Aslib Journal of Information Management

The role of arXiv, RePEc, SSRN and PMC in formal scholarly communication

Xuemei Li Mike Thelwall Kayvan Kousha

Received 27 March 2015
Revised 14 May 2015
Accepted 8 September 2015

The role of arXiv, RePEc, SSRN and PMC in formal scholarly communication

Xuemei Li
Peter F. Bronfman Library, York University, Toronto, Canada, and
Mike Thelwall and Kayvan Kousha
Statistical Cybermetrics Research Group,
School of Mathematics and Computer Science,
University of Wolverhampton, Wolverhampton, UK

Abstract

Purpose – The four major Subject Repositories (SRs), arXiv, Research Papers in Economics (RePEc), Social Science Research Network (SSRN) and PubMed Central (PMC), are all important within their disciplines but no previous study has systematically compared how often they are cited in academic publications. In response, the purpose of this paper is to report an analysis of citations to SRs from Scopus publications, 2000-2013.

Design/methodology/approach – Scopus searches were used to count the number of documents citing the four SRs in each year. A random sample of 384 documents citing the four SRs was then visited to investigate the nature of the citations.

Findings – Each SR was most cited within its own subject area but attracted substantial citations from other subject areas, suggesting that they are open to interdisciplinary uses. The proportion of documents citing each SR is continuing to increase rapidly, and the SRs all seem to attract substantial numbers of citations from more than one discipline.

Research limitations/implications – Scopus does not cover all publications, and most citations to documents found in the four SRs presumably cite the published version, when one exists, rather than the repository version.

Practical implications – SRs are continuing to grow and do not seem to be threatened by institutional repositories and so research managers should encourage their continued use within their core disciplines, including for research that aims at an audience in other disciplines.

Originality/value – This is the first simultaneous analysis of Scopus citations to the four most popular SRs.

Keywords Open access, Citations, Scholarly communication, ArXiv, RePEc, SSRN, PMC, Scopus, Subject repositories

Paper type Research paper

Introduction

Scholars can publicise their research in many ways, including online CVs (Kousha and Thelwall, 2014a), personal or departmental web sites (Más-Bleda *et al.*, 2014), social web sites (Skeels and Grudin, 2009; Thelwall and Kousha, 2014) and open access (OA) repositories (Björk *et al.*, 2010; Kim, 2010). OA repositories are web sites that host academic publications and grant free public access to them (Suber, 2012). There are two major OA channels: gold OA by publishing in OA journals or by paying for the OA option in non-OA journals (e.g. Springer Open Choice) (Harnad and Brody, 2004; Laakso, 2014), and green OA (Björk *et al.*, 2014; Laakso, 2014) by making preprints, working papers, postprints or accepted manuscripts publically available in another way, such as through subject repositories (SRs), institutional repositories (IRs) and

Aslib Journal of Information Management
Vol. 67 No. 6, 2015
pp. 614-635

2050-3806
DOI 10.1108/AJIM-03-2015-0049

personal homepages (Gargouri *et al.*, 2012). At least one SR allows publishers to charge for access and so not all are fully OA.

SRs seem to be very popular in some disciplines but may be undermined by gold OA publishing in journals and publishers that allow preprints to be deposited in IRs but not in SRs. For example, the Journal of the Association for Information Science and Technology copyright form, which is one of the standard Wiley-Blackwell forms, allows, "The right to self-archive on the Contributor's personal web site or in the Contributor's own web site or in the Contributor's institution's/employer's institutional repository or archive". SRs collect publications from one or more specific disciplines and can sometimes become standard points of access for academic literature (Björk, 2013). The arXiv.org e-print archive (arXiv), Research Papers in Economics (RePEc) and the Social Science Research Network (SSRN) SRs emerged in the 1990s with the rise of the internet, capitalizing on existing preprint dissemination traditions in physics and economics (Björk, 2013). In contrast, PubMed Central (PMC) archives full-text peer reviewed articles to fit the special needs in the biomedical and life sciences domain (Kling *et al.*, 2004; Kling and McKim, 2000). In comparison, IRs normally serve all of the subject areas within an individual academic institution (Brown, 2010a, b). They started to emerge around a decade later than arXiv in parallel with the early 2002 Budapest OA Initiative (Brown, 2010b). For example, the ePrints Soton archive at Southampton and the DSpace initiative at MIT are the two earliest IRs (Cullen and Chawner, 2011). In addition to research articles, IRs may also contain PhD or student theses, technical reports, video clips, images and data sets (Brown, 2010b). In the past decade, the establishment of new SRs has slowed down in comparison to the rapid growth of IRs (Björk, 2013; Pinfield *et al.*, 2014). Out of 2,728 repositories checked by OpenDOAR (2015), 83 per cent were IRs and only 11 per cent were SRs. This may underestimate the relative use of the two types because some SRs are huge and popular within their disciplines. SRs are not all larger than IRs, however, and the 56 studied SRs varied from holding over 100,000 items to less than 100 items (Björk, 2013; Cybermetrics Lab, 2014). Nevertheless, based on weighted webometric indicators (Aguillo *et al.*, 2010) the four highest impact repositories are all SRs: arXiv, SSRN, Europe PMC and RePEc (Cybermetrics Lab, 2014).

Despite publishers mainly allowing green OA archiving (about 80 per cent to IRs and 33 per cent to SRs) (Laakso, 2014), one study estimated that 12 per cent of published journal articles were green OA (Björk *et al.*, 2014) and another more systematic and recent study found that about half of all Scopus articles 2007-2012 were OA in one form or another, although with substantial disciplinary variations (Archambault *et al.*, 2014). The "build it and they will come" philosophy has not worked fully with the scholarly community except where there was an existing preprints culture (e.g. in physics and economics), or a strong mandate from an authoritative funding agency (e.g. NIH) (Björk *et al.*, 2014; Finch, 2012; Gargouri *et al.*, 2012; Poynder, 2012). The reason for partial OA uptake could be that print journals have largely migrated online and academics tend to rely on library electronic collections to access published journal articles (Tenopir *et al.*, 2011) and so may be confused about the need to widen access to articles that they can already see through their (transparent) institutional journal subscriptions (Spezi *et al.*, 2013). This may explain why the high percentage of OA awareness and generally positive altitudes in many surveys has not translated into universal OA uptake (Creaser *et al.*, 2010; Cullen and Chawner, 2011; Spezi *et al.*, 2013; Swan and Brown, 2005). Moreover, authors tend to cite published articles rather than OA versions, and many send them directly to their colleagues, with

posting to their own web sites, SRs and IRs being seen as less important (Cullen and Chawner, 2011; Larivière *et al.*, 2014; Morris, 2009).

Although SRs have been previously investigated for the relationship between OA publishing and citation counts, their level of use and scholars' attitudes towards them, little is known about cross-disciplinary uses of the major SRs, and trends in their level of uptake over time. Even if SRs are well known within a particular discipline, they may be ignored by other disciplines and hence dissemination strategies that rely upon SRs might be harmful for cross-disciplinary fertilization. Information about trends in uptake over time is needed to develop effective author guidelines and for publishers, research funders and institutions to develop research policies that are sensitive to the level of uptake of SRs. These issues can be addressed indirectly by examining formal citations in academic publications that mention SRs as the source of the cited article. Each such citation gives concrete evidence of the use of a SR to help future research. These citations form an unknown proportion of the uses of a SR, however, because articles can be read for other purposes than informing future research, and a citation in any case may not mention a SR as the source of the article. Nevertheless, the citations can be used to give indicators for the level of uptake that can be compared between disciplines and over time, as well as between SRs. Citations have three advantages over download statistics in this context: they are not affected by SR web site design issues, gaming or spam that may influence the number of downloads; they allow SRs to be compared against each other in a relatively impartial way (although different disciplines have differing proportions of their research in Scopus); and they give evidence of the discipline of the user (citing author). Conversely, downloads are more useful for directly estimating the usage of a SR because a paper may be found and read based on different degrees of information needs from a SR but cited in a different form, such as from its publishing journal (Kurtz and Bollen, 2010). This paper investigates simultaneously, for the first time, how the four most popular SRs have been cited in academic publications indexed in Scopus.

ArXiv, PMC, RePEc and SSRN

There had been at least three decades of systematically sharing preprints in particle physics when arXiv launched in 1991 (Kling *et al.*, 2004). ArXiv is dominated by authors from physics, mathematics and computer science, and 64 per cent of all arXiv articles are in Thomson Reuters Web of Science (WoS) (Larivière *et al.*, 2014). About 75 per cent of publishing condensed matter physicists deposit in arXiv (Moed, 2007), as do 81 per cent of mathematicians (Fowler, 2011). Physicists deposit to arXiv voluntarily and routinely search arXiv for new articles (Spezi *et al.*, 2013) or to stay current (Hemminger *et al.*, 2007). Fifteen years ago, 92 per cent of mathematics faculty and 67 per cent of physics-astronomy faculty used preprints to support their research at the University of Oklahoma (Brown, 1999), confirming their popularity within these subjects. More physics faculty in Southampton University archived with arXiv than with the university's IR (Xia, 2008), and 61 per cent of *Astrophysical Journal* papers are posted to arXiv after acceptance (Schwarz and Kennicutt, 2004), both underlining its value. The importance of arXiv is such that astronomers and physicists value peer review less than do researchers in other disciplines (Mulligan *et al.*, 2013), which allows them to cite arXiv articles even if they have not been refereed. ArXiv seems to be central to the fields of physics and mathematics to an extent that other SRs probably do not match.

PMC grew out of the E-biomed project, which was originally modelled on arXiv and hosted preprints and postprints of biomedical research articles. Nevertheless,

although biochemists and microbiologists are keen to share genomic and proteomic databases (Brown, 2003a), preprints are not acceptable as a viable research dissemination mode for chemists (Brown, 2003b) due to ethical concerns about posting non-peer reviewed articles or data in medicinal, pharmaceutical and biologic chemistry areas "where erroneous information can have life threatening implications" (Brown, 2003b). In recognition of this, E-biomed re-launched as PMC in 2000, giving access instead to refereed OA articles posted by sponsoring journals and scholarly societies (Kling *et al.*, 2004). It now also allows individual authors to submit articles accepted for publication but does not host unrefereed work. It subsequently generated PMC International, which is a partnership between the USA, UK and Canada for archiving life sciences literature. Europe PMC grew from UKPMC in 2012 (UKPMC was launched in 2007) while PMC Canada became operational in 2009 (PMC, 2014). Europe PMC and PMC Canada both include significantly more abstract records than full-text documents and are not exact mirror sites of PMC (Nariani and Fernandez, 2012).

Medical scientists tend to have their articles deposited to PMC or IRs but still rely upon traditional sources for published journal articles (Spezi *et al.*, 2013). The US National Institutes of Health (NIH) OA mandate requires NIH funded research articles to be open to public within 12 months of publication, and many publishers deposit the published copies by the end of embargo date. PMC is the largest SR in terms of archived items (Björk, 2013).

RePEc disseminates economics working papers, journal articles and software components. It was founded in 1997 as a follow up project to NetEc and WoPEc, which started in 1993 (Karlsson and Krichel, 1999; Walshe, 2001; Zimmermann, 2013). It claims to have brought commercial journal publishers and the open source community together to provide free access to research (Bátiz-Lazo and Krichel, 2012). Unlike the other three SRs, RePEc does not have funding support and relies upon volunteers. It also joins many decentralized archives together rather than hosting items on its own server (Karlsson and Krichel, 1999). As a result, it tends to link to full-text items archived elsewhere (Lyons and Booth, 2011) through its services such as IDEAS, EconPapers and the MPRA Personal RePEc Archive. By including unrefereed research, RePEc has affected the type of economics research that can be disseminated, including heterodox economics articles that would be discriminated against in major economics research journals (Novarese and Zimmermann, 2008). Nevertheless, although economists often archive free versions of their published articles online, only 27 per cent were found in RePEc in one study (Bergstrom and Lavaty, 2007) and so it does not seem to be universal in economics.

Unlike arXiv and PMC, RePEc generates and promotes its own usage metrics. RePEc (2014) IDEAS ranks top research items, series, authors and institutions based on citations, abstract views and downloads (Zimmermann, 2013). RePEc Journal Impact Factors have also been used as a research-related indicator (Gibson *et al.*, 2014), and are relatively robust for econometrics journals (Chang and McAleer (2013).

Originating from the Financial Economics Network, SSRN was established in 1994 as a cheap way to disseminate working papers globally (Jensen, 2012). Authors can upload their papers to SSRN as green OA but publishers and institutions are allowed to charge fees for downloading their SSRN papers. Hence SSRN is only partially OA and uses a different model to the other SRs. Business faculty tend to archive their working papers in SSRN rather than in RePEc or IRs because authors can remove their uploaded paper at any time (Hahn and Wyatt, 2014; Lyons and Booth, 2011).

Like RePEc, SSRN calculates usage statistics for its publications. SSRN (2014) ranks top papers, authors and institutions in business, economics and law based on citations and downloads. SSRN download counts have been found to correlate significantly with other traditional research indicators (Black and Caron, 2006), and seem to generate interest amongst academics (Cohen, 2008). Some law faculty have even worried that archiving in IRs might reduce their SSRN rankings (Donovan and Watson, 2011). Given the popularity of its download statistics, SSRN attempts to stop gaming (Edelman and Larkin, 2014). Since external links may inflate an article's download counts, SSRN only allows links to abstract pages to ensure that readers have a chance to view abstracts before downloading the full-text (Black and Caron, 2006). In contrast, PMC only has pages for full-text versions of articles whereas arXiv and RePEc allow linking to both abstract pages and full-text documents.

Related OA citation analysis

There is a rich literature on the apparent "citation advantage" of OA articles over non-OA articles (Craig *et al.*, 2007; Swan, 2010; Wagner, 2010). A number of studies suggest that access to OA full-text before publication and authors' "quality bias" when choosing which of their preprints or postprints to post online are the two major factors behind the apparent OA citation advantage (McVeigh, 2004; Miguel *et al.*, 2011; Moed, 2007).

There are few studies of citations to unpublished articles in SRs, presumably because editors prefer authors to cite published versions of articles (Brown, 2001). Frandsen (2009) found no OA advantage for unpublished RePEc economics working papers while Elleby and Ingwersen (2011) found that working papers received significantly fewer citations than did peer reviewed journal articles from the same research unit. Chu and Krichel (2007) compared citations from WoS and Google Scholar with download statistics for the top 200 most downloaded RePEc articles, finding the two indicators to be related. Brown (2003b) investigated the usage and acceptance of the Chemistry Preprint Server (launched by Elsevier and existing from 2000 to 2004 (Brown, 2010a)) and reported no WoS citations for a subset, although 32 per cent of the most viewed and discussed preprints were eventually published in peer reviewed journals. One recent study investigated WoS citations to, or documents in, arXiv and found that arXiv items, either published or unpublished (including those published in non-WoS indexed journals), receive fewer citations than do equivalent WoS indexed articles (Larivière *et al.*, 2014).

Research questions

The goal of this paper is to assess trends in the uptake of the four major SRs and their interdisciplinary usage. The following questions drive the investigation:

RQ1. Has the level of use of arXiv, RePEc, SSRN and PMC increased over time, including in recent years?

RQ2. Have arXiv, RePEc, SSRN and PMC attracted use from other disciplines or are they essentially disciplinary silos?

The evidence used to address the above questions is taken from explicit mentions of the four SRs in academic literature citations. Although these citations are only very partial indicators of SR use, they can be used for comparisons over time and, to some extent, comparisons between SRs. They can also suggest the share of use of SRs from within different disciplines.

Methods

Scopus was chosen to count how many documents cite the four SRs because Scopus covers more publications than does WoS (journals: 21,000 vs 12,000, and conference proceedings: 17,000 vs 14,800 at the time of writing) (Elsevier, 2014; Thomson Reuters, 2014) and the overlap between Scopus and WoS is large (Gavel and Iselid, 2008). The difference in the total number of individual items, such as articles, may not be the same, however. More importantly, Scopus allows more comprehensive searches within the cited reference fields than does the WoS Cited Reference Search (Kousha *et al.*, 2012). The following Scopus field codes were used:

(1) WEBSITE: to restrict the results to articles with a given URL in their cited references;

(2) REFSRCTITLE: to restrict the results to reference source titles;

(3) SUBJAREA: to limit the results to each of the four broad disciplinary areas:

- Social sciences (this encompasses the Scopus categories: business, management and accounting; social sciences; psychology; economics, econometrics and finance; decision sciences): SUBJAREA(soci OR psyc OR busi OR econ OR deci).
- Natural sciences (this includes engineering, formal sciences and some life sciences and encompasses the Scopus categories: agricultural and biological sciences; chemistry; mathematics; physics; materials science; engineering; earth and planetary sciences; multidisciplinary; environmental science; computer science; biochemistry, genetics and molecular biology; veterinary; chemical engineering; energy): SUBJAREA(chem OR math OR phys OR envi OR comp OR engi OR mate OR eart OR agri OR vete OR mult OR ceng OR ener OR bioc).
- Medical sciences (this excludes some life sciences and encompasses the Scopus categories: health professions; dentistry; pharmacology, toxicology and pharmaceutics; nursing; neuroscience; medicine; immunology and microbiology): SUBJAREA(medi OR nurs OR heal OR phar OR immu OR neur OR dent).
- Arts and humanities (this is the Scopus Arts and Humanities category): SUBJAREA(arts).

(4) PUBYEAR: to limit the publication year, for example from 2000 to 2013: (PUBYEAR > 1999) AND (PUBYEAR < 2014).

To illustrate the above, to identify documents published from 2000 to 2013 citing arXiv URLs from the arts and humanities, the following query was used: SUBJAREA (arts) AND WEBSITE(arxiv) AND (PUBYEAR > 1999) AND (PUBYEAR < 2014). EuropePMC and PMC Canada were not included in the PMC search because PMC is the original authoritative SR; although EuropePMC and PMC Canada are in partnership with PMC, they are more biomedical literature databases (more abstracts than full-texts) rather than OA SRs. Moreover, (WEBSITE(ukpmc) OR WEBSITE (europepmc) OR WEBSITE(pubmedcentralcanada)) AND (PUBYEAR > 1999) AND (PUBYEAR < 2014) only returns 68 results and so would have little impact on the findings. WEBSITE("*ncbi.nlm.nih.gov/pmc*") was used for PMC. Searches for documents citing SSRN and RePEc were similar to those for arXiv, except using WEBSITE(ssrn) and WEBSITE(*repec.org*), respectively. RePEc tends to link to

full-text documents on external servers which may include a "repec" string in their URLs. In total, 100 random citing documents were visited for each SR citation query to check whether the matching documents cited the SR in question. Many arXiv citing documents cited arXiv in a very casual way (e.g. arXiv: 1408.6543) with no hyperlink and no category. In addition, the mirror site http://xxx.lanl.gov/ was heavily cited as well. WEBSITE(arxiv) therefore misses many citing documents while REF(arxiv) would include too many irrelevant results (e.g. citing documents with arXiv in document titles or anywhere else beyond the reference list). To try to capture as many relevant results as practical, query (1) was used:

$$(WEBSITE(*arxiv*)\ OR\ WEBSITE(*xxx.lanl.gov*)\ OR\ REFSRCTITLE(arxiv))$$
$$AND\ (PUBYEAR > 1999)\ AND\ (PUBYEAR < 2014) \quad (1)$$

Random checks of 100 out of the 62,164 citing documents returned from the query (1) found one irrelevant citing document: Ivanov, P.P. (1940) Arxiv Xivinskix Xanov XIX V. Issledovanie i Opisanie Dokumentov s Istorièeskim Vvedeniem, p. 16. Leningrad: Izdanie Gosudarstvennoj Publiènoj Biblioteki from the query (*REFSRCTITLE(arxiv)*). To check for the prevalence of this problem, query (2) was run:

$$(REFSRCTITLE(arxiv)\ AND\ NOT\ WEBSITE(*arxiv*))\ AND$$
$$(PUBYEAR > 1999)\ AND\ (PUBYEAR < 2014) \quad (2)$$

This returns 6,389 unique citing documents from *REFSRCTITLE(arxiv)* alone. To check how many citing documents could possibly be missing using the query (1), query (3) was run:

$$(REF(arxiv)\ AND\ NOT\ ((WEBSITE(*arxiv*)$$
$$OR\ WEBSITE(*xxx.lanl.gov*)\ OR\ REFSRCTITLE(arxiv))))$$
$$AND\ (PUBYEAR > 1999)\ AND\ (PUBYEAR < 2014) \quad (3)$$

This returns 1,524 citing documents which are mixed with more error matches. Query (1) was used despite it missing a few results and returning a few incorrect results.

All Scopus searches were conducted in August 2014 (see Appendices 1 and 2). Presumably, the majority of articles from 2013 had been indexed in Scopus by this time. Nevertheless, Scopus only counts citing documents rather than the exact number of citations and so if an article cites a repository more than once then the additional citations are ignored.

To check how often each of the citing documents cited the SRs, each citing document must be visited to find out the exact number of citations. Given the number of citing documents involved for all the four SRs, it is impractical to visit each of them. Although a random sample of 160 is reasonable (Thelwall, 2004, p. 37), in order to limit the sampling error to ±5 per cent, a random sample size of 384 is necessary (Neuendorf, 2002, p. 89). After exporting all the citing documents from Scopus to Excel, the RAND() function generated a random sample of 384 citing documents for each of the four SRs. Duplicates were not checked for and removed because each sample should reflect the full spectrum of matching articles. Each of the citing documents was then visited to count the number of SR citations in order to record how many just cited the SR in question as a whole without pointing to any particular items archived by the SR and to

find out how many wrongly returned citing documents from the Scopus queries to report their effectiveness. In particular, the cited arXiv and RePEc abstract/full-text links were tracked as well as SR-specific information (e.g. how many RePEc software component citations and how many PMC citations pointed to gold OA journal articles). These samples were used only for the citing checks; the main analyses were performed on the whole of Scopus.

Results and discussion

The number of documents within the whole of Scopus citing each SR has grown quickly over time (Figure 1). The differing volumes may be due to different SR usage rates or differing sizes of the supporting scholarly communities. In addition, RePEc tends to link to full-text articles archived elsewhere rather than hosting copies of articles within the repository (Lyons and Booth, 2011), and was probably cited less as a result. PMC citing documents increased exponentially after 2009, perhaps due to NIH OA mandates since 2006.

Documents citing SRs at the broad disciplinary level

Unsurprisingly, arXiv attracted the most citing documents from natural sciences; both RePEc and SSRN attracted the most citing documents from the social sciences; and PMC attracted the most citing documents from the medical sciences (Figures 2-5). Medicine is in last place in the three non-medical SRs and so PMC is by far the dominant SR for medical research. Arts and humanities research is in second place in RePEc and SSRN, presumably due to the overlap between social science and humanities research within individual disciplines (and Scopus subject categories). Natural science research within RePEc and SSRN may stem from mathematics and physics research applied to economic modelling issues, for example, in econophysics and mathematical economics.

Documents citing SRs at the individual subject level

The subjects most citing each SR give more detailed insights (Figures 6-9). Unsurprisingly, arXiv is dominated by mathematics, physics and computer science, RePEC is dominated by economics, and PMC is mainly dominated by medical and health-related subjects. Contrasting RePEc and SSRN, both are dominated by

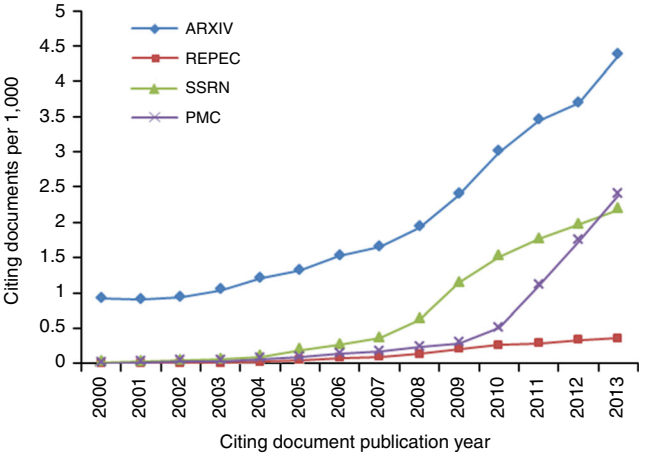

Figure 1.
Citing documents per 1,000 Scopus publications from 2000 to 2013

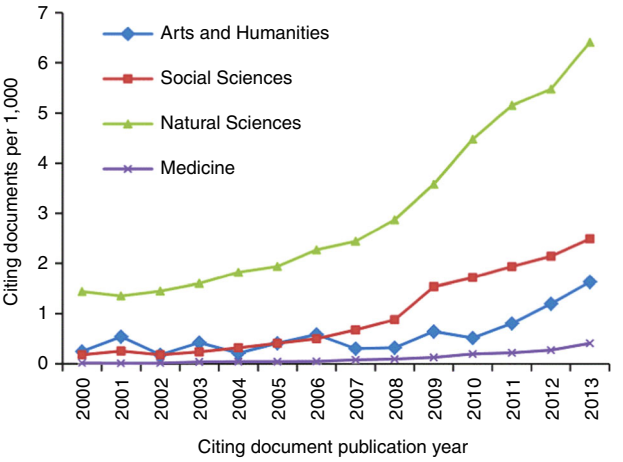

Figure 2.
Documents citing arXiv per 1,000 Scopus documents from the four broad disciplinary areas

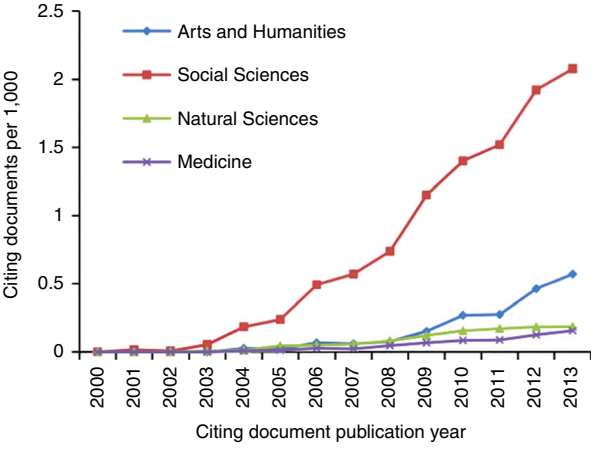

Figure 3.
Documents citing RePEc per 1,000 Scopus documents from the four broad disciplinary areas

economics but it is less dominant in SSRN. The profile of economics in SSRN is perhaps surprising, given the existence of a more specialist SR, although SSRN originated within financial economics. Within PMC, the wide range of subjects represented is perhaps surprising, although the non-medical subject areas have relevance to medicine. For example, biochemistry informs pharmaceutics, agriculture relates to the life sciences, and the environment can impact on health.

Perhaps most surprisingly, arXiv attracts significantly more citations from mathematics than from any other subject area. The dominance of mathematics is not evident in Larivière *et al.*'s (2014) study, which found that similar proportions of 2010-2011 WoS physics (20 per cent) and mathematics (21 per cent) papers were in arXiv (Larivière *et al.*, 2014, Figure 2) and a much higher proportion of references were to arXiv in WoS physics papers than in WoS mathematics papers (1995-2010). In addition, 1.4 per cent of references in WoS physics papers from 2011 and 1 per cent of references in WoS mathematics papers from 2011 cited arXiv preprints (Larivière *et al.*, 2014, Figure 6A). Given that these papers have multiple references

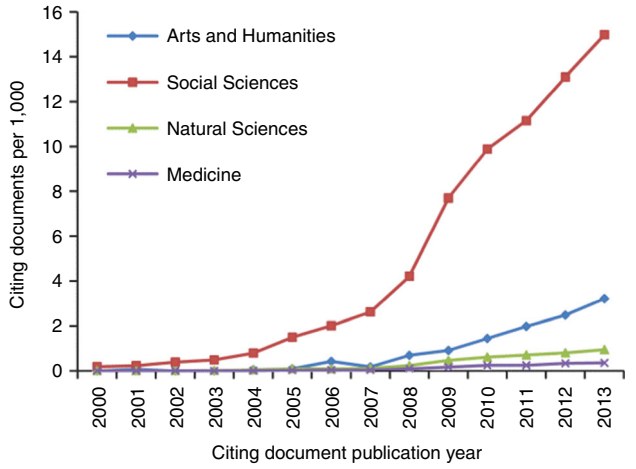

Figure 4.
Documents citing SSRN per 1,000 Scopus documents from the four broad disciplinary areas

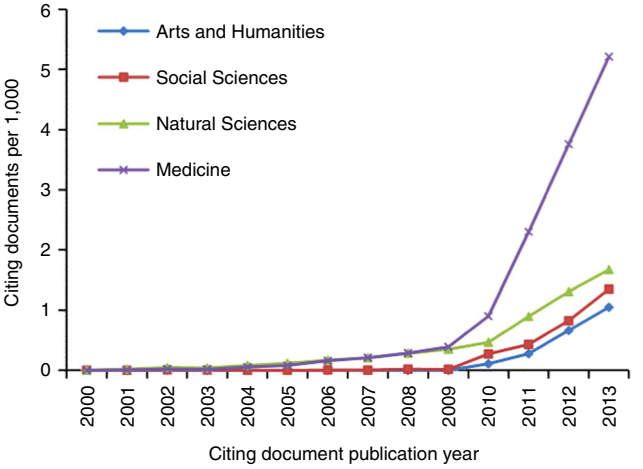

Figure 5.
Documents citing PMC per 1,000 Scopus documents from the four broad disciplinary areas

each, it is likely that this reflects a much higher proportion of papers in WoS citing arXiv. As illustrated in Figure 6, in 2011, the arXiv citing proportion is 2 per cent for mathematics and 1 per cent for physics. Both these numbers are much lower than could be expected from Larivière *et al.*'s (2014) study and also reverse the difference between mathematics and physics. The difference may be due to Larivière *et al.*'s (2014) method identifying ways of mentioning arXiv without using URLs, such as references with arXiv identifiers, that must have been more comprehensive than the combination of WEBSITE and REFSRCTITLE searches used here.

The physics/mathematics difference may also be due to classification and coverage differences between WoS and Scopus. Scopus covers more mathematics documents (1,447,750 at the time of writing by searching Scopus using SUBJAREA(math) AND (PUBYEAR > 1999) AND (PUBYEAR < 2014), and is 47.2 per cent of the number of Scopus physics articles) than does WoS (689,156 at the time of writing by searching WoS using *SU* = (mathematics) AND PY = (2000-2013), and is 35.7 per cent of the

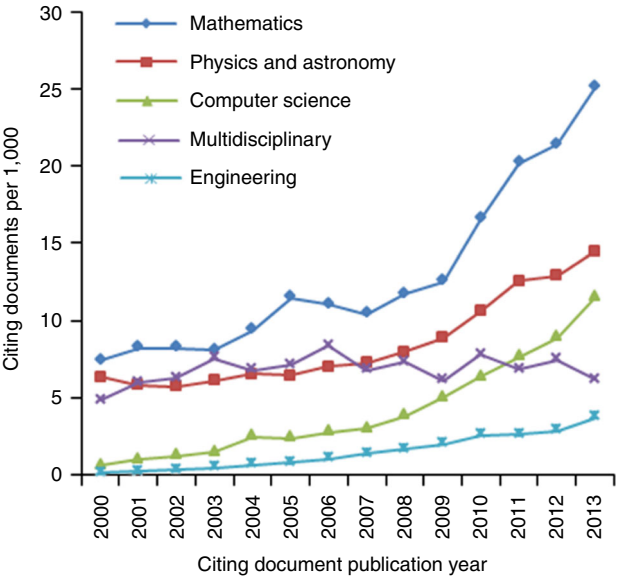

Figure 6.
Top subjects citing arXiv per 1,000 Scopus documents in the subject

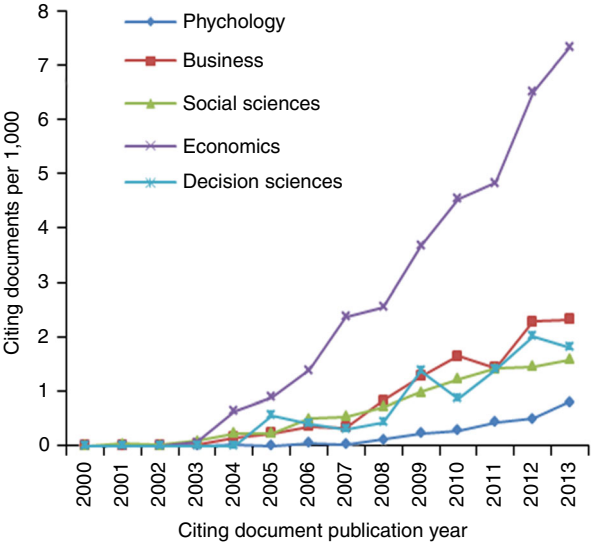

Figure 7.
Top subjects citing RePEc per 1,000 Scopus documents in the subject

number of WoS indexed physics articles) and hence there is a substantial content difference between Scopus and WoS. Scopus may tend to classify documents as mathematics that are not classified as mathematics in WoS and the opposite for physics. Scopus may also index more computer science and classify some of it as mathematics (e.g. Information Processing Letters) as well as dual classifying some computer science as mathematics (e.g. Lecture Notes in Computer Science) and also dual classifying some physics as mathematics (e.g. Physica A: Statistical Mechanics and its Applications). To illustrate this, query (4) returns all arXiv citing documents in

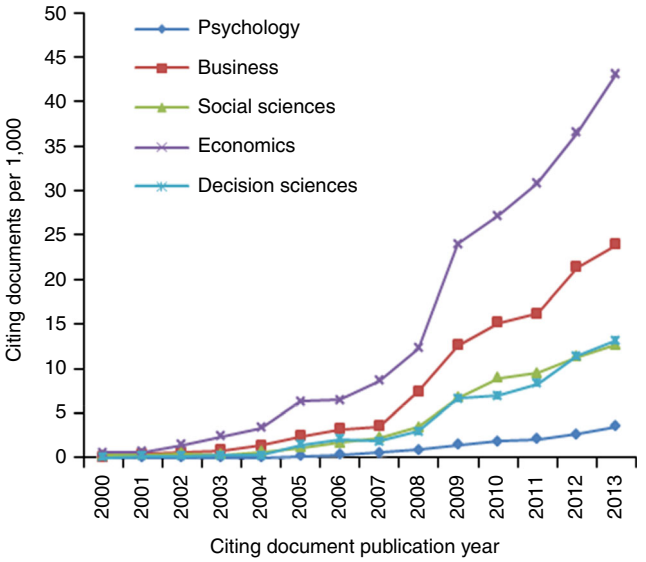

Figure 8.
Top subjects citing SSRN per 1,000 Scopus articles in the subject (journal and conference articles in English)

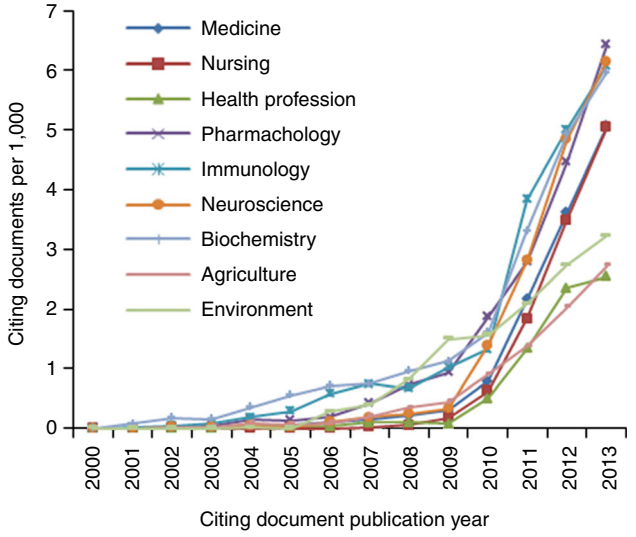

Figure 9.
Top subjects citing PMC per 1,000 Scopus articles in the subject (journal and conference articles in English)

Scopus-indexed mathematics publications. Out of the five journals most citing arXiv (see Table I), only articles from *Advances in Mathematics* are overwhelmingly categorized as mathematics in both WoS and Scopus. Articles from the two most citing journals *Lecture Notes in Computer Science* and *Communications in Mathematical Physics* are both dually classified as mathematics with computer science and physics, respectively. Both *Physical Review D Particles Fields Gravitation* and *Cosmology and IEEE International Symposium on Information Theory Proceedings* are not indexed in WoS, however, articles from the two journals are all partially mathematics although

Journal	Citing arXiv	WoS category and % of articles in journal classified by WoS in the category	Scopus category and % of articles in journal classified by Scopus in the category
Lecture notes in computer science	1,742	Maths: 5.4% Computing: 99.8%	Maths: 90.6% Computing: 99.8%
Communications in mathematical physics	1,205	Maths: 0% Physics: 100%	Maths: 100% Physics: 100%
Physical review D particles fields gravitation and cosmology	732	Not indexed	Math: 44.6% Physics: 100%
IEEE International Symposium on information theory proceedings	606	Not indexed	Maths: 47.8% Computing: 47.8% Engineering: 52.2%
Advances in mathematics	474	Math: 100%	Math: 100% Computing: 6.9%

Table I.
The five Scopus mathematics journals most citing arXiv 2000-2013

those from the former are also classified as physics, while those from the latter also as computer sciences and engineering:

$$(WEBSITE(*arxiv*)\ OR\ WEBSITE(*xxx.lanl.gov*)\ OR\ REFSRCTITLE(arxiv))$$

$$AND\ SUBJAREA(math)\ AND\ (PUBYEAR > 1999)\ AND\ (PUBYEAR < 2014) \quad (4)$$

Larivière *et al.*'s (2014) (probably better) method of relying upon the arXiv category that the article was uploaded to may also affect the results but the two major causes of the difference are probably the greater coverage of Scopus and the large number of citations to arXiv's mirror site http://xxx.lanl.gov/ that were not included in that study. This suggests, but does not prove, that arXiv is more important in formal scholarly communication to mathematics (at least in comparison to physics) than has previously been explicitly acknowledged.

SR citation frequencies per citing document in the random samples

Based upon the four random samples of matching documents, the Scopus queries used to search SR-citing documents were reasonably effective at returning correct citing documents. Only one arXiv citing document pointed to an irrelevant URL:

http://demoscope.ru/weekly/2005/0223/arxiv04.php

And only two SSRN-citing documents pointed to irrelevant URLs:

www.landesbioscience.com/journals/rnabiology/article/SuessRNA5-1.pdf

www.cisco.com/univercd/cc/td/doc/solution/esm/qossrnd.pdf

On average arXiv had the most citations per citing document (2.51) followed by SSRN (1.7), RePEC (1.27) and PMC (1.08) (Tables II and III). One article cited arXiv 37 times out of 52 references, which was more than double the maximum for the other SRs. Over 93 per cent of the sampled citing documents only cited PMC once, in comparison to RePEc (87.8 per cent) and then SSRN (72.9 per cent) while less than 58 per cent of the sampled citing documents cited arXiv only once (Table III).

Out of the 965 arXiv citations from the random sample of 384 articles matching the arXiv Scopus query, 70 per cent were in arXiv physics categories and 17 per cent

were in arXiv mathematics categories. Out of the 384 random documents citing arXiv, 44 per cent were categorized by Scopus at least once as physics while 36 per cent were categorized at least once as mathematics, although the smaller difference for Scopus may be due to the way in which its journals are classified. Many of the arXiv citations are in a short format like arXiv:1011.3370 (arXiv e-print ID) rather than exact URLs. These were classified as pointing to arXiv abstracts, although the authors could assume that the link would also lead to the full-text OA versions. There were 162 (17 per cent) citations with full-text arXiv article links. Eight arXiv citations pointing to arXiv articles without indicating the article ID.

Almost all (97 per cent) of the 487 RePEc citations in the random sample pointed to either IDEAS (393; 81 per cent) or Econpapers (81; 17 per cent). Most IDEAS and Econpapers citations pointed to external full-text download URLs and only 13 pointed to full-text documents, 12 of which were outside IDEAS and Econpapers. Two-thirds (321; 66 per cent) of the RePEc citations pointed to working papers, a substantial minority (73; 16 per cent) cited software components (uniquely amongst the SRs here), and a few (39; 8 per cent) pointed to non-OA full-text documents such as subscription-based journal articles. A fifth (105; 22 per cent) of the RePEc citations pointed to university archives through either IDEAS or Econpapers, and the rest pointed to working paper series from the World Bank, the IMF, the NBER Working Papers, EconWPA and others. Although working papers are clearly central to RePEc, economics researchers may also get notified of new working papers through NEP (the free New Economics Papers e-mail notification services). For example, WEBSITE ("nber.org/papers") OR REFSRCTITLE("NBER working paper") returns 18,981 citing documents (also from year 2000 to 2013) for NBER working papers alone.

Only a few (48; 7 per cent) of the 652 SSRN citations in the random sample point to SSRN articles at SSRN Working Papers or ssrn.com without article IDs or links. In total, 12 of the SSRN citations had disappeared, perhaps due to journal requests to remove them after submission, although faculty may also remove articles (Hahn and Wyatt, 2014).

Table II.
Number of citations per paper for articles in the four random samples of 384 articles matching each respository query

	arXiv	RePEc	SSRN	PMC
Total citations	965	487	652	414
Mean	2.51	1.27	1.70	1.08
Median	1	1	1	1
Maximum	37	7	15	7
Minimum	0	1	0	1

Table III.
Frequencies of 1-4 citations per citing document for articles in the four random samples of 384 articles matching each respository query

Citations	arXiv	RePEc	SSRN	PMC
1	221 (57.6%)	337 (87.8%)	280 (72.9%)	360 (93.8%)
2	69 (18.0%)	25 (6.5%)	51 (13.3%)	22 (3.6%)
3	32 (8.3%)	7 (1.8%)	14 (3.6%)	1 (0.3%)
4	16 (4.2%)	5 (1.3%)	12 (3.1%)	0 (0.0%)

Almost all (393; 95 per cent) of the 414 PMC citations from the random sample pointed to full-text pdf links, although PMC provides different versions of full-text links, including HTML. Most (258; 62 per cent) of the PMC citations pointed to gold OA journal articles (see: http://doaj.org/). the main journals were *PLOS ONE* (75 citations), the *World Journal of Gastroenterology* (69) and *Environmental Health Perspectives* (46). It is not clear why these authors cited the PMC archived articles rather than the OA journal sites.

Overall, arXiv was cited the most frequently in each citing document followed by SSRN, RePEc and PMC based on both the mean and frequency statistics from the four random samples, and all SR citations overwhelmingly pointed to particular articles, either their abstracts or full-texts, rather than citing a SR as a whole (exceptions: two articles cited RePEc and two cited PMC). Whilst arXiv allows links to its articles' abstracts or full-texts; RePEc hosts abstracts and points to full-text to external servers from a wide range of working paper series; SSRN sets abstract page as the default link of an article and readers need to view the abstract page before reaching the full-text download page to ensure robust downloading counts; and PMC points to various versions of full-text articles. Not surprisingly in this context, RePEc and SSRN citations were dominated by abstract pages, a minority (17 per cent) of arXiv citations pointed directly to full-text versions and almost all (95 per cent) PMC citations pointed directly to full-text pdfs. Despite the substantial differences in the type of document linked to, it seems possible that the links serve broadly similar purposes for most authors, who may read the title and abstract first and then decide whether to read the full text of a paper.

Limitations

Scopus does not cover all research publications and it is possible that some important sources of publications are missing, for example perhaps book chapters and Chinese journals. In addition, the Scopus queries seem to return the majority SR citations but do not return all of them. Moreover, as the analysis of mathematics suggests, the results are likely to be due to some extent to the coverage and subject classifications of Scopus, so that comparisons between fields may be unfair if Scopus has wider coverage of one. The grouping of subjects into four broad disciplinary areas is an oversimplification to some extent. For example, biochemistry is important to PMC but was categorized within the natural sciences. The citing differences by subject, discipline and repository over the years are all based on citing documents rather than actual citations.

Most importantly, however, it seems likely that most citations to documents found in these repositories would not mention the repositories, especially for published articles, but would use a traditional citation instead. Hence, the figures reported here are likely to be substantial underestimates. In addition, articles seem to be commonly referenced in arXiv with identifiers instead of URLs, further undermining the figures, despite the use of the REFSRCTITLE command to catch some of these. Moreover, since RePEc does not have a single centralized archive, authors may also cite other archives that RePEc redirects them to.

Finally, the way in which the relatively new Scopus WEBSITE command indexes documents may have changed during the period studied, for example to be applied more comprehensively over time. Tests with this command suggested that it has been applied retrospectively to documents that were published long before it was introduced, however. For example, a search for WEBSITE(com) returned small numbers of (false) matches from as far back as 1977, before the web began and before internet domain names were used.

Conclusions

In answer to Question 1, direct citations to arXiv, RePEc, SSRN and PMC in Scopus-indexed scholarly publications have all increased steadily from 2000 to 2013, although at different rates. The low initial number of citations to PMC is not surprising as it was launched later than the others, in 2000. The exponential growth in articles citing PMC after 2008 may have been caused by the NIH OA mandates since 2006. The small number of citations to RePEc may be caused by RePEc often linking to full-text versions of articles on external servers. The increasing number of citations to all of the SRs forms useful evidence that they all continue to be an important part of the scholarly infrastructure, despite publishers' apparent preferences for IRs. Hence, researchers in relevant disciplinary areas should continue to use them and policymakers do not yet need to encourage or plan for a wholesale migration to IRs. These findings are about the trends in uptake of the SRs, as evident from citations to them in published articles and are based upon the assumption that these citations reflect the much higher usage of them by researchers, even though the vast majority of articles found in SRs and cited in published work are presumably not cited via the SR. Perhaps most importantly, the findings assume that researchers cite a uniform proportion of papers discovered in SRs with SR references. This assumption is somewhat problematic because it seems possible that researchers have become increasingly likely to cite SRs to acknowledge their role or to help readers to find the articles.

In answer to Question 2, there are substantial disciplinary differences in citing the four SRs. At the broad disciplinary level, each repository was most cited within its own area. At the subject level, arXiv seems to be cited the most by mathematics, RePEc and SSRN are both cited most by economics, and PMC is cited the most by a group of biomedical subjects. Perhaps most importantly, however, the evidence of substantial use of each SR outside of its disciplinary area is valuable evidence of the utility of SRs for supporting this kind of wider uptake. Researchers seeking interdisciplinary audiences for their research can therefore use SRs for this.

The comparison between the SRs found some substantial differences. For example, 16 per cent of the RePEc citations pointed to software components, showing that it is uniquely successful at hosting information about software, and other SRs might also wish to consider making provisions for hosting non-standard academic outputs. A total of 62 per cent of the PMC citations pointed to gold OA journal articles, confirming that gold OA is particularly important for biomedical and life sciences researchers (Gargouri *et al.*, 2012; Sotudeh and Horri, 2007).

In terms of methods, the new Scopus WEBSITE reference search facility has made it possible to investigate citations to online archives because it was possible to construct queries with few false matches. Nevertheless, it was not possible to identify all relevant citations with this method due to shorthand arXiv citation formats, which was only partially compensated for with the REFSRCTITLE command. Despite this and the differences in strategies of the different repositories in terms of whether to accept unrefereed articles and whether to present league tables based upon download statistics, arXiv, RePEc, SSRN and PMC clearly play an important and growing role in scholarly communication within their fields.

For future work, the WEBSITE reference search facility from Scopus can also be applied to other types of web site, also following up previous studies of investigated web pages (Kousha and Thelwall, 2014b) and YouTube (Kousha *et al.*, 2012).

References

Aguillo, I., Ortega, J., Fernández, M. and Utrilla, A. (2010), "Indicators for a webometric ranking of open access repositories", *Scientometrics*, Vol. 82 No. 3, pp. 477-486.

Archambault, E., Amyot, D., Deschamps, P., Nicol, A., Provencher, F., Rebout, L. and Roberge, G. (2014), "Proportion of open access papers published in peer-reviewed journals at the European and World levels – 1996-2013", Science-Metrix, available at: http://science-metrix.com/files/science-metrix/publications/d_1.8_sm_ec_dg-rtd_proportion_oa_1996-2013_v11p.pdf (accessed 12 January 2015).

Bátiz-Lazo, B. and Krichel, T. (2012), "A brief business history of an on-line distribution system for academic research called NEP, 1998-2010", *Journal of Management History*, Vol. 18 No. 4, pp. 445-468.

Bergstrom, T.C. and Lavaty, R. (2007), "How often do economists self-archive?", Department of Economics, UCSB, CA, available at: http://escholarship.org/uc/item/69f4b8vz (accessed 23 August 2014).

Björk, B.-C. (2013), "Open access subject repositories: an overview", *Journal of the Association for Information Science and Technology*, Vol. 65 No. 4, pp. 698-706. doi: 10.1002/asi.23021.

Björk, B.-C., Laakso, M., Welling, P. and Paetau, P. (2014), "Anatomy of green open access", *Journal of the Association for Information Science and Technology*, Vol. 65 No. 2, pp. 237-250.

Björk, B.-C., Welling, P., Laakso, M., Majlender, P., Hedlund, T. and Guðnason, G. (2010), "Open access to the scientific journal literature: situation 2009", *PLoS ONE*, Vol. 5 No. 6, p. e11273. doi: 10.1371/journal.pone.0011273.

Black, B. and Caron, P. (2006), "Ranking law schools: using SSRN to measure scholarly performance", *Indiana Law Journal*, Vol. 81 No. 1, pp. 83-139.

Brown, C. (1999), "Information seeking behavior of scientists in the electronic information age: astronomers, chemists, mathematicians, and physicists", *Journal of the American Society for Information Science*, Vol. 50 No. 10, pp. 929-943.

Brown, C. (2001), "The E-volution of preprints in the scholarly communication of physicists and astronomers", *Journal of the American Society for Information Science and Technology*, Vol. 52 No. 3, pp. 187-200.

Brown, C. (2003a), "The changing face of scientific discourse: analysis of genomic and proteomic database usage and acceptance", *Journal of the American Society for Information Science and Technology*, Vol. 54 No. 10, pp. 926-938.

Brown, C. (2003b), "The role of electronic preprints in chemical communication: analysis of citation, usage, and acceptance in the journal literature", *Journal of the American Society for Information Science and Technology*, Vol. 54 No. 5, pp. 362-371.

Brown, C. (2010a), "Communication in the sciences", *Annual Review of Information Science and Technology*, Vol. 44 No. 1, pp. 285-316.

Brown, D.J. (2010b), "Repositories and journals: are they in conflict? A literature review of relevant literature", *Aslib Proceedings*, Vol. 62, pp. 112-143.

Chang, C.-L. and McAleer, M. (2013), "Ranking leading econometrics journals using citations data from ISI and RePEc", *Econometrics*, Vol. 1 No. 3, pp. 217-235.

Chu, H. and Krichel, T. (2007), "Downloads vs citations: relationships, contributing factors and beyond", available at: http://eprints.rclis.org/handle/10760/11085 (accessed 21 August 2014).

Cohen, N. (2008), "Now Professors Get Their Star Rankings, Too", *The New York Times*, 9 June, p. 4, available at: www.nytimes.com/2008/06/09/business/media/09link.html (accessed 28 August 2014).

Craig, I.D., Plume, A.M., McVeigh, M.E., Pringle, J. and Amin, M. (2007), "Do open access articles have greater citation impact? A critical review of the literature", *Journal of Informetrics*, Vol. 1 No. 3, pp. 239-248.

Creaser, C., Fry, J., Greenwood, H., Oppenheim, C., Probets, S., Spezi, V. and White, S. (2010), "Authors' awareness and attitudes toward open access repositories", *New Review of Academic Librarianship*, Vol. 16 No. S1, pp. 145-161.

Cullen, R. and Chawner, B. (2011), "Institutional repositories, open access, and scholarly communication: a study of conflicting paradigms", *The Journal of Academic Librarianship*, Vol. 37 No. 6, pp. 460-470.

Cybermetrics Lab (2014), "WORLD|Ranking Web of Repositories", available at: http://repositories.webometrics.info/en/world (accessed 24 August 2014).

Donovan, J.M. and Watson, C.A. (2011), "Will an institutional repository hurt my SSRN ranking: calming the faculty fear", *AALL Spectrum*, Vol. 16 No. 6, pp. 12-13.

Edelman, B.G. and Larkin, I. (2014), "Social comparisons and deception across workplace hierarchies: field and experimental evidence", Organization Science, available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1346397 (accessed 28 August 2014).

Elleby, A. and Ingwersen, P. (2011), "Do open access working papers attract more citations compared to printed journal articles from the same research unit?", *Proceeding of the ISSI 2011 Conference, Presented at the 13th International Conference of the International Society for Scientometrics & Informetrics*, *Durban*, *July 4-7*, pp. 327-332.

Elsevier (2014), "Content overview: Scopus", available at: www.elsevier.com/online-tools/scopus/content-overview (accessed 21 February 2014).

Finch, D.J. (2012), "Accessibility, sustainability, excellence: how to expand access to research publications", available at: www.researchinfonet.org/publish/finch/ (accessed 2 January 2015).

Fowler, K.K. (2011), "Mathematicians' views on current publishing issues: a survey of researchers", available at: http://conservancy.umn.edu/handle/11299/109309 (accessed 9 March 2014).

Frandsen, T.F. (2009), "The effects of open access on un-published documents: a case study of economics working papers", *Journal of Informetrics*, Vol. 3 No. 2, pp. 124-133.

Gargouri, Y., Larivière, V., Gingras, Y., Carr, L. and Harnad, S. (2012), "Green and gold open access percentages and growth, by discipline", *Proceedings of STI 2012, Presented at the 17th International Conference on Science and Technology Indicators*, *Montreal*, *5-8 September*, available at: http://arxiv.org/abs/1206.3664 (accessed 10 August 2014).

Gavel, Y. and Iselid, L. (2008), "Web of science and scopus: a journal title overlap study", *Online Information Review*, Vol. 32 No. 1, pp. 8-21.

Gibson, J., Anderson, D.L. and Tressler, J. (2014), "Which journal rankings best explain academic salaries? Evidence from the University of California", *Economic Inquiry*, Vol. 52 No. 4, pp. 1322-1340.

Hahn, S.E. and Wyatt, A. (2014), "Business faculty's attitudes: open access, disciplinary repositories, and institutional repositories", *Journal of Business & Finance Librarianship*, Vol. 19 No. 2, pp. 93-113.

Harnad, S. and Brody, T. (2004), "Comparing the impact of open access (OA) vs. non-OA articles in the same journals", *D-lib Magazine*, Vol. 10 No. 6, available at: http://eprints.ecs.soton.ac.uk/10207 (accessed 21 August 2014).

Hemminger, B.M., Lu, D., Vaughan, K.T.L. and Adams, S.J. (2007), "Information seeking behavior of academic scientists", *Journal of the American Society for Information Science and Technology*, Vol. 58 No. 14, pp. 2205-2225.

Jensen, M.C. (2012), "ABOUT SSRN: from the desk of Michael C. Jensen, Chairman", available at: www.ssrn.com/update/general/mjensen.html (accessed 28 August 2014).

Karlsson, S. and Krichel, T. (1999), "RePEc and S-WoPEc: internet access to electronic preprints in economics. presented at the third ICCC", *IFIP Conference on Electronic Publishing in Ronneby*, May, pp. 10-12.

Kim, J. (2010), "Faculty self-archiving: motivations and barriers", *Journal of the American Society for Information Science and Technology*, Vol. 61 No. 9, pp. 1909-1922.

Kling, R. and McKim, G. (2000), "Not just a matter of time: field differences and the shaping of electronic media in supporting scientific communication", *Journal of the American Society for Information Science*, Vol. 51 No. 14, pp. 1306-1320.

Kling, R., Spector, L.B. and Fortuna, J. (2004), "The real stakes of virtual publishing: the transformation of E-Biomed into PubMed central", *Journal of the American Society for Information Science and Technology*, Vol. 55 No. 2, pp. 127-148.

Kousha, K. and Thelwall, M. (2014a), "Disseminating research with web CV hyperlinks", *Journal of the American Society for Information Science and Technology*, Vol. 65 No. 8, pp. 1615-1626, available at: www.researchgate.net/publication/256433340_Disseminating_Research_with_Web_CV_Hyperlinks/file/3deec5228643f90bff.pdf (accessed 4 March 2014).

Kousha, K. and Thelwall, M. (2014b), "Web impact metrics for research assessment", in Cronin, B. and Sugimoto, C.R. (Eds), *Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact*, MIT Press, Cambridge, pp. 289-306.

Kousha, K., Thelwall, M. and Abdoli, M. (2012), "The role of online videos in research communication: a content analysis of youtube videos cited in academic publications", *Journal of the American Society for Information Science and Technology*, Vol. 63 No. 9, pp. 1710-1727.

Kurtz, M.J. and Bollen, J. (2010), "Usage bibliometrics", *Annual Review of Information Science and Technology*, Vol. 44, pp. 1-64.

Laakso, M. (2014), "Green open access policies of scholarly journal publishers: a study of what, when, and where self-archiving is allowed", *Scientometrics*, Vol. 99 No. 2, pp. 475-494, available at: http://dx.doi.org/10.1007/s11192-013-1205-3 (accessed 1 March 2014).

Larivière, V., Sugimoto, C.R., Macaluso, B., Milojević, S., Cronin, B. and Thelwall, M. (2014), "Arxiv E-prints and the journal of record: an analysis of roles and relationships", *Journal of the Association for Information Science and Technology*, doi: 10.1002/asi.23044.

Lyons, C. and Booth, H.A. (2011), "An overview of open access in the fields of business and management", *Journal of Business & Finance Librarianship*, Vol. 16 No. 2, pp. 108-124.

McVeigh, M.E. (2004), "Open access journals in the ISI citation databases: analysis of impact factors and citation patterns a citation study from thomson scientific", Thomson Reuters, available at: http://ip-science.thomsonreuters.com/m/pdfs/openaccesscitations2.pdf (accessed 17 January 2012).

Más-Bleda, A., Thelwall, M., Kousha, K. and Aguillo, I.F. (2014), "Successful researchers publicizing research online: an outlink analysis of European highly cited scientists' personal websites", *Journal of Documentation*, Vol. 70 No. 1, pp. 148-172.

Miguel, S., Chinchilla-Rodriguez, Z. and de Moya-Anegón, F. (2011), "Open access and scopus: a new approach to scientific visibility from the standpoint of access", *Journal of the American Society for Information Science and Technology*, Vol. 62 No. 6, pp. 1130-1145.

Moed, H.F. (2007), "The effect of 'open access' on citation impact: an analysis of ArXiv's condensed matter section", *Journal of the American Society for Information Science and Technology*, Vol. 58 No. 13, pp. 2047-2054.

Morris, S. (2009), "Journal authors' rights: perception and reality", *Publishing Research Consortium*, available at: www.publishingresearch.org.uk/documents/JournalAuthorsRights.pdf (accessed 1 March 2014).

Mulligan, A., Hall, L. and Raphael, E. (2013), "Peer review in a changing world: an international study measuring the attitudes of researchers", *Journal of the American Society for Information Science and Technology*, Vol. 64 No. 1, pp. 132-161.

Nariani, R. and Fernandez, L. (2012), "Open access publishing: what authors want", *College & Research Libraries*, Vol. 73 No. 2, pp. 182-195.

Neuendorf, K.A. (2002), *The Content Analysis Guidebook*, SAGE Publications Inc., London.

Novarese, M. and Zimmermann, C. (2008), "Heterodox economics and dissemination of research through the internet: the experience of RePEc and NEP", *On The Horizon-The Strategic Planning Resource for Education Professionals*, Vol. 16 No. 4, pp. 198-204.

OpenDOAR (2015), "OpenDOAR – Charts – Worldwide", available at: www.opendoar.org/find.php?format=charts (accessed 2 January 2015).

Pinfield, S., Salter, J., Bath, P., Hubbard, B., Millington, P., Anders, J.H.S. and Hussain, A. (2014), "Open-access repositories worldwide, 2005-2012: past growth, current characteristics and future possibilities", *Journal of the American Society for Information Science and Technology*, Vol. 65 No. 12, pp. 2404-2421, available at: http://eprints.whiterose.ac.uk/76839/ (accessed 16 February 2014).

PMC (2014), "PMC International", available at: www.ncbi.nlm.nih.gov/pmc/about/pmci/ (accessed 3 March 2014).

Poynder, R. (2012), "Open access mandates: ensuring compliance", *Open and Shut*, available at: http://poynder.blogspot.fi/2012/05/open-access-mandates-ensuring.html (accessed 2 January 2015).

RePEc (2014), "IDEAS: rankings", available at: http://ideas.repec.org/top/ (accessed 9 August 2014).

Schwarz, G.J. and Kennicutt, R.C. Jr (2004), "Demographic and Citation Trends in Astrophysical Journal papers and Preprints", arXiv:astro-ph/0411275, available at: http://arxiv.org/abs/astro-ph/0411275 (accessed 2 March 2014).

Skeels, M.M. and Grudin, J. (2009), "When social networks cross boundaries: a case study of workplace use of Facebook and LinkedIn", *Proceedings of the ACM 2009 International Conference on Supporting Group Work, ACM*, pp. 95-104.

Sotudeh, H. and Horri, A. (2007), "The citation performance of open access journals: a disciplinary investigation of citation distribution models", *Journal of the American Society for Information Science and Technology*, Vol. 58 No. 13, pp. 2145-2156.

Spezi, V., Fry, J., Creaser, C., Probets, S. and White, S. (2013), "Researchers' green open access practice: a cross-disciplinary analysis", *Journal of Documentation*, Vol. 69 No. 3, pp. 334-359.

SSRN (2014), "Home: SSRN", available at: www.ssrn.com/en/ (accessed 9 August 2014).

Suber, P. (2012), *Open Access*, MIT Press, Boston, MA, available at: http://cyber.law.harvard.edu/hoap/Open_Access_(the_book) (accessed 11 January 2015).

Swan, A. (2010), "The open access citation advantage: studies and results to date", available at: http://eprints.soton.ac.uk/268516/ (accessed 2 March 2014).

Swan, A. and Brown, S. (2005), "Open access self-archiving: an author study", available at: http://cogprints.org/4385 (accessed 2 March 2014).

Tenopir, C., Mays, R. and Wu, L. (2011), "Journal article growth and reading patterns", *New Review of Information Networking*, Vol. 16 No. 1, pp. 4-22.

Thelwall, M. (2004), *Link analysis: An Information Science Approach*, Emerald Group Pub Ltd, New York.

Thelwall, M. and Kousha, K. (2014), "Academia.edu: social network or academic network?", *Journal of the Association for Information Science and Technology*, Vol. 65 No. 4, pp. 721-731.

Thomson Reuters (2014), "Web of Science Core Collection Help", available at: http://images.webofknowledge.com/WOKRS517B4/help/WOS/hp_database.html (accessed 1 March 2014).

Wagner, B. (2010), "Open access citation advantage: an annotated bibliography", *Issues in Science and Technology Librarianship*, No. 60, available at: www.istl.org/10-winter/article2.html (accessed 3 January 2014).

Walshe, E. (2001), "Creating an academic self-documentation system through digital library interoperability: the RePEc model", *New Review of Information Networking*, Vol. 7 No. 1, pp. 43-58.

Xia, J. (2008), "A comparison of subject and institutional repositories in self-archiving practices", *The Journal of Academic Librarianship*, Vol. 34 No. 6, pp. 489-495.

Zimmermann, C. (2013), "Academic rankings with RePEc", *Econometrics*, Vol. 1 No. 3, pp. 249-280.

Appendix 1

Table AI. Scopus citing documents queries

SR	Query
arXiv	(WEBSITE(*arxiv*) OR WEBSITE(*xxx.lanl.gov*) OR REFSRCTITLE(arxiv)) AND (PUBYEAR > 1999) AND (PUBYEAR < 2014)
RePEc	WEBSITE(*repec.org*) AND (PUBYEAR > 1999) AND (PUBYEAR < 2014)
SSRN	(WEBSITE(*ssrn*) OR REFSRCTITLE(ssrn)) AND (PUBYEAR > 1999) AND (PUBYEAR < 2014)
PMC	WEBSITE("*ncbi.nlm.nih.gov/pmc*") AND (PUBYEAR > 1999) AND (PUBYEAR < 2014)

Appendix 2

Table AII. Scopus subject area codes used in the SUBJAREA() command

Subject area code	Subject area description
agri	Agricultural and biological sciences
arts	Arts and humanities
bioc	Biochemistry, genetics and molecular Biology
busi	Business, management and accounting
ceng	Chemical engineering
chem	Chemistry
comp	Computer science
deci	Decision sciences
dent	Dentistry
eart	Earth and planetary sciences
econ	Economics, econometrics and finance
ener	Energy
engi	Engineering
envi	Environmental science
heal	Health professions
immu	Immunology and microbiology
mate	Materials science
math	Mathematics
medi	Medicine
neur	Neuroscience
nurs	Nursing
phar	Pharmacology, toxicology and pharmaceutics
phys	Physics and astronomy
psyc	Psychology
soci	Social sciences
vete	Veterinary
mult	Multidisciplinary

Corresponding author
Dr Xuemei Li can be contacted at: lixuemei@yorku.ca