	U (%)	U_1%
Quickmod	4	0	0	0	0	x	
MSight	81	12	5	3	15	20	4
MIAPEGelDB	7	1	0	0	1	0	0
MALDIPepQuant	5	2	0	0	2	0	0
Make2D-DB II	15	3	2	0	3	0	0
HCD/CID spectra merger	38	7	0	0	7	0	0
GlycoSuiteDB	120	17	4	1	18	6	1
FindPept *	45	13	31	28 (26)	41 (39)	68*(66)	
FindMod *	182	39	30	25 (23)	64 (62)	39* (37)	
PeptideMass*	175	59	91	59 (54)	118 (114)	50* (48)	
MARCOIL	101	20	12	1	21	5	1
T-coffee	2706	820	x	x	x	x	
tagident	16	0	26	26	26	100	61
Swiss-PdbViewer	5910		x	x	x	x	
SwissParam	2	1	0	0	1	0	0
RAxML	902	167	x	x	x	x	
PaxDb	0	0	0	0	0	x	
OpenStructure	3	0	0	0	0	0	0
MyHits	20	6	18	18	24	75	47

M = articles containing NEXTBIO in text references but no SD citations; C = Scopus cites included in Science Direct; *As mentioned in the methodological section the analysis for these four applications is incomplete and the real percentage of underestimation is therefore expected to be considerably lower.

Unfortunately the NEXTBIO software has some limitations, which makes it impossible to do the same analysis for the more popular PROSITE database. In contrast to the small numbers of articles in which HAMAP or SWISS-2Dpage

[17] For authors using bibliometric data it may be interesting that of the 518 SD publications that were found through NEXTBIO to mentioning the use of the Scopus databases in their full text, only 12 included the URL (though in some the URL may have been in the reference list).

were mentioned, a total of 1730 publications (in Science Direct journals) were found that mention PROSITE somewhere in the full text (minus the references). Unfortunately the software only shows a limited number of around 776 of these 1730 bibliographic references. It was therefore not possible to repeat the analysis conducted for the other databases. In total, the source articles in which the PROSITE database was introduced, received 4643 from publications included in Scopus. 1000 and 661 of these were made in Elsevier journals.

For Peptidecutter, Peppepsearch, NextProt and Masssearch an appropriate source article could not be identified. Some applications also had to be excluded such as compute pi/MW, sulfonator, myristoylator, blast, biochemical pathways, allall, pROC, PRATT and TCS because they gave too many unrelated hits due to name ambiguity similar to the "Enzyme database". Multiident received 153 Scopus citations and 28 SD citations. Given these numbers one would have expected a considerable number of in-text references as found through NEXTBIO. However none were found – though with the alternative spelling "multi-ident" one in-text reference was identified as well as five unrelated articles as the name was not sufficiently unambiguous. For this reason this application was also excluded from table 4.

The four applications Findpept, FindMod, PeptideMass and Peptidecutter have, apart from in the article analysed, also been introduced in a book chapter. The URLs giving access to these applications suggest this book chapter as a potential reference. This chapter, which is not included in Scopus and could therefore not be studied, received over 1400 Google scholar citations. Part of these citations is likely to have come from Scopus SD journals. This suggests that a considerable number of the articles with an in-text reference as found through NEXTBIO which were not found to have a corresponding SD citation may have included citations to the book chapter. While they are mentioned in the table, these results are therefore not considered reliable. For the applications Findpept, FindMod, and PeptideMass an alternative M was created through a manual search of the reference lists of these alternatives. Where a reference to the book chapter was found this was deducted from the original M and presented between brackets in the table. The rate of underrepresentation remains high, but would have been lower if it would have been possible to assess the number of Scopus cites (C) to the book chapter. As was the case for Swiss-2DPage part of this under-representation is caused because authors refer to the URL rather than including a formal citation.

Some applications such as Swiss-model, RaXML, Swiss-PDBviewer and T-coffee were too popular to be studied through this approach as was the case for the PROSITE database. They received 7707; 2706, 5910 and 902 Scopus citations respectively, but the in-text references yielded by NEXTBIO could not be analysed in detail. For the applications Glycanmass, Glycomod, GPSDB, PLcarber; protscale; protparam the suggested reference is the same general article. This article received 924 citations in Scopus and 204 citations in Science direct journals. However, some of the applications yielded too many NEXTBIO in text

references so that this "group" of applications could not be studied either as was the case for PROSITE. The reason why they are included in the table is that this helps to make an assessment of the relative share of SD citations in the total Scopus citation coverage in this field.
Some applications such as Pax-DB, OpenStructure, Quickmod, MIAPEgelDB and MALDIPepQuant and HCD/CID spectramerger, do not yield any in text references through the NextBio search. As a consequence the estimated rate of under-representation of acknowledged use is zero. One potential explanation is that some of these applications were introduced very recently and there has not yet been much time to cite these in either the references of articles or in the text. This reasoning lies behind the exclusion of PaxDb of which the source article was published in 2012, which is after the period in which the citations were measured.
The rate of underestimation of the other applications studied was, 5% for Marcoil, 6% for GlycoSuiteDB to around 20% for MSight. The underestimation of the acknowledge use of both MyHits (75%) and Tagident (100%) is high in comparison to the other applications as well as the databases studied in table 4. In the case of Tagident all citations were made in non SD journals. While there appears a strong underestimation of acknowledged use in the case of this application, in reality it can never be 100%, as the source article is referred to in non-SD journals. For this reason we adapt our indicator somewhat to provide a lower bandwith of the estimated underestimation (U_l). For this we take instead of "C" the number of Scopus citations. In the case of Tagident U_l is 61 %, in the case of Myhits it is 47 %, indicating that the underestimation of Tagident lies between at least 61 and 100% and the underestimation of Myhits lies between Myhits lies between at least 47 and 75%. According to this /(very conservative) estimate of underestimation the lower boundaries of the underestimation of HAMAP and Swiss-2Dpage would be 2 and 3 %.

Discussion and Conclusion

While citations appear systematically related with usage measured through unique visitors, it is not yet clear how these indicators are related. We find that a considerable share of the acknowledged use in research articles is not captured by citation analyses. The degree of underestimation varies between the databases and applications studied.
Both observations raise some concern over the accuracy, completeness and suitability of the sole use of citation analyses for measuring the impact of e-research infrastructures. This concern also potentially extends to other types of knowledge claims. The observed variations may be explained using existing citation theories. Publications that have already received a large number of citations may be more citable than those cited less, a derivation of the Matthew effect (Merton, 1995). Conversely, if the technology has become ubiquitous, researchers may consider that they no longer need to cite knowledge claims which have become "common knowledge". This echoes an argument made in Martin & Irvine (1983). A combination of these explanations might be used to explain the

observed relation between usage (as measured through weblog analysis) and citations. Neither of these explanations, however, can explain the variation in the rates of underestimation of acknowledged use through citation analysis between applications. It does appear from table 4 that the underestimation of young applications which have not yet received a substantial number of citations tends to be zero.

A more in depth exploration of the instances of acknowledged use that were not reflected in citations for the case of the Swiss 2Dpage database, revealed that in a large share of these instances, the authors had referred to the URL that provided access to the application in either the reference list or inside the text. This type of acknowledgement is more difficult to analyse than formal citations, but it may nonetheless be a common way for researchers to refer to electronic databases and applications.

The approaches highlighted in this paper: 1) "web usage statistics derived from the analysis of web logs", 2) "citation analyses" and 3) "the analysis of in-text references to specific research infrastructures" do not provide a complete insight in the actual scholarly usage of e-research infrastructures and their impact. Not all usage will be acknowledged by researchers in the reference list or as in-text reference. Furthermore, researchers may also be using technologies without being fully aware of it. A discussion of the HAMAP database studied in this paper will serve to explain this. It is important to realize that there is a difference between 1) first order users, who make direct use of, for example, the HAMAP rule book and 2) second order users who, while not making use of the rule book or HAMAP database, do make use of the information of HAMAP annotated proteins contained in other protein databases. When referring to usage, this paper only referred to the first order users. However it is important to realize that the actual use and impact of such technologies may be extended beyond its direct use.

This is one of the first articles that introduce an (exploratory) comparative analysis of in text reference analysis and citation analysis. The main part of the analysis is limited to journals included in the Science Direct database. It is clear that the proposed approach to the analysis of in-text references through the use of NEXTBIO has its limitations: especially with reference to name-ambiguity and popular applications. The second limitation can probably be solved relatively easily through alternative approaches to the analysis of in-text referencing. The first limitation is more difficult to solve. In the case of Tagident the underestimation appears to be 100 %. This is not an accurate reflection of reality since citations have been made in non-SD journals. This suggests a weakness of the proposed approach when dealing with applications which had still received only a very small number of citations in the period measured. In-text reference analysis which is not restricted to SD journals will not face this problem.

Analysts have argued that it is somehow "unfair" to compare citations to reviews with those to theoretical or empirical papers. Some may argue that this argument can be extended to publications introducing new methods, research instruments or research infrastructures. Normalisation is often used to account for differences in

the average frequency of citation to different document types (Moed et al, 1995, Rehn & Kronman, 2008). Due to the structure of the bibliometric databases methodological papers, papers introducing research instruments or research infrastructures are normally not identified as such. Therefore they are also not normally subjected to such normalisations. Furthermore a complete theoretical justification for assigning a different value to citations received by different document types is still lacking. The differential underestimation of "acknowledged use" via citation measurement might provide part of such a justification if the rate of under-acknowledgement differs systematically between types of knowledge claims. In this paper an indication is found that citation analysis underestimates the acknowledged use of some types of knowledge claims (in this case biological databases). Further analysis of the varying degree of underestimation of different knowledge claim types could provide a way forward to a more complete justification for both citation normalisation and/or the use of alternative metrics in assessing the impact of different knowledge claim types. As highlighted in a recent Nature materials editorial (2012), the merit of the latter should be evaluated with care for: "Not everything that can be counted counts and not everything that counts can be counted". This oft used and paraphrased quote is sometimes attributed to Cameron (1963), but often also to Albert Einstein's blackboard writing.

Acknowledgements

The Spanish Ministry of Economics and Competitiveness funded the project of which this paper forms part through the grant: CSO2011-23508. The research of the third author was supported by the Juan de la Cierva (JDC)-MICINN program of the same Ministry. SIB Swiss Institute of Bioinformatics allowed for the use of the server web log data used for part of this analysis. We would also like to thank Felix de Moya Anegón for introducing us to the NEXTBIO application "section search" and Isidro F Aguillo for advice on the use of Quest's Funnelweb software. Researchers at the Centre for Science and Technology Studies of Leiden University (NL) provided stimulating ideas in discussions during a research stay of one of the authors. The usual disclaimer applies.

References

Ball, R., Tunger, D. (2006) Science indicators revisited - Science Citation Index versus Scopus: A bibliometric comparison of both citation databases, *Journal Information Services and Use*, 26: 293-301

Bairoch A. (2000) *The ENZYME database in 2000*, *Nucleic Acids Res* 28:304-305.

Bornmann, L. & Daniel, HD (2008) What do citation counts measure? A review of studies on citing behavior, *Journal of Documentation,* 64 (1): 45-80

Cameron, WB. (1963) *Informal Sociology: A Casual Introduction to Sociological Thinking*, New York, Random House

Cozzens, SE (1989) What do Citations count? The rhetoric-first model, *Scientometrics*, 15 (5-6): 437-447

Cronin, B. (1984) *The Citation Process. The Role and Significance of Citations in Scientific Communication*, Taylor Graham, Oxford.

De Jong, S., van Arensbergen, P. Daemen, F., van der Meulen, B., van den Besselaar, P. (2011) Evaluating research in its context: an approach and two cases. *Research Evaluation* 20 (1): 61-72.

De Solla Price, D. (1976) A General Theory of Bibliometric and Other Cumulative Advantage Processes, *Journal of the American Society for Information Science*, 27 (5-6): 292-306 1976

Duin, D., King, D., van den Besselaar, P. (2012) Identifying Audiences of E-Infrastructures - Tools for Measuring Impact, *PLOS ONE*, 7(12)

Editorial (2012) Alternative metrics, *Nature Materials,* 11, 907

Fingerman, S. (2006) Web of Science and Scopus: Current Features and Capabilities, *Issues in Science and Technology Librarianship*, DOI:10.5062/F4G44N7B

Garfield, E. (1998) Random thoughts on citationology, its theory and practice, *Scientometrics*, 43 (1): 69-76

Gilbert, N. (1977) Referencing as Persuasion, *Social Studies of Science*, 7 (1)

Gorraiz, J., & Schlögl, C. (2007) Comparison of two counting houses in the field of pharmacology and pharmacy. In *Proceedings of the international conference of the international society for scientometrics and informetrics,* 11: 854–855.

Hoogland C., Mostaguir K., Sanchez J.-C., Hochstrasser D.F., Appel R.D. (2004) SWISS-2DPAGE, ten years later. *Proteomics*, 4(8), 2352-2356.

Jacso, P. (2006) Evaluation of citation enhanced scholarly databases. *Journal of Information Processing and Management*, *48*(12), 763–774.

Jonkers, K., De Moya Anegon, F., Aguillo, F. (2012) Measuring the use of research infrastructures as an indicator of research activity, *Journal of the American Society of Information Science and Technology*, 63 (7): 1374–1382

Lima, T., Auchincloss, AH., Coudert, E., Keller, G., Michoud, K., Rivoire, C., Bulliard, V., de Castro, E., Lachaize, C., Baratin, D., Phan, I., Bougueleret, L., Bairoch, A. (2009) HAMAP: a database of completely sequenced microbial proteome sets and manually curated microbial protein families in UniProtKB/Swiss-Prot, *Nucl. Acids Res.* 37 (suppl 1): D471-D478. doi: 10.1093/nar/gkn661

López-Illescas, C., Moya-Anegón, F., Moed, HF. (2008) Coverage and citation impact of oncological journals in the Web of Science and Scopus. *Journal of Informetrics*, *2 304–316*

Lowry, OH., Rosebrough, NJ, Farr, AL, Randall, RJ. (1951) Protein Measurement with the Folin Phenol Reagent, *Journal of Biological Chemistry*, 193:265-275 .

Martin, BR., Irvine, J. (1983) Assessing basic research: Some partial indicators of scientific progress in radio astronomy, *Research Policy*, 12 (2): 61–90

Merton, RK (1995) The Thomas Theorem and the Matthew Effect, *Social Forces*, 74 (2)

Moed, H.F.; Colledge, L.; Reedijk, J.;Moya-Anegón, F.; Guerrero-Bote, V.; Plume, A.; Amin, M. (2012**)** Citation-based metrics are appropriate tools in journal assessment provided that they are accurate and used in an informed way**.** *Scientometrics*, 92 (2) 367-376**.**

Moed, H.F., De Bruin, RE., Van Leeuwen, TN (1995) New bibliometric tools for the assessment of national research performance: Database description, overview of indicators and first applications, *Scientometrics* 33 (3): 381-422

Moya-Anegon, F., Chinchilla-Rodríguez, Z., Vargas-Quesada, B., Corera-Álvarez, E., Muñoz-Fernández, F. J., González-Molina, A. (2007) Coverage analysis of Scopus: A journal metric approach. *Scientometrics*, 73(1), 53–78.

NEXTBIO (2012) section search, retrieved from http://www.applications.sciverse.com/action/appDetail/293416

Quest*. (*2010*). Funnel Web Analyzer®—overview. Retrieved from http://www.quest.com/funnel-web-analyzer/index.asp*

Rehn, C., & Kronman, U. (2008). *Bibliometric handbook for Karolinska Institutet V1.05* http://ki.se/content/1/c6/01/79/31/bibliometric_handbook_karolinska_institutet_v_1.05.pdf.

Senker, J. (1995) Tacit knowledge and Models of Innovation, Industrial and Corporate Change,

Scopus (2012) http://www.scopus.com

Science direct journal coverage (2012) http://www.info.sciverse.com/sciencedirect/content/journals/titles

Sigrist C.J.A., de Castro E, Cerutti L, Cuche BA, Hulo N, Bridge A, Bougueleret L, Xenarios I. (2012) *New and continuing developments at PROSITE*, *Nucleic Acids Res.* doi: 10.1093/nar/gks1067

Thomson Reuters (2012) http://thomsonreuters.com

Van Raan, A. (2005) Measuring Science, in Moed, HF., Glänzel, W., Schmoch, U., *Handbook of Quantitative Science and Technology Studies,* Dordrecht: Springer

Weingart, P. (2005) Impact of bibliometrics upon the science system: Inadvertent consequences? *Scientometrics*, 62(1): 117-131