Measuring the scientific impact of e-research infrastructures:

a citation based approach[1]

Jonkers K[a,] , Derrick GE[a,b], Lopez-Illescas C[d]. Van den Besselaar, P.[d]

[a]CSIC Institute for Public Goods and Policies, Department of Science and Innovation dynamics, C/Albasanz 26-28, 28033 Madrid, Spain, correspondence: koen.jonkers@csic.es

[b] Health Economics Research Group (HERG), Brunel University, London, United Kingdom

[c]SCImago group. Department of Information Science. University of Extremadura, Badajoz, Spain.

[d] VU University Amsterdam, Department of Organization Science and Network Institute, Amsterdam, The Netherlands

Abstract.

This micro-level study explores the extent that citation analysis provides an accurate and representative assessment of the use and impact of bioinformatics e-research infrastructure. The bioinformatic e-research infrastructure studied offers common tools used by life scientists to analyse and interpret genetic and protein sequence information. These e-resources therefore provide an interesting example with which to explore how representative citations are as acknowledgements of knowledge in the life sciences. The examples presented here suggest that there is a relation between number of visits to these databases and number of citations; however, a parallel finding shows how citation analysis frequently underestimates acknowledged use of the resources offered on this e-research infrastructure. The paper discusses the implications of the findings for various aspects of impact measurement and also considers how appropriate citation analysis is as a measurement of knowledge claims.

Keywords: citation analysis, research infrastructure, evaluation, bioinformatics

Introduction

[1] Version of November 2013

Submitted for publication.

An earlier version was published as K Jonkers, GE Derrick, C Lopez Illescas, P van den Besselaar, Are citations a complete measure for the impact of e-research infrastructures. In: Juan Gorraiz, Edgar Schiebel (eds), *Proceedings ISSI 2013*. Vienna: AIT, 136-151.

This paper explores to what extent citation analysis provides an accurate assessment of the usage of e-research infrastructures in scientific articles. In general, citations are used to measure the "impact" of knowledge claims due to the easy accessibility 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 article level metrics via social media, have been promoted as ways to assess the broader impact of research, among many others (De Jong et al, 2011). However citation based indicators are still the dominant approaches for measuring the scholarly impact of research.

Although the scholarly use and impact of research technologies, as with scientific knowledge claims, can be assessed using citation analysis, for many research infrastructures, citations may not be a sufficient way with which to represent 'impact'. Where using citations can 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. These include log-files that measure the website visits to research infrastructures (Jonkers et al, 2012; Duin et al 2012). Patents are also a method of considering the importance of research instruments in biotechnological innovation processes (Senker, 1995). Indeed a full assessment of the impact of e-research infrastructures should include an analysis of the references in patents.

This article aims to investigate the extent that citations provide an adequate and representative assessment of the use and impact of bioinformatics e-research infrastructures. Firstly, we investigate whether citations to original articles introducing a research infrastructure provide an accurate representation of use and impact. This is analysed by studying whether the "intensity of use" (as measured by the number of visits to the URLs of the infrastructures' domains) is related to the number of citations to articles in which these research infrastructures were introduced. If a strong positive relationship exists, citations would therefore be a strong indicator of usage. Secondly, publications may include in-text references to the research infrastructure.- This article investigates the extent that acknowledged use of research technologies is neglected when only citation counts are used. These research questions were explored, using research e-resources (databases with biological information and bioinformatic tools) hosted by ExPASy (described below in the Methodology section).

Theoretical background: Why citations?

In order to sufficiently interpret any results obtained in this study, we examine the literature underpinning the theories of citations, citing behavior and consider what citations represent. Two main bodies of theory underlie the use of citations for analysing research output. First, the "normative theory of citations" simply states that researchers cite documents that are, (1)

relevant to their topic, (2) provide useful background for their research, and/or (3) acknowledge an intellectual debt (Bornmann & Daniel, 2008). In this theory, science is viewed as rewarding quality in research by concentrating on rewarding the traits of individual articles (Baldi, 1998). Cronin (1984) argues that citations perform a scholarly communication function between texts in line with the normative theory of citations. Here, citations can also indicate a measure of reward for past work or scientific status (Martin & Irvine, 1983). In summary, the normative theory of citations emphasises the awarding of citations based on *"what is said"* (Lokker et al, 2012).

The second theory, whilst not mutually exclusive to the first, emphasises that the decision to award a citation to other articles are not free from personal or social influences. Therefore the "social-constructivist theory of citations" states that citations represent a social process. As such, citations are used as an aid for persuasion (Gilbert, 1977; Cozzens, 1989) and reflect factors such as the social hierarchy of a field or an author's traits (Lokker et al, 2012). The social constructivist theory provides an explanation for why people would add additional citations, beyond those that could be expected on the basis of the normative view of citations. In contrast to the normative theory, the social constructivist theory concentrated on the "who one is" factor for the decision to award a citation.

Additionally, where citation analysis is becoming an increasingly prominent feature of scientific evaluation, authors are inclined to cite to raise the visibility of their own work and/or that of their collaborators, colleagues or members of their own "citation cartel". The motivations to recognise a knowledge claim via awarding a citation, differs between authors and references.

It is too simplistic to think that citing behaviour can be sorted into just two theories. Some authors consider it impossible to develop a convincing, overarching 'theory of citations' (Weingart, 2005), as citing behaviour; and citations as indicators for quality are considered as two separate issues. The more aggregated, the more citations become detached from actual citing behaviour. Therefore, the more useful they can be for investigating research quality. Despite the limitations of citations as quality measures being extensively debated and highlighted in the literature, there are still a number of citation characteristics that contribute to our understanding of what they actually represent. Understanding these characteristics can guide decisions about the appropriate application of citation analysis or, alternatively, when a complementary of different evaluation tool is required. Unlike previous contributions, this article is solely concerned with the issue of when authors may neglect to cite particular knowledge claims and yet still explicitly state their usage within their article.

There are a number of potential explanations for these "missed citations" to knowledge claims in the academic literature. For example, the origins of knowledge claims can be lost over time

as new (arguably improved) claims emerge. These original knowledge claims may then be absorbed into the common knowledge of a research discipline or even of the general public (Martin & Irvine, 1983). This "knowledge" thereby becomes general and seemingly not requires further referring to the original knowledge claim. Researchers who subsequently use the original knowledge claim may either: (1) not be aware of the existence of a citable item related to the original knowledge claim; or (2) consider it superfluous. Simple "forgetting" is another obvious motivation for not including a citation; however alternatively it may be that the author regards the knowledge claim in question as not worthy of a citation, therefore not awarding a citation. Finally, and most importantly, authors may be using alternative, less traditional forms of academic acknowledgements.

Not all types of knowledge claims receive an equal number of citations (Martin & Irvine, 1983). Reviews, for example, tend to receive more citations than articles (Asknes, 2005; Moed et al, 1995) and articles containing information about methods receive more citations than articles presenting new data or arguments. Peritz (1983), for example, showed that methodological papers in sociology were more frequently cited when compared to non-methodological papers. The same relationship is also seen in the life sciences.. Indeed, the most cited article of all time (*Protein measurement with the folin phenol reagent*) was published in 1951 and had gained 299,133 "WoS citations" by Dec 2012. This article outlines a commonly used method in biochemistry used to determine protein concentrations i.e. The Lowry method (Lowry et al, 1951; Garfield, 1998). The Lowry paper presents a commonly used biochemical technique, and whenever this method is applied by other researchers, they are *obliged* to reference the original knowledge claim (the Lowry paper) in their subsequent scientific articles. This scientific obligation in part, explains the high number of citations accumulated by Lowry.

This paper aims to analyse the extent to which citations to original articles provide an accurate representation of the usage of the e-resources (databases and applications). The cases selected are the e-resources hosted by ExPASy, a commonly used server hosting specialised proteomics e-resources used in the life sciences. It is expected that the "usage intensity" (as measured in number of visits to the URL domains) is systematically related to the frequency of citations to the articles in which these research technologies are introduced. If so, then citations would be considered as a strong indicator of usage. In other words, the ratio of use (measured as visits to the site) to citations is expected to be equal for the four databases analysed in this part of the study. The second aim of this article is to explore the extent to which citations are an adequate representation of the in-text references to e-research technologies. By addressing this aim, the extent that the acknowledged use of these research technologies is neglected when measuring citations alone is investigated. In addition, the extent that this differs between several resources is examined. It is expected that the number of references to articles introducing these e-

resources is proportional to the number of mentions to these technologies made in the text of articles.

Data and Methodology

Introducing ExPASy

This study used e-resources (databases and applications) hosted by the Expert Protein Analysis Server, ExPASy. These e-resources, developed and maintained by the Swiss Institute of Bioinformatics, are used by life scientists to analyze and interpret genetic and protein sequence information for their research. These e-resourcess therefore provide a unique opportunity with which to consider how the knowledge claims entailed in research technologies are transmitted within the life science community. The databases under study in the first part of this paper are; (1) PROSITE, (2) SWISS 2D-PAGE, (3) HAMAP, and (3) ENZYME. A full list and description of the databases used in this part of the study are presented in Table 1 and are described below.

Table 1. Descriptions of databases used in this study

	Name	Description of database contents
1	PROSITE	*Contains informationabout protein families, domains and functional sites*
2	Swiss 2D-PAGE	*Contains information about proteins identified on 2D- & 1D- PAGE maps*
3	HAMAP	*Classification and annotation system for protein sequences. Contains manually curated protein family profiles and annotation rules*
4	ENZYME	*Contains information about the nomenclature of enzymes*

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 2D-PAGE database assembles data on proteins identified on two-dimensional polyacrylamide gel electrophoresis (2D-PAGE). This database was created and maintained by the University Hospital of Geneva in collaboration with the Department of Medical Biochemistry of Geneva University. Its information is currently available via the EXPASY server (Appel et al, 1994). Each SWISS 2D-PAGE entry contains textual and image data for a protein. This includes 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. These are known as the "HAMAP families". HAMAP is based on manually created family rules and is applied to bacterial, archaeal and plastid-encoded proteins (Lima et al, 2009). ENZYME is a repository of

information relative to the nomenclature of enzymes. It is based on the recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (IUBMB). Specifically, ENZYME describes each type of characterized enzyme for which an EC (Enzyme Commission) number has been provided (Bairoch, 2000).

Two of the selected databases, PROSITE and SWISS 2D-PAGE, contain large amounts of data. In the case of PROSITE this has been developed in house by the Swiss Institute of Bioinformatics, in the case of SWISS 2D-PAGE the data has been generated by researchers worldwide and then collected and maintained by the Swiss Institute of Bioinformatics. The other two databases, HAMAP and ENZYME, each contain a set of rules which are used to classify information in protein sequence databases.

These four databases were selected based on their exclusive accessibility through the ExPASy server. This makes recording the number of visits feasible when one has access to the original log files. A small part of the user traffic is channeled through ExPASy mirror servers in among others China, Australia, and Japan. These mirror servers were especially important in the times before quick internet facilitated easy access to the server based in Switzerland. They no longer play a role at present. The size of the weblogs of these mirror servers is dwarfed by the size of the main server of ExPASy. 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.

Measuring 'Usage Intensity'

For both research questions, measures were needed that reflected the frequency of researcher use for a database as well as a measure of the frequency of their citations. The measure for database usage was based on the number of visitors to each directory giving access to these databases. To analyse the ExPASy server weblog, the free software Funnel Web Analyzer developed by QUEST (2010) was utilized (Jonkers et al, 2012). This data allowed for the construction of an indicator reflecting the number of visitors to these databases for the time period 2003-2008. This indicator was then used as a proxy for 'usage intensity'. 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 accounts for a substantial share of the reported web-traffic.

Tracking citations to each database investigated

Within the guidelines for using these biological databases, users are requested to include at least one of a number of suggested references in any resulting publications. These suggested references for users are listed on each of the database websites. There have also been a number of additions to the list of suggested references where researchers have published articles with updates of, and extensions to the databases. In this study, all relevant articles were used in order to cover all relevant references. For HAMAP we found two core references, for SWISS 2D-PAGE thirteen, for PROSITE fifteen and for ENZYME five core references (see Table 2).

Using Scopus and SCI, all papers citing these articles in the period 2000-2011 (time of download June 2012) were retrieved. Both SCI and Scopus provide powerful analytical tools for citation analysis. However, although *Scopus* is a database with journal inclusion criteria similar to *Thomson Reuters' WoS* (SCI), and similar in its coverage on the world level (Moya-Anegón et al., 2007, p. 76), coverage differences between the two remain. In particular, WoS indexes less journals (breadth) than Scopus but has indexed its journals over a longer period of time (depth) as Scopus sources are not commonly indexed prior to 1996. The implications of these two distinctions between databases (depth versus breadth) has been investigated extensively (Fingerman, 2006; Ball & Tunger, 2006). For the purposes of this paper, Scopus was selected for use due to its better coverage of *Science Direct* journals. This coverage has direct implications for this study, due to the nature of the text analysis component of this study described below.[2]

Identifying and collecting in-text mentions

The number of in-text mentions to the e-resources were analysed using "section search" tool of the software, NEXTBIO (2012), which is offered via the SCIVERSE platform. This software allows for the analysis of the full text of articles contained in the *Science Direct* database, which comprises journals owned by Elsevier. The software searches the article sections: *Title*, *Abstract*, *Introduction*, *Methods*, *Results*, *Discussion*, *Summary* and *Captions,* but does not cover the bibliography.

[2] 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). *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). *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), and www.thomsonscientific.com/products/ccr (1986-present), plus archives 1840 - 1985 from INPI." (Thomson Reuters, 2012.).

Table 2 Suggested source publications

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_briefingsbioinformatics_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			

The text analysis for this study yielded a list of articles and where at least one of the databases was mentioned in the text[3]. However, there are a number of important characteristics of the results generated by this text analysis software to consider. For example, a search for the keyword "enzyme" results in a large number of false positives generated. This is because the word "enzyme" is not only used to refer to the database ("ENZYME") but also to a protein characteristic. Further, any search for the words "enzyme database" also yields false positives, as there are several other existing enzyme databases that are identified in this kind of search.

As the software tool NEXTBIO, described above, only analyses *Science Direct* journals, the analysis of citations was refined by collecting the smaller set of references listed in Science Direct journals for the life sciences. In addition, the journals included were checked for inclusion in Scopus. This was confirmed. This confirmation also implied that the citation counted by Scopus is conducted for all journals included in the NEXTBIO analysis described above. This was also the case for any potentially additional references listed in journals that are not indexed by the *Science Direct* database.

Citation analysis

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

The number of publications that refer to one of the e-resources under investigation in the article's full text and the citations to source articles found in Scopus was compared.

An assessment of the level of underestimation of acknowledged use of knowledge claims was estimated by comparing citations made in *Science Direct* journals to the articles found via NEXTBIO's "section search." The citations that were found through the citation analysis (M) were disregarded. As such, the following formula used:

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

Where *U* refers to Rate of 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 (this does not include publications also appearing in C).

The citation behaviour of authors publishing in *Science Direct* journals was expected to be similar to those of authors publishing in other journals. Therefore the expected total number of citations can be inferred, taking into account that all acknowledged reports of usage would have been reflected in the citations.

The databases presented in Tables 3 and 4 below were selected, because they are only accessible via the ExPASy server. This means that the number of visits can be determined exactly. The tables show the potential of the use of weblog analyses.

To explore the usefulness of the methodology proposed above, an additional 36 bioinformatic e-resources hosted on the ExPASy server were studied. The ExPASy website (2012) provides access to and descriptions of each of the additional applications. 13 of these e-resources could be analysed within the framework of the methodological approach proposed in this paper: Msight, MIAPEGelDB, MALDIPepQuant, Make2D-DB II, HCD/CID Spectra merger, GlycosuiteDB, OpenStructure, MyHits, tagident, SwissParam and MARCOIL. In addition to the proteomic tools and databases that are the central focus of this paper, we also tested the method for a non-expasy bioinformatic tool Scratchpads.

Results

An alternative measure for database use, which is independent of monitoring academic literature, has previously been introduced (Jonkers et al, 2012; Duin et al 2012,). Table 3 shows that the database showing the highest usage intensity, or the highest number of visits for the period 2003-2008, is also the database cited most frequently (PROSITE). Unfortunately, a small sample size prevents a correlation analysis being conducted using this data. However, the

data shown is consistent with an expected pattern where the number of unique visitors is 10 to 30 times higher than the number of citations.

Table 3. 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 4 presents the following data: (1) the number of citations made to source articles when the four databases were indexed by Scopus 2000-2011; and (2) the same data for *Science Direct* journals indexed by Scopus for 2000-2011. Table 4 also includes the number of publications found via the full text section searches using the software tool NEXTBIO. It was expected that the majority of these in-text mentions of acknowledged database use would occur within the methods section. This, however, was not always the case.

Table 4. Citations and in-text mentions of the databases (2000-2011)

	PROSITE	HAMAP	SWISS-2DPAGE
Citations by articles/reviews all Scopus	4634	102	575
Citations to SD journals in Scopus	1000	16	52
In-text mentions of SD articles(without bibliography))	1730	7	29
In-text mentions without a formal reference in Scopus	-	2	20
Total in-text mentions + cites in SD journals in Scopus	-	18	72
Rate of underrepresentation (U)	-	11.1%	27.8%
Expected number of cites and in-text mentions in entire Scopus	-	113	735

- : data not available

Analysis of the Rate of Underestimation (U) found underestimation for two of the four databases; U equalled 11.1% and 27.8% respectively. This indicated a substantial level of under-estimation for the acknowledged use of e-research technologies (knowledge claims) through citation analysis and also a considerable variation in the extent that this underestimation occurs between databases.

Eleven articles/reviews in *Science Direct* journals mentioned the HAMAP database in their full text. One of these articles is the original source, but ten other articles also mentioned HAMAP. Of these ten, seven were published before 2012. The year 2013 was excluded as the online versions of Scopus had not yet provided complete records for this year.

The total number of articles in Scopus which cite one of the two HAMAP source articles was found to be 110. Of these citations, 102 occurred before 2012 and 16 were in *Science Direct* journals. In total, five out of the ten that contain in-text mentions of the HAMAP database do not cite either of the two HAMAP source articles in their reference lists. If 2012 is excluded, this figure becomes 2/7 articles that do not correctly cite the suggested HAMAP references in their bibliographies. In addition, 18 articles from *Science Direct* journals cite either one of the suggested source articles from the HAMAP database, or mention it in the text (in-text mentions). Finally, a total of 16 citations were found to the suggested source articles in *Science Direct* journals. Here, only a small rate underestimation was found (11%). It was assumed that the citing behavior for Scopus-indexed and other Elsevier journals is similar. Therefore an expected calculation of 113 articles was expected to have either directly cited or used in-text mentions to HAMAP for the entire Scopus database.

When analyzing SWISS-2DPAGE, a similar approach was used to investigate the extent of citation and in-text mentions for acknowledged use. A total of 575 articles are identified in Scopus referring to one of the 13 suggested source articles listed in Table 2. The software tool used to investigate in-text mentions, NEXTBIO, found a total of 52 in-text mentions of Swiss-2DPage (two false positives were excluded). From these results, 20 articles did not contain a corresponding, formal reference in Scopus. The rate of underestimation therefore was substantially higher at 27.8%. As authors who publish in *Science Direct* journals are assumed to cite their sources in a similar way to authors publishing in *non-Science Direct* journals, 735 articles were expected from Scopus that would either cite or contain an in-text mention of SWISS 2D-PAGE.

Considering the relatively large rate of underestimation for the acknowledged use of knowledge claims using formal citations, a manual analysis was performed of articles mentioning but not citing any of the thirteen suggested source articles for SWISS 2D-PAGE. It was possible that, given the nature of the database, which provides access to the empirical results of previous research studies, a non-citing article would simply refer to previous research study rather than one of the suggested source articles. This was found not to be the case. Rather than including a formal citation, thirteen of these articles instead provided the URL to the Swiss 2D-PAGE site. Out of these articles, two could not be accessed, and only five mentioned but did not provide a formal acknowledgement of SWISS 2D-PAGE in the text.

In contrast to the small number of articles that mentioned HAMAP or SWISS 2D-PAGE, a total of 1730 publications (in *Science Direct* journals) mentioned PROSITE in the full text. However, only 776 of these were included in the list of references. Due to limitations of the NEXTBIO software, the same analysis could not be conducted for the PROSITE database. For the

suggested source articles introducing PROSITE, 4643 citations were received from publications listed in Scopus and 1000 of these were in Elsevier journals.

Table 5. Underestimation of acknowledge usage by citation analysis for other EXPASY-hosted databases (2000-2011)

	Scopus cites	C	NEXTBIO	M	C+M	U (%)	U_2%
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
Prosite	4634	1000	1730	X	X	X	
Hamap	102	16	7	2	18	11	2
Swiss-2Dpage	575	52	29	20	72	28	3

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.

The analysis for the additional bioinformatics databases and applications hosted by ExPASy is shown in Table 5. For Peptidecutter, Peppepsearch, NextProt and Masssearch appropriate (suggested) source articles could not be identified (in the case of NextProt, the SIB indicated before submission of this paper that a source article had been published in 2011). Some additional databases were excluded as they resulted in too many unrelated returned results due to ambiguity related to the search term similar to "enzyme database" previously discussed. These included, compute pi/MW, sulfonator, myristoylator, blast, biochemical pathways, allall, pROC, PRATT and TCS. The e-resource, Multiident, received 153 Scopus and 28 *Science Direct* citations. It was expected that a considerable number of in-text mentions would also be found using NEXTBIO but none were. When the alternative spelling, "multi-ident" was used, one in-text mention was identified as well as five unrelated articles. This application was therefore excluded from the e-resources considered and listed in Table 5.

Of the e-resourceslisted in Table 5, four (Findpept, FindMod, PeptideMass and Peptidecutter) were introduced in a book chapter rather than in a journal article. The URLs for these e-resources suggest this book chapter as a source reference. However as this book chapter is not indexed by Scopus and could not be considered in this study, despite it receiving over 1400 citations in Google Scholar of which some may be from *Science Direct* journals. This suggests

that a considerable number of the articles with an in-text mention identified by NEXTBIO and not having a corresponding *Science Direct* citation still may have included a citation to this suggested book chapter source. These results are included in Table 5, but they are not considered reliable.

For the e-resources: Findpept, FindMod, and PeptideMass, an alternative value for *M* was developed using a manual search of reference lists. For this alternative M, when a reference to the book chapter was found, this was deducted from the original M. This alternative M value is shown in brackets in Table 5. For these values, the rate of under-representation still remains high. Arguably, this value would be lower if the number of Scopus cites (C) to the book chapter could be accurately assessed.

E-resources such as Swiss-model, RaXML, Swiss-PDBviewer and T-coffee were too popular to be studied using this approach. This was also the case for the PROSITE database. Each of these popular E-resources shown in Table 5 received 7707, 2706, 5910 and 902 Scopus citations respectively, however the in-text mentioned identified by NEXTBIO could not be analysed. For Glycanmass, Glycomod, GPSDB, PLcarber, protscale and protparam, the suggested source reference is the same article for each e-resource. In total, this source reference received 924 citations from Scopus and 204 citations from *Science Direct* journals. However, some of these e-resources yielded too many in-text mentions using NEXTBIO. Therefore, these 6 e-resources or applications were not investigated. The reason why they have nonetheless been listed in Table 5 is to allow for the assessment of the relative share of *Science Direct* citations in total Scopus citation coverage for this field.

In addition, applications such as Pax-DB, OpenStructure, Quickmod, MIAPEgelDB and MALDIPepQuant and HCD/CID spectramerger, did not result in any in-text mentions via NEXTBIO. As a consequence, the estimated rate of under-representation of acknowledged use is zero. One potential explanation for this may be related to the relatively young age of these applications, with some introduced quite recently. This would mean that an insufficient amount of time has passed for citations in references, URLs or in-text citations to accumulate. For this reason, PaxDb was excluded from analysis as its suggested source article was only published in 2012. This would have been too short a period of time to analyse citations.

The rate of underestimation (U) for the remaining e-resources was found to be 5% for Marcoil; 6% for GlycoSuiteDB; and 20% for MSight. The underestimation of the acknowledge use of both MyHits (75%) and Tagident (100%) is relatively high compared to the other e-resources listed in Table 5. For Tagident, all citations were made in *non-Science Direct* journals. Whereas there appears a strong rate of underestimation for this application, it is unlikely to ever be 100%. This is because the suggested source article is referenced in *non-Science Direct* journals. For

this reason, the indicator was adapted to provide a lower bandwidth of the estimated rate of underestimation (U_2). This was calculated using C, the number of Scopus citations. Therefore the U_2 for Tagident was 61 % and 47% for Myhits. This indicates that the estimated rate of underestimation of Tagident (U_{a2}) would be between 61-100% and 47-75% for Myhits. Using this conservatively estimated rate of underestimation, the lower boundaries of the underestimation for HAMAP and Swiss-2Dpage would be 2% and 3 % respectively.

The proportionate share of *Science Direct* citations to the total Scopus citations was found to be 23 %. As was expected, for the indicators Scopus and C, both have a statistically significant Spearman rank correlation of r=0.902 (N=14)[4]. The spearman rank correlation tests indicated that there was a significant correlation between the number of years since the publication of the first suggested source article (time = t) and (1) the number of Scopus (-0.856**), (2) *Science Direct* citations (r=-0.823**), (3) in-text mentions (r=-0.747**) and (4) in-text mentions not identified using citation analyses (M) (r=-0.599*). No significant rank correlation was found between time (t) and the rate of under-representation (U). A significant positive correlation between the number of Scopus references and M was found (0.628*). No significant correlation was found however between the number of *Science Direct* references and M.

Analysis of the non-expasy bioinformatics platform Scratchpads – a product of the ViBRANT project - points to some further limitations of this and alternative strategies. Scratchpads is a platform to support communities in biological systematics and in biodiversity research to share and publish the data and information about species and their habitat. The papers in which Scratchpads was introduced received 62 Scopus citations, but only one of those is in a Science Direct journal. Furthermore for the in-text mentions of Scratchpads, the same problem emerged as in the case of the ENZYME database: Scratchpads and ViBRANT are too general words to identify the platform. The term 'scratchpads' often leads to computer science articles (scratchpad memory) or brain science articles (visio-spatial scratchpad). And the term 'vibrant' is a regularly used adjective and adverb. This, by the way, points at a strategic issue: as proving impact is increasingly crucial for acquiring funding, project leaders may think of names that can be more easily distinguished, and also promote clearly specific publications users of the infrastructure should refer to.

An alternative strategy for in-text mentions of Scratchpads (and ViBRANT) using Google Scholar resulted in a very high rate of false positives: only 5-10 percent of the first 200 results was correct. Google scholar therefore did not appear to offer a suitable alternative to the NEXTBIO software used in this paper. However, the share of false positives may be reduced very strongly, by combining search terms. Unfortunately, this is not possible in NEXTBIO

[4] For these analyses we did not include Quicmod, Findpept, Findmod, PeptideMass, T-Coffee, Swiss-PdbViewer, RAxML and Prosite

software. In Google Scholar, we tested the combination of search terms e.g., *Scratchpad* AND biodiv** or *Scratchpad* AND taxon**. This resulted in a much cleaner set. Nearly all the references to Scratchpad identified through Google Scholar are made in open access journals and conference proceedings that are not included in Science Direct. In contrast to the approach developed in this paper, the Google Scholar approach to 'in-text search' depends on an unknown universe of journals and grey literature. This makes it more difficult to evaluate conclusions about the rate of under-representation. A general lesson is that further tool development for this kind of evaluation studies would therefore be important.

Discussion and Conclusion

This paper illustrated that by considering the rate of underestimation of different knowledge claims, a more complete justification for both citation normalisation and/or the use of alternative metrics in assessing the impact of knowledge claims may be reached. The results presented in this paper show that while citations were related with usage as measured through unique visitors, it is not yet clear how these indicators are related. This may partially be because a considerable share of the acknowledged use of research is not captured by citation analyses alone. Indeed, the rate of underestimation between the e-resources analysed was also found to vary. These observations raise concerns over the accuracy, completeness and suitability of citation analyses as the sole tool for evaluating the impact of e-research infrastructures. This concern also has the potential to extend to considerations of evaluating other types of knowledge claims using citation analysis alone.

Existing research into theories of citations and citing behaviour provides some insights into how these variations may be explained. Citations are known to beget citations: a highly cited publication tends to receive more citations than papers of similar quality because they are more visible or perceived as more "citable" than those cited less, a derivation of the Matthew effect (Merton, 1995). However, if a technology has become ubiquitous, researchers may no longer consider the need to cite this knowledge claim considering it to be "common knowledge." This echoes an argument made by Martin & Irvine (1983). For the purposes of the results presented in this paper, combining these two explanations may explain the relationship between usage (as measured through weblog analysis) and citations observed in this study. Neither explanation, explains the variation in the rates of underestimation of acknowledged use through citation analysis between e-resources. For younger e-resources, however, the rate of underestimation does tend towards zero.

In the case of SWISS 2D-PAGE, a more in-depth exploration of acknowledged use not reflected in citations revealed that many authors had instead referred to a URL to the database either in the reference list or within the text. This type of acknowledgement is more difficult to analyse

than formal citations, but it still represents an alternative method of acknowledging the use of these research infrastructures.

In this paper, 3 alternative approaches to assessing the use of e-research infrastructures are highlighted: 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. Neither approach when used in isolation provides a complete reflection of the actual scholarly usage of e-research infrastructures as not all usage is acknowledged using the reference list or in-text mentions. In addition, as a description of the HAMAP database (Lima et al, 2009) illustrates, researchers may be using technologies without being fully aware of them. There is a difference between: 1) first order users, who make direct use of the HAMAP rule book; and 2) second order users who, while not directly using the rule book or HAMAP database, do use information about HAMAP annotated proteins through other protein databases. When considering usage, this paper only referred to these first order users, therefore it is important to also understand that the actual impact of these technologies may be indirect.

This paper presents one of the first (exploratory) comparative analyses of in text mentions and citation analysis. However using the section search tool within NEXTBIO to analyse the in-text mentions does have limitations. In particular, there are limitations related to name-ambiguity and e-resources (applications/databases) that are popular. The later limitation could be addressed by using alternative approaches to the analysis of in-text mentions or improvements in NEXTBIO. The first limitation, however, is more difficult. In the case of Tagident, the rate of underestimation appears to be 100 % but this is not accurate as the citations appeared in non-SD journals. This suggests a weakness of the proposed approach when dealing with applications which had received only a small number of citations in the time period under consideration. As the example of Scratchpad shows, for some applications the scientific user community can be found to publish almost entirely outside Science Direct journals. Analysis of the in-text mentions that is not restricted to *Science Direct* journals does not have this limitation.

Previous research has argued that comparing citations to reviews, with citations to theoretical or to empirical papers is unfair. Some argue that the same inequality of comparison extends to citations to publications introducing new methods, research instruments or research infrastructures, as the researchers do not cite methods in the same way that they cite other types of knowledge. Citation normalisation is often used to account for differences in the average frequency of citation to different document types such as reviews, letters, editorials and articles (Moed et al, 1995, Rehn & Kronman, 2008). Unfortunately the existing structure of the bibliometric databases does not identify methodological papers, or papers introducing research infrastructures as a different type of document. Therefore, the process of normalisation cannot

be applied to these types of documents in the same way as it is applied to reviews or letters. Furthermore, it is still theoretically unclear why a different value for citations received by different document types is justified. The different levels of the rate of underestimation for acknowledged use of knowledge claims using citation measurement could form part of this justification. This is especially the case if the rate of under-acknowledgement systematically differs between types of knowledge claims. The results presented in this paper show that using citation analysis alone to evaluate e-research bioinformatics resources underestimates the true impact of these knowledge claims to a variable extent. Acknowledging that citations do not reflect the full use of certain knowledge claims must be understood when considering the impact of these e-resources in the biological sciences. Acknowledgement behaviour differs between scientific fields and these differences can be analyzed by building on the approach described in this paper.[5]

Acknowledgements

A shorter version of this paper was presented at the ISSI 2013 conference in Vienna (Jonkers et al, 2013) and at the IWBBIO 2013 conference in Granada. The Spanish Ministry of Economics and Competitiveness funded the project of which this paper forms part through the grant: CSO2011-23508. The first three researchers also received funding from the Ramón y Cajal programme (MINECO) the JAE-DOC programme (CSIC) and the Juan de la Cierva programme (MINECO) of the Spanish Ministry of Economics and Competiveness and the Spanish Research Council (CSIC). The last author acknowledges the EC funded ViBRANT project (grant RI-261532). 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 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 with respect to those contributions. The first author worked on this article at the CSIC institute for Public Goods and Policies. The information and views set out in this chapter do not necessarily reflect the opinion of the first author's current employer. This employer does not guarantee the accuracy of the data included in this study. Neither his current employer nor any person acting on its behalf may be held responsible for the use which may be made of the information contained herein.

References

Baldi, S. (1998). Normative vs. Social Constructivist processes in the allocation of citations: a network-analytic model. *American Sociological Review*, 63(6), 829-46.

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 Research,* 28, 304-305.

[5] Bibliometric researchers, for example, often do not acknowledge the Web of Science or Scopus by including a URL in their publications, let alone a formal citation to the articles in which these databases were first introduced. Of the 518 SD publications that were found through NEXTBIO to mention the use of the Scopus databases in their full text, only 12 included the URL (though in some articles the URL may have been in the reference list).

Bornmann, L. & Daniel, H.D. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation*, 64(1), 45-80.
Cameron, W.B. (1963). *Informal Sociology: A Casual Introduction to Sociological Thinking.* New York: Random House.
Cozzens, S.E. (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.* Oxford:Taylor Graham.
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, 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.
Duin, D., King, D., van den Besselaar, P. (2012). Identifying Audiences of E-Infrastructures - Tools for Measuring Impact. *PLOS ONE*, 7(12), e50943. doi:10.1371/journal.pone.0050943
Editorial. (2012). Alternative metrics, *Nature Materials,* 11,907 doi:10.1038/nmat3485
ExPASy (2012). http://www.expasy.org/proteomics. Accessed June 2012.
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), 113-122.
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,* 1 (pp. 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.
Jonkers, K., Derrick, GE, Lopez-Illescas, C., Van den Besselaar, P. (2013) Are citations a complete measure for the usage of e-research infrastructures. In Juan Gorraiz & Edgar Schiebel et al. (eds), *Proc ISSI 2013,* Vienna, 2013: 136-151
Lima, T., Auchincloss, A.H., 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. *Nucleic Acids Research,* 37 (1): D471-8, doi: 10.1093/nar/gkn661
Lokker, C., Haynes, R.B., Chu, R., McKibbon, K.A., Wilczynski, N.L., Walter, S.D. (2012). How well are journal and clinical article characteristics associated with the journal impact factor? A retrospective cohort study. *Journal of the Medical Library Association,* 100(1), 28-33.
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(4), 304–316.
Lowry, O.H., Rosebrough, N.J., Farr, A.L., Randall, RJ. (1951). Protein Measurement with the Folin Phenol Reagent. *Journal of Biological Chemistry,* 193, 265-275.
Martin, B.R., Irvine, J. (1983). Assessing basic research: Some partial indicators of scientific progress in radio astronomy. *Research Policy,* 12(2), 61–90.
Merton, R.K. (1995). The Thomas Theorem and the Matthew Effect. *Social Forces*, 74(2), 379-424.

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, R.E., Van Leeuwen, T.N. (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. http://www.applications.sciverse.com/action/appDetail/293416. Accessed June/October 2012.

Quest*. (*2010*)*. Funnel Web Analyzer®—overview. *http://www.quest.com/funnel-web-analyzer/index.asp*. Accessed June 2012.

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. Accessed June 2012.

Senker, J. (1995). Tacit knowledge and Models of Innovation. *Industrial and Corporate Change,* 4, 425-447.

Scopus. (2012). http://www.scopus.com. Accessed June/October 2012.

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

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

Thomson Reuters. (2012). http://thomsonreuters.com. Accessed June/October 2012.

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.