Article

JIS

Journal of Information Science
XX (X) pp. 1-13

DOI: 10.1177/016555150000000
jis.sagepub.com
SAGE

Measuring researchers' use of scholarly information through social bookmarking data: a case study of BibSonomy

Ángel Borrego
Facultat de Biblioteconomia i Documentació, Universitat de Barcelona, Spain

Jenny Fry
Department of Information Science, Loughborough University, United Kingdom

Abstract
This paper explores the possibility of using data from social bookmarking services to measure the use of information by academic researchers. Social bookmarking data can be used to augment participative (e.g. interviews and surveys) and other non-participative (e.g. citation analysis and transaction logs) methods to measure the use of scholarly information. We use BibSonomy, a free resource sharing system, as a case study. Results show that published journal articles are by far the most popular type of source bookmarked followed by conference proceedings and books. Commercial journal publisher platforms are the most popular type of information resource bookmarked followed by websites, records in databases and digital repositories. Usage of open access information resources is low in comparison to toll access journals. In the case of open access repositories, there is a marked preference for the use of subject-based repositories over institutional repositories. The results are consistent with those observed in related studies based on surveys and citation analysis, confirming the possible use of bookmarking data in studies of information behaviour in academic settings. The main advantages of using social bookmarking data are that is an unobtrusive approach, it captures the reading habits of researchers who are not necessarily authors and data is readily available. The main limitation is that a significant amount of human resources are required in cleaning and standardising the data.

Keywords
BibSonomy; Information behaviour; Scholarly communication; Scientific information; Social bookmarking

1. Introduction

Studies of researchers' use of scientific and scholarly information are of interest to all stakeholders in the scholarly communication system: authors, publishers, librarians, research managers, and policy makers etc. Researchers' information practices have been studied using a variety of approaches from those that use more naturalistic and interpretive methods, such as diaries, observation and interviews, to more broadly-based participative methods such as questionnaire surveys, through to large-scale non-participative methods, such as bibliometrics and deep-log analysis, which are based on data generated as a by-product of information seeking, dissemination or publication activities.

In the naturalistic tradition studies have resulted in holistic understandings of the information practices of researchers where all aspects of information seeking and use are considered in the context of specific research cultures. For example, in the context of interdisciplinary humanities researchers Palmer and Neumann [1] revealed the translation and boundary crossing techniques that researchers developed in order to locate and make sense of information from diverse disciplinary information landscapes. More recently, Benardou et al. [2] have demonstrated

Corresponding author:
Ángel Borrego, Universitat de Barcelona, Facultat de Biblioteconomia i Documentació, Melcior de Palau, 140, 08014 Barcelona, Spain.
Email: borrego@ub.edu.

the different ways in which researchers in the arts and humanities treat information sources as primary data, highlighting fine-grained details such as the reading and annotating of texts and the different ways in which these annotated texts are shared with other researchers. The series of reader surveys conducted by Tenopir and King [3] in the US since the 1970s depict patterns of use for both formal (journal articles, books, etc.) and informal (meetings, seminars, etc.) sources of information.

Citation analysis, traditionally based on bibliometric data obtained from published journal articles, is an unobtrusive way of capturing authors' reading habits, but the results of such studies need to be interpreted with caution due to the variety of motivations that can underpin the inclusion of a citation in a published work [4] and authors do not necessarily cite every useful information source that they have read. Furthermore, studies based on citation analysis of published outputs exclude those researchers for whom the publication of their research outcomes is not a matter of course, e.g. readers from industry or the public sector. Studies based on usage data, transactions recorded in the servers that host electronic information resources, depict patterns of behaviour on a large-scale in terms of number of visits, time spent, browsing sequences, dates and hours of the visits, etc. [5] These large-scale studies of actual, as opposed to self-reported, behaviour have the potential to augment other approaches to understanding information behaviour, such as those described above. Usage data, however, are not widely available as transaction records are held in the servers of publishers and are commercially sensitive. Additionally, local download statistics compiled by individual libraries are difficult to interpret since COUNTER[1] reports offer no evidence of where the usage originates.

Bibliometric-based methods have an application beyond research into information behaviours. They have been applied in a variety of evaluation contexts to measure, amongst other things, the research output and impact of individuals and Higher Education institutions. In-turn results might be used to inform tenure, promotion and resource allocation decisions. Although citation analysis has been the approach typically applied to the evaluation of research, in the context of the Internet and digital information resources bibliometric-based methods have been extended to include a wider-range of behaviours relating to information seeking, dissemination and publication activities. For example, techniques have been developed to evaluate the use and readership of specific journal titles based on download data. Consequently, new usage indicators that are calculated in a way that is similar to citation indicators have been developed, such as the Journal Usage Factor [6].

Global download usage data for journal titles is not publicly released by publishers, which is the main limitation in calculating Journal Usage Factors. In order to overcome the lack of global download statistics Haustein and Siebenlist [7] have recently proposed to estimate global journal usage by analysing data from social bookmarking services. Social bookmarking services enable users to store, search and share online information resources. During the last few years social bookmarking services that are targeted at the academic research community have emerged, which allow researchers to manage bibliographic metadata of scholarly and scientific literature. In 2010 four social bookmarking services were being used within academic research communities [8]. With the discontinuation of Elsevier's 2collab in April 2011, three services remain: BibSonomy (www.bibsonomy.org), CiteULike (www.citeulike.org) and Connotea (www.connotea.org).

Amongst stakeholders in the scholarly communication system there has been a growing interest in evaluating the use and impact of open access resources and sources, both within and beyond the academic research community [9]. There are two main routes to open access, the "Gold Road" that is based on the model of open access journals, and the "Green Road" that is based on the model of digital repositories of papers published in traditional subscription-based journals, typically after an embargo period. Digital open access repositories can be defined as openly accessible web-based databases of research materials that may or may not be peer-reviewed and may or may not be published in a subscription-based journal. There are two main types of open access repositories: institutional repositories, containing the research outputs of research active employees; and subject-based repositories that hold research outputs relating to particular disciplines or sub-disciplines.

Traditional approaches to citation analysis based on published journal articles are not an effective means by which to measure use and impact in the context of the "Green Road" to open access, since typically authors do not cite unpublished versions of articles, preferring to cite the final published version, albeit there are disciplinary differences in this pattern of behaviour [10]. Whilst the "Gold Road", e.g. open access journals, is well-developed in the Life and Medical Sciences, in other broad-disciplinary areas, such as Physics and Earth Sciences, the "Green Road" has been adopted to a greater extent [11]. Consequently, in order to achieve a holistic picture of the use and impact of open

[1] *"Launched in March 2002, COUNTER (Counting Online Usage of Networked Electronic Resources) is an international initiative serving librarians, publishers and intermediaries by setting standards that facilitate the recording and reporting of online usage statistics in a consistent, credible and compatible way." http://www.projectcounter.org.*

access it is important to develop methodological techniques that are effective in the context of the "Green Road" to open access. The characteristics of social bookmarking services lend themselves to exploration of their potential to contribute to such methodological techniques.

This paper aims to expand on the possibility of using social bookmarking services to measure the use of both published and unpublished information by academic researchers using BibSonomy, a free resource sharing system, as a case study. The research is underpinned by the following questions:

- What types of sources (journal articles, conference proceedings, books, etc.) are being bookmarked by users and do any of these types predominate?
- What are the types of information resources (commercial publishers' platforms, databases, repositories, catalogue records, etc.) being bookmarked by users?
- Are any of these resources available via open access, and if so can a pattern be determined as to predominant open access resources?
- Is social bookmarking data useful for measuring information use within the scholarly communication system? If so;
- How might such data be incorporated into existing approaches for understanding information behaviour in order that stakeholders have access to a complementary set of techniques that collectively provide a holistic picture of information behaviours with regard to informal and formal scholarly communication?

2. Methods

We use BibSonomy as our data source. BibSonomy is a free resource sharing system that allows the sharing of bookmarks and bibliographic references [12]. When a user discovers a website or a publication on the web, they can store the bookmark on BibSonomy's server adding tags in order to retrieve it more easily in the future. The system works in a way that is similar to any Internet browser's bookmark options, but allows users to access data from anywhere and share their bookmarks with other users.

BibSonomy was selected from among the three existing social bookmarking tools serving the academic research community, because it offers a free dataset of the database for research purposes in the form of a SQL dump. According to Reher and Haustein [8] all three existing tools (BibSonomy, CiteULike and Connotea) have a comparable amount of users as measured by traffic statistics. Although BibSonomy is a product targeted at academic researchers, it is important to highlight that it is a freely available tool that anyone can use. BibSonomy distinguishes between two different types of bookmarks: websites and publications. For the purpose of this study we worked with the publications.

We used the dataset corresponding to 1st January 2011 [13]. The dataset included 370.585 records with a URL bookmarked between December 2005 and December 2010. Each bibliographic reference was classified into nine categories according the type of information resource they pointed to:

1. Bookseller
2. Book publisher
3. Catalogue record
4. Database
5. Digital library
6. Digital repository
7. Journal publisher
8. Open access journal
9. Website

When the URL was not self-evident for classification purposes, the website was visited. If the URL was no longer available, pointed to a local disk (C:) or was password protected, the record was removed from the database. Every effort was made to classify all the records in the dataset, even where the URL pointed to an information resource requiring a subscription.

The classified records had been created by 3,168 users. Of the 3,168 users, 17 were responsible for creating more than 1,000 records each and these users were analysed individually. The 281,590 records associated with 14 out of these 17 users had been uploaded within a narrow time period —always less than five days— and referred to a highly limited range of information resources —usually a single catalogue or digital repository. It was assumed that records created by these 14 users had been uploaded by managers of digital libraries in order to enhance use and did not

represent typical behaviour of researchers using BibSonomy, therefore, such records (281,590) were removed from the sample. The analysis reported in this paper is based on the remaining 81,683 records that were created by 3,154 users.

As shown in Figure 1, there is a linear relationship between the number of users and the number of bookmarks for the nine categories of resources considered in this study. This indicates that results are not affected by outlying users that bookmark great amounts of a specific kind of resource.

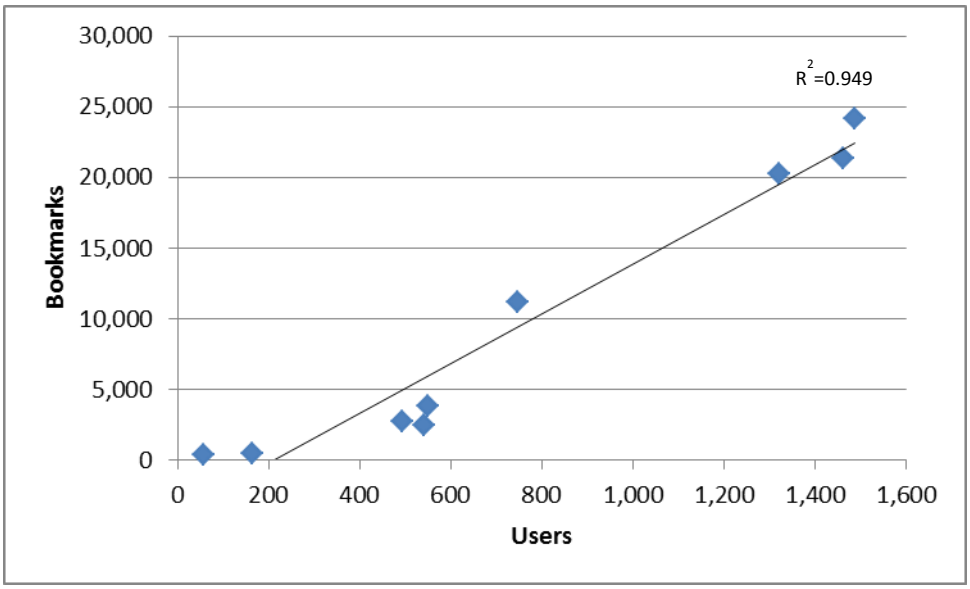

Figure 1. Distribution of resources bookmarked by number of users

3. Results

When bookmarking a bibliographic reference in BibSonomy, users can specify the type of information source the bookmark links to and 61,649 records included this information. As shown in Table 1, the most popular type of source bookmarked, was journal articles, which accounted for nearly half (47%) of the sources bookmarked. A quarter of the bookmarks (25%) linked to conference proceedings, whereas 16% of records referred to books or book chapters.

Table 1. Typology of information sources bookmarked

	Bookmarks	**%**
Journal articles	28,885	46.85
Conference proceedings	15,418	25.01
Books	7,362	11.94
Book chapters	2,447	3.97
Technical reports	1,872	3.04
PhD and Master Thesis	738	1.20
Other	4,927	7.99
Total	61,649	100.00

Table 2 presents the type of information resources bookmarked according to the URL that the bibliographic reference pointed to. Of the bookmarks nearly one third (30%) link to records in commercial journal publishers' platforms, followed by websites (26%), records in databases (25%) and digital repositories (7%). The remaining categories each account for less than 5% of the bibliographic references.

Table 2. Typology of resources bookmarked

	Bookmarks	**%**
Commercial Journal Publishers	24,167	29.59
Websites	21,353	26.14
Databases	20,230	24.77
Repositories	6,082	7.45
Booksellers	3,868	4.74
Catalogue Records	2,710	3.32
Open Access Journals	2,461	3.01
Book Publishers	478	0.59
Digital Libraries	334	0.41
Total	81,683	100,00

Table 3 presents the list of the 12 most popular journal publishers amongst the 209 different commercial journal publishers that were bookmarked. Since the number of bookmarks (B_p) is likely to be influenced by the number of journal titles associated with each publisher (J_b), the number of bookmarks has been weighted to reflect the average number of bookmarks per journal title (B_p/J_b). The number of unique users per publisher (U_p) and the average number of bookmarks per user (B_p/U_b) have also been recorded. As can be observed, there are no significant differences in the average number of bookmarks per journal (B_p/J_b) or the average number of bookmarks per user (B_p/U_b) across the different publisher platforms. There are two exceptions to this, the American Physical Society (83 bookmarks per journal) and, to a lesser extent, Nature Publishing (15 bookmarks per journal) that account for a large amount of bookmarks despite the relatively low number of journal titles in their platforms.

Table 3. Commercial publishers' platforms bookmarked

	B_p	**J_b**	**B_p/ J_b**	**U_p**	**B_p/ U_b**
Elsevier ScienceDirect	4,694	797	5.89	570	8.24
SpringerLink	4,537	785	5.78	757	5.99
IEEE Xplore	1,521	173	8.79	307	4.95
Wiley Online Library	1,474	409	3.60	328	4.49
JSTOR	1,210	377	3.21	193	6.27
American Physical Society	1,073	13	82.54	111	9.67
American Institute of Physics	847	92	9.21	156	5.43
SAGE Publications	794	180	4.41	182	4.36
Oxford Journals	690	78	8.85	112	6.16
Taylor & Francis Online	690	257	2.68	216	3.19
Nature Publishing	657	43	15.28	157	4.18
Ingenta Connect	517	240	2.15	161	3.21
197 Publishers with less than 500 bookmarks	5,463	n/a	n/a	n/a	n/a

B_p: number of bookmarks to the publisher's platform
J_b: number of unique journals bookmarked from the publisher
U_p: number of unique users who bookmark journals from the publisher

After commercial journal publishers, the second most popular type of information resource bookmarked are websites. This category included bookmarks of information sources that were available from websites within the Higher Education, Government and Industry sectors. In the first instance, these are bookmarks of sources that are located at the websites of institutions and organisations where research is performed. For example, the websites of university departments (including the websites of research groups, research projects, and individual academics), government agencies, and the research departments of private companies, that provide copies of staff publications. In the second instance, this category included websites of congresses and conferences that offer online access to their proceedings and there was a wide variation in the types of information source available. Sources ranged from

conference abstracts, posters, and presentation slides to the full-text of talks and papers given. Finally, the website category also included bookmarked references to non-scholarly/scientific sources and resources, such as newspapers and magazines, and bookmarks to other websites, such as Wikipedia articles, blog entries, etc., which strictly speaking would be more appropriate in the 'bookmarks' section of BibSonomy since they are not publications per se.

The third most popular type of information resource to be bookmarked are records in bibliographic databases. The top-four most frequently bookmarked resources in the databases category (Table 4) accounted for 85% of this type of bookmark and it is of note that all four allow for searches in freely available databases. The top-three most popular databases specialise in computer science. Strictly speaking, the most popular resource (ACM Digital Library) could be considered as a publisher platform, since it includes the full text of every article published by the Association for Computer Machinery (ACM), but since it also offers bibliographic records from other publishers it has been considered a database for the purpose of this study. In contrast, DBLP and CiteSeer are produced by academic departments of computer science. The position of Google Scholar in sixth place in the list is due to the creation of bookmarks directly from Google Scholar's results page (URLs such as http://scholar.google.com/scholar?q=keywords). Typically, BibSonomy users bookmark the full-text documents that they access leaving no trace of the information resource that they used to locate the publication. There were, however, users who did bookmark the search results pages of the information resource they used and this was particularly the case for users of Google Scholar. Given the potential idiosyncrasies amongst users in the stage of their information search at which they choose to create a bookmark, the result in Table 4 showing that Google Scholar accounted for 3% of the bookmarks to databases may actually be masking a much greater rate of use and conversely may mean that it is disproportionately represented compared to other popular general search engines.

Table 4. Databases bookmarked

	URL	n	%
ACM Digital Library	portal.acm.org	6,732	33.28
DBLP Computer Science Bibliography	dblp.uni-trier.de	4,377	21.64
CiteSeer	citeseer.ist.psu.edu	3,084	15.24
PubMed	www.ncbi.nlm.nih.gov	2,913	14.40
EBSCOhost	search.ebscohost.com	744	3.68
Google Scholar	scholar.google.com	675	3.34
High-Energy Physics Literature Database	www.slac.stanford.edu/spires/	406	2.01
Citebase	www.citebase.org	240	1.19
sociotech-lit.de	www.sociotech-lit.de	195	0.96
ProQuest	www.proquest.com	164	0.81
ERIC	www.eric.ed.gov	117	0.58
28 bibliographic databases with less than 100 records		583	2.88

In BibSonomy, bookmarks of documents available in open access repositories represent 7% of the bookmarks created by users of the tool. This is not necessarily an accurate representation of the extent of use of open access repositories amongst users of BibSonomy due to differences in the way that bibliographic databases link to the full-text of documents. Most of the records bookmarked in databases include a link to the full-text of the document, whether it be available via a toll fee or open access. Some of the databases, such as CiteSeer, link most of the records to an open access source, whereas other databases, such as CiteBase, are created by harvesting the metadata and full-text links to documents held in open access repositories, such as arXiv. On this basis it can probably be assumed that some users searching for literature in bibliographic databases bookmark the records that include links to open access repositories where they can access the full-text of the document. In these cases, users will probably not bookmark the repository record itself even if this was how they accessed the full-text. In addition to the 7% of total bookmarks that link to documents held in open access repositories, consideration also needs to be given to the large number of links to pre and post prints hosted on institutional websites (usually departmental websites that host the publications of the academics in the Department), research projects' websites, research groups' websites or academics' personal websites, since these websites represent resources that often play an important open access role in the scholarly communication system.

When analysing the most popular bookmarked repositories (Table 5), arXiv accounts for nearly half (48%) of the bookmarks to publications in repositories. In addition, it should be highlighted that the top five most popular

bookmarked repositories, which account for 68% of the bookmarks, are subject-based, as opposed to institutional. The first institutional repository to appear in the list is that of University College London, that comes seventh with 2% of the bookmarks. In fact, just 25% of the documents bookmarked in repositories are hosted in institutional repositories while the remaining 75% are hosted in subject-based repositories. In terms of scatter, the limited number of bookmarked documents held in institutional repositories is distributed amongst 224 repositories, showing a low number of documents bookmarked per repository, whereas bookmarks to documents held in subject-based repositories are concentrated in just 35 repositories.

Table 5. Repositories bookmarked

	URL	n	%
arXiv	www.arxiv.org	2,876	47.81
Pedocs	www.pedocs.de	552	9.18
ACL Anthology: A Digital Archive of Research Papers in Computational Linguistics	aclweb.org/anthology-new	376	6.25
Social Science Research Network	www.ssrn.com	160	2.66
E-LIS. E-prints in Library and Information Science	eprints.rclis.org	153	2.54
PubMed Central	www.ncbi.nlm.nih.gov/pmc/	130	2.14
University College London	eprints.ucl.ac.uk	122	2.01
252 repositories with less than 100 documents bookmarked		1,713	28.17

Another element that is of interest is whether there are any differences in the type of open access repository sources bookmarked compared to those bookmarked from other information resources. If we compare the typology of open access repository sources bookmarked in BibSonomy (Table 6), with those of Table 1, the percentage of articles (34%) and conference proceedings (20%) bookmarked in repositories is lower when compared to the entire population of bookmarks in the dataset (47% and 25% respectively). Correspondingly, a higher percentage of other types of open access sources have been bookmarked compared with the entire population of books, namely: book chapters, technical reports and dissertations. There are a number of open access repositories that specialise in dissertations, such as OhioLINK ETD Center (http://etd.ohiolink.edu), TEL (Thèses en Ligne) (http://tel.archives-ouvertes.fr), Tesis Doctorals en Xarxa (http://www.tdx.cat), and a number of sources in these repositories have been bookmarked.

Table 6. Typology of open access repository sources bookmarked

	Bookmarks	% in OA repositories	% for all bookmarks
Journal articles	1,071	34.22	46.85
Conference proceedings	640	20.45	25.01
Book chapters	382	12.20	3.97
Books	352	11.25	11.94
Technical reports	217	6.93	3.04
PhD and Master Thesis	175	5.59	1.20
Other	293	9.36	7.99
Total	3,130	100.00	100.00

In terms of other types of open access information resources, 3% of the bookmarks in BibSonomy pointed to articles published in open access journals. The most popular publishers of open access journals, according to the number of BibSonomy bookmarks received, were BioMed Central[2] (10% of the bookmarks) and PLoS[3] (Public Library of Science) (9%). The majority of the bookmarks, however, point to individual journal titles published by academic departments, scientific societies and professional bodies, etc. A high-degree of scattering in the bookmarks to open access journals was expected given the relatively recent adoption of open access journals, which have been more

[2] *See BioMed Central at: http://www.biomedcentral.com/*

[3] *See Public Library of Science at: http://www.plos.org/*

popular in some disciplines than others, and the relative immaturity of the reputation of many open access journals, compared to traditional subscription-based journals. This expectation was not realised, however, with 25% of the open access journals accounting for 80% of the bookmarks to this type of resource. This distribution was identical to the distribution of bookmarks to journal titles in the total population of bookmarks, where 26% of journal titles accounted for 80% of the bookmarks.

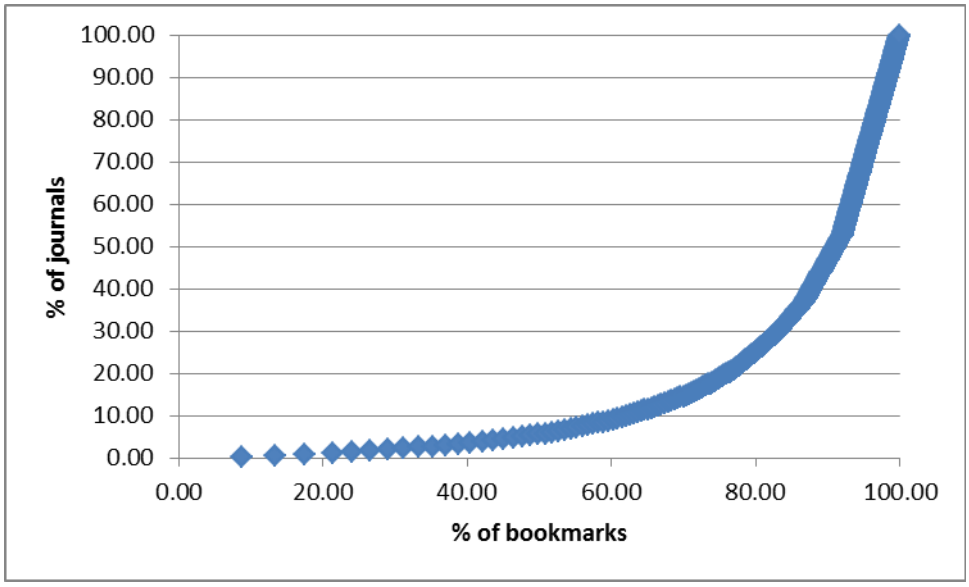

Figure 2. Scattering of bookmarks to open access journals

The BibSonomy dataset corresponding to 1st January 2011 included publications bookmarked from December 2005 to December 2010. When we analyse the evolution in the number of publications bookmarked during this period we observe an exponential increase in the use of the service until 2009 when the total number of bookmarks peaked at 26,740. During 2010 the number of bookmarks started to decline with the total number of bookmarks decreasing to 20,130. When we compare the evolution in the number of bookmarks by type of information source it becomes clear, however, that the number of bookmarks to open access repositories and open access journals has remained stable, as shown in Figure 2. Further research with longer temporal trends will be necessary in order to determine whether this data indicate an increase in the use of open access resources and sources compared to other types of resources and sources.

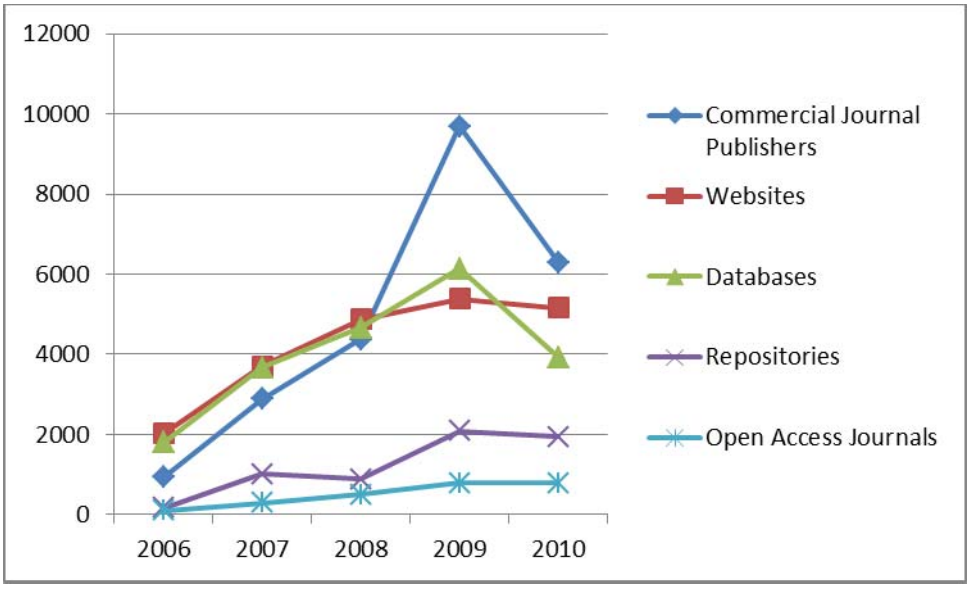

Figure 3. Types of information resources bookmarked by date of creation of the bookmark

"Booksellers" were a target for 5% of the bookmarks in BibSonomy, the majority of which comprised of Amazon (75%) and Google Books (24%). Despite the fact that Google Books is not a bookseller, it has been included in this category since, from the point of view of the information it offers to its users, it offers a service similar to that of "booksellers" e.g. full-text searching of monographs and access to excerpts of their contents.

Finally, 3% of the bookmarks point to catalogue records, mainly from German academic libraries —BibSonomy has been integrated into several of these catalogues—, 0.6% to book publishers and 0.4% to digital libraries —i.e. collections of digitised documents—, with Gallica, http://gallica.bnf.fr, (25% of the bookmarks) being the most popular resource.

The 81,683 bookmarks in BibSonomy had been created by 3,154 users. There is a high degree of concentration of use amongst a few of these users: 15% of the most active users (472 people) are responsible for 80% of the bookmarks. In fact, 1,849 users (59%) have only between one and five bookmarks in their accounts.

4. Discussion

4.1. What types of sources are bookmarked by users?

Published journal articles are by far the most popular type of source bookmarked by users of BibSonomy. This result is consistent with what has been systematically observed in studies of information behaviour in academic settings. Tenopir and King [3], in a monograph synthesizing the results of successive readership surveys in American universities since the 1970s, observe that in every survey they have conducted, scientists were found to read many more scholarly articles than any other type of document. More recently, studies of information behaviour show that scholarly journal articles account for over 90% of scholars' information sources [14, 15].

The fact that conference proceedings appear as the second preferred type of document bookmarked is probably related to the fact that BibSonomy seems to be especially popular amongst academics in computer science, as suggested by the disciplinary scope of the databases and repositories bookmarked. In fact, a comparative study performed by Reher and Haustein [8] concluded that BibSonomy is the most commonly used social bookmarking service in the discipline of computer science. The preference for conference proceedings as an important channel for scholarly communication in computer science has been well documented. A recent bibliometrics study conducted by Wainer et al. [16] found that around 40% of the bibliographic references in ACM papers were to conference proceedings, with 30% to published journal articles and 8% to books.

BibSonomy distinguishes between 'bookmarks' (any website a user wants to bookmark) and 'publications' (articles, books, thesis, etc.). We have analysed the 'publications' only, but we have come across many websites bookmarked in this category that are not publications as understood in the broader context of scholarly communication (e.g. bookmarks to blog entries and Wikipedia articles, etc.). This seems to mean two things; firstly that users do not understand the distinction imposed by BibSonomy or are not concerned about following such a distinction; and secondly, that websites which are not publications, such as blogs and Wikipedia etc., are of great interest to them, since they bookmark many of these resources.

This result is consistent with recent observations by Niu et al. [17] who conducted a study of five universities based in the US. They found that websites were amongst scholars' primary information resources and were used on a daily basis to support research activities. As they point out, most survey-based information behaviour studies prior to 2000 did not include web pages as a possible information resource. Traditionally, scholars have used formal (publications) and informal (face-to-face interactions with colleagues, seminars and conferences etc.) channels of communication for dissemination and collaboration purposes. In the academic research context the Internet has been adopted to such an extent that available tools (e.g. email, VoIP, blogs and personal websites) afford informal communication that is on a par with, or in some cases superior to, face-to-face informal communication. Nicholas et al. [18] showed that a greater number of UK-based researchers rated websites and blogs as being important, than those who rated conference or working papers as important, illustrating a shifting pattern in researchers' information behaviours with regard to informal scholarly communication.

4.2. What types of information resources are bookmarked by users?

Commercial journal publisher platforms were the most popular type of information resource bookmarked by users of BibSonomy and accounted for almost one third of the bookmarks to information resources. The popularity of commercial journal publisher platforms amongst users is consistent with journal articles being the most popular type of source to be bookmarked. Although this result is in-line with what might be expected based on the documented predominance of the published journal article as an information source, one of the advantages of using social bookmarking data to measure the use of information by researchers is that it allows comparisons to be made across different publishers and journals that would otherwise not be possible since usage data from commercial publishers are not widely available.

Bibliographic databases were also a popular information resource to be bookmarked, which reflects their popularity for literature searches. The most popular bookmarked databases had two characteristics in common: they allowed searches to be made in freely available databases and they provided the link to the full-text of the information source. It seems feasible that users search bibliographic databases and bookmark the relevant records they retrieve. These records may include a link to the full-text of the source available most commonly either from a publisher's platform or an open access repository, but also including other types of information resource. The position of Google Scholar as the sixth most popular bookmarked bibliographic database suggests extensive use of this Internet search engine, but since we have only counted the bookmarks created directly from its results page, this may be a disproportionate representation since the usual bookmarking behaviour would be to link to the source once it has been retrieved. Previous related studies based on either referring webpages [19] or surveys [20] have shown that Internet search engines, especially Google and Google Scholar, are now amongst the most important sources of information for researchers.

4.3. To what extent are open access information resources and sources bookmarked by users?

Bookmarks can be a good way to measure the use of open access resources and sources since even though some open access repositories and open access journals offer 'usage' statistics, comparative data across different resources is not readily available, not least because of the different approaches used to measure 'usage'. As a result, most of the research regarding the use of repositories seeks to understand adoption and usage patterns by contributors to repositories, e.g. those who place documents into repositories, but there is little research on the perceptions and experiences of readers (sometimes referred to as end-users in the related literature). According to Jean et al. [21], despite the widespread recognition of the central importance of readers to the ultimate success of open access repositories, we know very little about them. One reason is that repository managers have been more concerned about overcoming barriers to depositing content, than the reader experience, and another is the practical difficulty of identifying and surveying repository users.

According to our results, the usage of open access information resources is low in comparison to toll access journals. Open access repositories represent 7% of the bookmarks in BibSonomy and open access journals 3%. In the case of open access repositories, there is a marked preference for the use of subject-based repositories over institutional repositories. This finding is likely to be related to the volume and type of sources these two different types of repositories hold, but does appear to corroborate current understanding with regard to the use of institutional repositories. For example, according to Asunka et al. [22] there is growing certainty that most institutional repositories remain largely empty, ineffective, or underutilized. Previous research on the use of repositories has also found low levels of usage by readers. For instance, Organ [23] analysed the download statistics at the institutional repository of the University of Wollongong (Australia). During the six month period of the study, 6.2% of the sources received greater than 50 full-text downloads, with the vast majority (79.5%) within the 1 - 50 full-text download range and 14.3% not being downloaded at all. It is worth noting that 95.8% of the downloads were referred from Google.

Caution is required, however, when using the results of bookmarking services analyses as a proxy for the use of open access repositories. The actual use of open access repositories amongst users of BibSonomy is likely to be higher than the 7% of total bookmarks suggests. It should be taken into account that most of the records bookmarked in databases include a link to the full-text of the document, either toll or open access. Some of the databases link most of the publications to an open access source, whereas other databases are created by harvesting the metadata and links to the full-text from open access repositories. It can probably be assumed, therefore, that some users searching for literature in bibliographic databases bookmark the records that include links to repositories where they can get access to the full-text of the document. In these cases, BibSonomy users are unlikely to bookmark the repository record, even if this represents the full-text version that they accessed. In addition to the links to published sources in open access repositories, the large number of links to pre and post prints hosted on institutional websites (usually departmental websites that host the publications of the academics in the Department), research projects' websites, research groups'

websites or academics' personal websites should be considered when measuring the use and impact of open access resources and sources as they represent a large proportion of scholarly outputs available via open access.

When we compare the typology of sources bookmarked in open access repositories with those bookmarked from other information resources, we find that the proportion of journal articles and conference proceedings bookmarked in repositories is underrepresented compared to that of the entire population of bookmarks. This is balanced out by the higher proportion of book chapters, technical reports and dissertations bookmarked from open access repositories. This finding suggests that the impact of open access repositories may be greatest in terms of grey literature and this warrants further investigation.

In addition to journal articles held in open access repositories, 3% of the bookmarks to open access information resources pointed to articles published in open access journals. Since the successful adoption of open access journals is relatively recent and there are a large number of open access journals that do not yet have well-established reputations, it was expected that there would be a high-degree of scattering in the bookmarks to open access journals. This expectation, however, was not realised, with just 25% of open access journal titles accounting for 80% of the bookmarks.

4.4. Is social bookmarking data useful for measuring information use in the scholarly communication system?

Social bookmarking data can be used to augment participative (e.g. interviews and surveys) and other non-participative (e.g. citation analysis and transaction logs) methods to measure the use of scholarly information. Our results concerning the type of information resources and sources bookmarked by users are consistent with those observed in other studies based on surveys and citation analysis, confirming the possible use of bookmarking data in studies of information behaviour in academic settings. It should be highlighted, however, that although BibSonomy is a product targeted at academic researchers, it is a freely available tool that anyone can use.

Being a non-participative approach the main advantage of using bookmarking data, compared to survey data, is that it is an unobtrusive strategy and not vulnerable, therefore, to low response rates. Compared to citation analysis, the main advantage is that it captures the reading habits of all kind of readers, not only those who are authors. In relation to usage data, the main advantage is that bookmarking data is readily available, whereas transaction logs are recorded in private servers. In addition, whereas logs indiscriminately record all types of uses (including, for instance, the download of articles that are quickly examined and dismissed) the action of creating a bookmark implies a certain degree of interest from a reader who wants to keeps the URL for future reference.

When considering the analysis of social bookmarking services to develop understanding of the use of scholarly information some limitations need to be recognised. A great deal of the metadata provided in the bookmark records proved to be incomplete or erroneous, with many discrepancies in spelling. A significant amount of human resources were required in order to manually code the records and to standardise data such as journal titles. Despite the fact that BibSonomy offers scrapers, which allow for the automatic extraction of metadata from selected resources, the results of Reher and Haustein [8] suggest that these do not always work well and often fail to correctly import metadata. The fact that social bookmarking services offer automatic metadata scrapers probably makes users more inclined to bookmark resources that make use of this facility. In turn, this could influence the results of our study and as a point of reference, as of July 2011 BibSonomy offered scrapers for 79 information resources, which included the major scholarly publishers and databases. Furthermore, the existence of links from some platforms to BibSonomy might affect the representativeness of bookmarking data, since users might be more inclined to bookmark publications when a link is readily available. As of December 2011, two of the most frequently bookmarked commercial publishers' platforms (American Institute of Physics and Ingenta Connect), one of the most frequently bookmarked databases (CiteSeer) and one of the most frequently bookmarked repositories (arXiv) provided a link to BibSonomy.

Another issue that should be taken into account is that records created in social bookmarking services can be automatically uploaded by managers of digital libraries in order to enhance their use. These 'users' do not represent the typical behaviour of researchers using the tool, and needed to be removed from the sample. For instance, in our study we removed 281,590 records from the original dataset that were uploaded by just 14 users. All of these records had been uploaded within a very narrow time frame and usually referred to a single catalogue or repository. This reinforces the point that in order to use social bookmarking data to measure the use of scholarly information attention needs to be paid to cleaning and standardising the data.

5. Conclusions

Social bookmaking data can be used to augment other methods to measure the use of scholarly information, such as surveys, citation analysis or server transaction logs. It is a non-participative approach that offers readily available data that captures the reading habits of all kind of readers, not only those readers who are authors. A large amount of work is required, however, in order to standardise incomplete or erroneous metadata. Additionally, it needs to be borne in mind that some social bookmarking services are especially popular amongst academics in specific disciplines or countries, which may have an influence on the results. Since the dataset provided by Bibsonomy is anonymised it was not possible to determine the disciplinary identity of users (e.g. through their departmental affiliation or through surveying them), if this were possible it could be very enlightening to reveal patterns of behaviour along disciplinary dimensions.

The results of this research are consistent with what has been observed in systematic studies of information behaviours in academic settings, which confirms the possible use of social bookmarking data in studies concerned with researchers' information behaviours. It should be borne in mind, however, that although BibSonomy is a product targeted at academic researchers, it is a freely available tool that anyone can use.

Social bookmarks can be a good way to measure the use of open access resources from the point of view of the perceptions and experiences of readers. According to our results, the usage of open access resources is low in comparison to toll access journals. There is a clear preference for the use of subject-based repositories and it is likely that the use of open access repositories amongst users of BibSonomy is probably higher than our results suggest. It should be taken into account that most of the records bookmarked in databases include a link to the full-text of the document available in open access, whether that happens to be an open access repository or a researcher's website etc.

The results reported in this paper are based on a single case study. Further research should be carried out using data from other social bookmarking services.

Acknowledgement

This research was conducted whilst Ángel Borrego was an academic visitor in the Department of Information Science at Loughborough University. The visit was supported by the Catalan Agency for Management of University and Research Grants (2010 BE1 01025) and the Spanish Ministry of Science and Innovation (CSO2008-04762/SOCI).

References

[1] Palmer CL and Neumann LJ. The information work of interdisciplinary humanities scholars: exploration and translation. *Library Quarterly* 2002; 72: 85-117.

[2] Benardou A, Constantopoulos P, Dallas C and Gavrilis, D. Understanding the information requirements of arts and humanities scholarship. *The International Journal of Digital Curation* 2010; 5: 18-33.

[3] Tenopir C and King DW. *Towards electronic journals: realities for scientists, librarians and publishers*. Washington: Special Libraries Association, 2000.

[4] Wouters P. The signs of science. *Scientometrics* 1998; 41: 225-241.

[5] Huntington P, Nicholas D, Jamali HR and Watkinson A. Obtaining subject data from log files using deep log analysis: Case study OhioLINK. *Journal of Information Science* 2006; 32: 299-308.

[6] CIBER. *The journal usage factor: exploratory data analysis. Final report*, http://www.projectcounter.org/documents/CIBER_final_report_July.pdf (2011, accessed September 2011).

[7] Haustein S and Siebenlist T. Applying social bookmarking data to evaluate journal usage. *Journal of Informetrics* 2011; 5: 446-457.

[8] Reher S and Haustein S. Social bookmarking in STM: putting services to the acid test. *ONLINE: Exploring Technology & Resources for Information Professionals* 2010; 34: 34–42.

[9] Houghton J, Rasmussen B, Sheehan P, Oppenheim C, Morris A, Creaser C et al. *Economic implications of alternative scholarly publishing models: Exploring the costs and benefits. Final report*, http://ie-repository.jisc.ac.uk/278/ (2009, accessed September 2011).

[10] Fry J, Probets S, Creaser C, Greenwood H, Spezi V, and White S. *PEER Behavioural Research: Authors and Users vis-à-vis Journals and Repositories. Final report*. http://www.peerproject.eu/fileadmin/media/reports/PEER_D4_final_report_29SEPT11.pdf (2011, accessed December 2011).

[11] Björk BC, Welling P, Laakso M, Majlender P, Hedlund T and Gudnason G. 'Open access to the scientific journal literature: situation 2009'. *PLoS ONE* 5(6), http://dx.doi.org/10.1371/journal.pone.0011273 (2010, accessed September 2011).

[12] Benz D, Hotho A, Jäschke R, Krause B, Mitzlaff F, Schmitz C et al. 'The social bookmark and publication management system BibSonomy' *VLDB Journal* 19 (6), http://www.kde.cs.uni-kassel.de/pub/pdf/benz2010social.pdf (2010, accessed September 2011).

[13] Knowledge and Data Engineering Group, University of Kassel. Benchmark Folksonomy Data from BibSonomy, version of January 1st, 2011.

[14] King D, Tenopir C, Choemprayong S and Wu L. Scholarly journal information-seeking and reading patterns of faculty at five US universities. *Learned Publishing* 2009; 22: 126-144.

[15] Tenopir C, King D, Edwards S and Wu L. Electronic journals and changes in scholarly article seeking and reading patterns. *Aslib Proceedings: New Information Perspectives* 2009; 61: 5-32.

[16] Wayner J, Przibisczki de Oliveira H and Anido R. Patterns of bibliographic references in the ACM published papers. *Information Processing & Management* 2011; 47: 135-142.

[17] Niu X, Hemminger BM, Lown C, Adams S, Brown, C, Level A et al. National Study of Information Seeking Behavior of Academic Researchers in the United States. *Journal of the American Society for Information Science and Technology* 2010; 61: 869-890.

[18] Nicholas D, Williams P, Rowlands I and Jamali HR. Researchers' e-journal use and information seeking behaviour. *Journal of Information Science* 2010; 36: 494-516.

[19] Davis PM. Information-seeking behavior of chemists: A transaction log analysis of referral URLs. *Journal of the American Society for Information Science and Technology* 2004; 55: 326−332.

[20] Ollé C and Borrego A. A qualitative study of the impact of electronic journals on scholarly information behavior. *Library & Information Science Research* 2010; 32: 221-228.

[21] Jean BS, Rieh SY, Yakel E and Markey K. Unheard Voices: Institutional Repository End-Users. *College & Research Libraries* 2011; 71: 21-42.

[22] Asunka S, Chae HS and Natriello G. Towards an understanding of the use of an institutional repository with integrated social networking tools: A case study of PocketKnowledge. *Library and Information Science Research* 2011; 33: 80-88.

[23] Organ, M. 'Download Statistics - What Do They Tell Us? The Example of Research Online, the Open Access Institutional Repository at the University of Wollongong, Australia'. *D-Lib Magazine* 12(11), http://www.dlib.org/dlib/november06/organ/11organ.html (2006, accessed September 2011).

