Successful Researchers Publicizing Research Online: An Outlink Analysis of European Highly Cited Scientists' Personal Websites[1]

Amalia Más-Bleda*, Mike Thelwall**, Kayvan Kousha**, Isidro F. Aguillo*

* The Cybermetrics Lab, Institute of Public Goods and Policies (IPP), Spanish National Research Council (CSIC), Madrid, Spain.

** Statistical Cybermetrics Research Group, School of Technology, University of Wolverhampton, Wolverhampton, United Kingdom.

Abstract

Purpose - This study explores the link creating behaviour of European highly cited scientists based upon their online lists of publications and their institutional personal websites.
Methods - A total of 1,525 highly cited scientists working at European institutions were first identified. Outlinks from their online lists of publications and their personal websites pointing to a pre-defined collection of popular academic websites and file types were then gathered by a personal web crawler.
Findings - Perhaps surprisingly, a larger proportion of social scientists provided at least one outlink compared to the other disciplines investigated. By far the most linked-to file type was PDF and the most linked-to type of target website was scholarly databases, especially the Digital Object Identifier website. Health science and life science researchers mainly linked to scholarly databases, while scientists from engineering, hard sciences and social sciences linked to a wider range of target websites. Both book sites and social network sites were rarely linked to, especially the former. Hence, whilst successful researchers frequently use the Web to point to online copies of their articles, there are major disciplinary and other differences in how they do this.
Research limitations – The number of women in the sample is very low, and the sample inherits any discipline-bias and language-bias of the database used (ISIHighlyCited.com) for the identification of the scientists. For part of the analysis, we did not analyze all links created by the scientists, but only outlinks pointing to specific types of websites that were common hyperlink targets from researchers' web pages.
Practical implications – The differences found in the way that scientists use the web to publicize research will inform the design of future web based research evaluation and open access indicators so that they can be sensitive to different but valid strategies for web research publicity. The understanding of how successful scientists use the web to publicize their research and which web resources have the most success in each discipline will also help current researchers to plan and evaluate their own web strategies.
Originality/value - This is the first study to analyse the outlinking patterns of highly cited researchers' institutional web presences in order to identify which web resources they use to provide access to their publications.

Keywords: Link analysis, outlinks, highly cited researchers, open access.
Article Classification: Research paper.

1. Introduction

The web has provided scientists with a historically unique opportunity to present a public face to the world through their home page or online CV. Hosted on a university website or elsewhere, and under the control of the individual scholar or their institution, a personal website (i.e., a personal home page and related web pages, such as publication lists) can exhibit information about the scientist and their

[1] . This is a preprint of an article to be published in the Journal of Documentation © copyright Emerald Group Publishing Limited 2013.

research to promote their achievements and skills. Since research dissemination is important for scholars, (it would be useful to know how personal websites are used in practice to provide access to online information. Understanding both standard practice and common variations for personal websites would help researchers to identify the kinds of content that they could or should publish and can also inform the future development of indicators to assess the impact of personal home pages on the scholarly communication process.

The potential to extend bibliometric analyses to the Internet has long been recognized (Bar-Ilan, 2000, 2001; Chen *et al*., 2009; Pitzek 2002). As part of this, many studies have argued for the importance of Open Access (OA) academic information for scientific progress (e.g., Harnad, 2011; Rossini, 2007; Swan, 2007) and others have examined the value of different web sources for scholarly impact assessment (e.g., Kousha and Thelwall, 2007). It is also known that scientists use the web in different ways (Barjak, 2006), but no previous study has investigated how researchers use their personal home pages and online CVs or publication lists to link to content that may be useful to visitors. Given the widespread use of personal websites and controversies about OA publishing (see below), this seems to be an important omission.

Successful researchers are particularly important for scholarly progress and highly cited researchers presumably tend to be regarded as the most successful, however success is defined. For pragmatic purposes, this article focuses only on researchers from Europe. Hence, this study investigates the web resources (e.g., OA repositories, databases, fee-based publisher sites and social network sites) linked to by European Highly Cited (EHC, defined more precisely below) scientists from their institutional personal home pages and online publication lists to provide access to their academic publications and other resources.

2. Related research

2.1 Link analysis in academic web space

Since 1996, many investigations have been conducted on Web links to develop indicators for research assessment to supplement traditional bibliometric measures. However, an important issue for link analysis research is still how to count web links in a proper systematic manner to assess aspects of research communication. Researchers have worked on link analysis from different angles since its origins within information science (Ingwersen, 1998; Rodríguez Gairín, 1997). Wilkinson, Harries, Thelwall and Price (2003) investigated motivations for academic web site interlinking, concluding that metrics based upon link counts reflect an agglomeration of connections related to scholarly activities in a wide variety of ways. Several studies have analyzed the patterns of interlinking between university web sites to see whether there were geographic (Thelwall, 2002c), linguistic (Thelwall *et al*., 2003) and disciplinary (Tang and Thelwall, 2003) trends (see also Seeber *et al*., 2012). Others have compared sources of links for academic Web Impact Factor calculations (Thelwall, 2002a), have investigated new links metrics to complement the existing Web Impact Factor (WIF) (Ingwersen, 1998) for academic web sites and have examined the theoretical framework for interpreting link analysis, concluding that no single method is perfect for link interpretation and thus the use of multiple methods (method triangulation) is required (Thelwall, 2006). A recent investigation (Thelwall *et al*., 2012) seems to show a new trend, by assessing URL citations and title mentions as possible replacements for hyperlinks, when commercial search engines are used for data collection.

There are also many studies related to link analysis in national or regional academic web spaces (e.g., Bar-Ilan, 2004; Chung *et al*., 2009; Jalal *et al*., 2010; Ortega and Aguillo, 2007; Thomas and Willet, 2000; Qiu *et al*., 2004; Vaughan and Thelwall, 2005; Seeber *et al*., 2012), one common characteristic being that they have investigated interlinking between universities or departments**.** Early studies counted either individual hyperlinks or individual pages containing links and the latter was often the only choice if search engines were used for raw data; however, Thelwall (2002b) showed that this was not necessarily the best approach and proposed other alternatives, subsequently used for a range of studies (Chung *et al*., 2009; Payne and Thelwall, 2004, 2009). These document-based models were: web page, directory, domain and university or website (Thelwall, 2002b). For example, the website model counts the number of websites containing links rather than the number of linking pages.

2.2 Outlink analyses: Experiments and methods

Although most web link studies have investigated inlinks (incoming links directed to a website of interest) because they have been available from commercial search engines, outlink analyses have been less common in scholarly communication research and the investigations have had different purposes. Ajiferuke and Wolfram (2004) investigated the frequency distribution of outlinks within web pages, while Vaughan and Wu (2004) focused on links to commercial websites. Nevertheless, a subject of particular interest lately seems to be co-outlinking, which has been used, for instance, for checking if networks of sites can map triple helix structures (García-Santiago and Moya Anegón, 2009) and for showing common interests in municipal websites (Holmberg, 2011).

Link analyses have often used commercial search engines for collecting the raw hyperlink data via their Applications Programming Interfaces - APIs (Thelwall *et al*., 2012). These APIs allow automated data collection by letting programmers write code to access search engine results (Thelwall and Sud, 2012). However, in April 2011, Yahoo! withdrew all support for automatic searches, as normally used in webometric research, leaving no remaining automatic source of link data from search engines (Thelwall and Sud, 2011). For this reason other alternatives to search engines for gathering link data were necessary, and one of these alternatives was the personal web crawler. According to Yang and Qin (2008), self-developed crawling software could be applied to check the stability, reliability and web coverage of search engines.

A web crawler automatically follows hyperlinks to find and download web pages (Thelwall, 2001) and can crawl either an area of the Web or a set of websites. One important role of personal crawlers is to cross-check the results of commercial search engines (Thelwall, 2001), although a limitation is that they can only be used on a modest scale because of the time and computer resources needed (Thelwall *et al*., 2012). Linked pages can also be missed "if the links are in a form that the crawler does not understand, if the pages are password protected, if the server is temporarily down or if the owner requests that the page is not to be crawled" (Thelwall *et al*., 2005, p. 92). Moreover, crawlers can incur financial penalties to the owners of the websites crawled (Thelwall and Stuart, 2006) and unethical crawlers can generate spam and denial of service attacks (Sun, 2008). These are some reasons why it is necessary to follow ethical guidelines for web crawler use (Giles *et al*., 2010; Koster, 1993). Two important limitations for using web crawlers are that dynamic pages may not be gathered and so the results may not be comprehensive (Payne and Thelwall, 2007) and it is very difficult to interpret web link data because of the variety of reasons why web links are created (Wilkinson *et al*., 2003).

2.3 Highly cited scientists

It is known that not all scientists use the Web in the same way as each other. For example, Barjak (2006) showed that, almost a decade ago, younger scientists used the Internet more often for informal communication than older scientists, that male scientists were more likely to have their own web page than female scientists. Moreover, the higher the research productivity of scientists, the more they rely on the Internet for their informal communication, although scientists with very high research output did not use it more than scientists with high research output. He also showed that scientists that participated in research collaborations used Internet more intensely than scientists that did not collaborate.

The present study is concerned with the most influential and important scientists and as a proxy for this we focused on highly cited researchers, operationalized as being one of the 250 most cited authors of journal papers in each of the 21 disciplines identified by the Institute for Scientific Information (ISI) - now Thomson Reuters - between 2000 and 2008. Previous investigations have focused on similar populations. Batty (2003), for instance, analyzed the distribution of these scientists by country, place and institution and Basu (2006) used this population to obtain a country level indicator of citation excellence. Their social characteristics (Parker *et al*., 2010) and participation in the top leadership of universities (Ioannidis 2010) have also been studied. We are particularly interested in highly cited researchers working at European institutions. A recent study (Mas-Bleda and Aguillo, in press) revealed their distribution by country and discipline, indicating that 64% of EHC scientists had a personal website, with a larger proportion in Denmark, Israel and the United Kingdom and from Economics, Mathematics, Computer sciences and Space Sciences. Furthermore, "Life Sciences and Health Sciences researchers, especially those of Molecular Biology and Genetics, preferred to use the research group website rather than personal website to divulge their academic and research activities" (Mas-Bleda, and Aguillo, in press).

2.4 Open access and online publishing

The Web is one of the popular sources for OA publishing. OA is the "immediate, permanent, free online access to the full text of all refereed research journal article" (Harnad, 2005). According to Suber (2004, p. 1) "OA literature is not only free of charge to everyone with an internet connection, but free of most copyright and licensing restrictions", although this may no longer be the norm. OA Initiatives are growing stronger; the association of All European Academies (ALLEA) recently described an optimal panorama for Open Science in the 21st century for Europe (ALLEA 2012). The Budapest Open Access Initiative (BOAI), created in 2002, was designed to "accelerate progress in the international effort to make research articles in all academic fields freely available on the internet" (BOAI, 2002) and has recently created recommendations for the next ten years (BOAI, 2012).

There are two recognized complementary ways to make research freely accessible online: OA journals (also named "Gold OA") publish all articles on the web with unrestricted access (e.g. through journal websites); and self-archiving (also named "Green OA") by authors - publishing a preprint of their article online with publisher permission (Gargouri *et al*., 2012). In both cases, but presumably more in the latter, scientists can be expected to link to their own publications to make it easier for their home page visitors to access them. Nowadays OA research (either published works or grey literature) can be found in several channels: authors' websites, research group websites, institutional repositories, discipline-specific archives, journal websites, journal platforms and other portal sites (Matsubayashi *et al*, 2009).

A researcher's personal website is not only a useful platform to publicize his or her peer-reviewed publications (Barjak, 2006; Barjak *et al*., 2007; Björk, *et al*., 2010), but could also provide access to grey literature created as a result of research activities, such as preprints, working papers, technical reports, PowerPoint presentations related to conferences, lectures, seminars and teaching. In addition, a CV could provide links related to the scholar's education, career and research projects, or any software or other electronic resources created. Nevertheless, this way has been considered unstable for OA "because the availability of articles depends on the authors' voluntary contribution" (Matsubayashi *et al*, 2009, pp. 5-6). These publications can play an important role; for example the grey literature is essential for effective medical meta-analyses (McAuley *et al*., 2000).

Some grey literature repositories allow authors to publicly archive unpublished work in place that is highly visible to peers, perhaps bypassing the need for home page publicity. As an example, the arXiv repository plays an essential role in scholarly communication within physics (McKiernan, 2000; Pinfield, 2001) and the RePEc repository within economics (Seiler and Wohlrabe, 2011). Finally, scholars may also link to social network sites, such as Facebook, Academia.edu or LinkedIn, that may provide visitors a convenient way to communicate with them or that may provide additional information to that which is available on the home page. The assessment of OA publishing through an outlink analysis of EHC researchers could provide a better understanding of overall outlinking patterns to web resources such as OA repositories, databases, fee-based publishers' sites and social network sites.

3. Research questions

This paper reports an outlink analysis of EHC scientists' online lists of publications in order to get raw data to help future judgments about how to develop future research evaluation and OA indicators. Moreover, since researchers may not have a specific web page for their publication record, but may list them instead within a personal website, we also report an outlink analysis of EHC scientists' personal websites and combined results from the two sources.

From a scientist's list of publications, outlinks may point to abstracts, metadata or full text copies. However, from a scientist's personal website, outlinks may also point to other sources, such as the website of the research group, department, faculty or university to which he/she belongs, research projects, funding sources, raw data, teaching and/or informative materials as well as to web resources (repositories, publishers' websites, databases and social network sites). In this paper we focus on the latter.

The primary research questions guiding the present research are:

- Which web resources are linked to from EHC scientists' online lists of publications and institutional personal websites (e.g., OA repositories, fee-based publishers' websites, publication databases -either OA or non-OA, book databases or booksellers, social network sites)?
- Are rich file formats (e.g., PDF), which suggest significant research resources, commonly linked to by these researchers?
- Are there differences between disciplines in the results of the above research questions?

4. Methods

4.1 Selection of highly cited scientists

The first step was the identification of highly cited researchers working at European institutions and their institutional personal websites. The complete list of 45 European countries was chosen from the official website of the European Union (http://europa.eu/about-eu/countries/index_en.htm): the 27 member states of the European Union, 5 candidate countries and another 13 European countries. We listed the highly cited researchers working in these countries, based on the ISIHighlyCited.com database created by ISI/Thomson Reuters. This database contained the 250 most highly cited researchers during 1981-2008 in each of 21 disciplines, grouped into five broader areas: engineering (computer science, engineering, geosciences, materials science), hard sciences (chemistry, mathematics, physics, space sciences), health sciences (clinical medicine, immunology, microbiology, neuroscience, pharmacology), life sciences (agricultural sciences, biology and biochemistry, ecology/environment, molecular biology and genetics, plant and animal science) and social sciences (economics/business, psychology/psychiatry, general social science). Only a minority of the academics worked at European institutions and the remainder were manually identified and removed. The highly cited researchers only belonged to 22 of the 45 selected countries. For each scientist we collected, if available, the following information: name, discipline, country of affiliation, country of birth, citizenship, year of birth, gender and institutional personal website URL.

Before the data collection was complete (September 2011), the ISIHighlyCited.com database was replaced by a new online directory (http://researchanalytics.thomsonreuters.com/highlycited/). The new directory only listed the researcher's name, affiliation and discipline, failing to provide the additional information. As both tools - database and online directory - were therefore incomplete and outdated, we adopted the following search strategies to fill any gaps: 1) the Web of Knowledge (WoK) publication and citation database (created by ISI/Thomson Reuters and covering sciences, social sciences, arts and humanities) was used to find the full name of some researchers, 2) Google was used to locate the institution to which they belonged and their institutional personal websites, and 3) researchers' personal websites were consulted to obtain personal information such as their date and country of birth.

There are several sample selection limitations. The discipline-bias of the database used (ISIHighlyCited.com) is likely to cause the social sciences to be underrepresented and the humanities to be completely excluded. There is also a language-bias in this database towards English speaking countries, especially toward the United States (Batty 2003, Leeuwen *et al*., 2001). Although this database provides the 250 most highly cited researchers in each of 21 disciplines, the total number of researchers in these disciplines is unknown, and so the representation of highly cited researchers may have additional disciplinary biases, even within the hard sciences. A fourth limitation is that the study is restricted to highly cited scientists working at European institutions, which represented less than a quarter of the highly cited researchers identified by ISI/Thomson Reuters (Mas-Bleda and Aguillo, in press).

The number of women in the population was very low (5% of the EHC researchers) which may be due to systematic bias inherent in the method used. A previous investigation (Parker *et al*, 2010) into the most highly cited environmental scientists and ecologists had a similarly low percentage (5.5%) and the reason may be that "career prospects for female university researchers are clearly worse than for their male counterparts" (Danell and Hjerm, 2013, p. 1004), that female researchers tend to be less cited than men (Aksnes *et al*. 2011; Kretschmer *et al*., 2012; Pudovkin *et al*., 2012) and that the higher the academic position, the lower the presence of women (Mauleón and Bordons 2006; Prpic 2002; Torres-

Salinas *et al*., 2011). Moreover, since women tend to have larger career breaks than men for childbirth (Reskin and Bielby, 2005) and also disproportionately take on extra caring responsibilities (e.g., for aging parents) (Carmichael and Charles, 2003) any measure based upon total citations is biased, on average, against women in a way that, for example, measures based upon the average number of citations per publication would not be.

In order to partially correct for this source of gender bias Microsoft Academic Search (http://academic.research.microsoft.com/) was used to increase the percentage of women, since it was the only citation database that offered rankings of scientists ordered by total citations received. This database provides 15 field rankings (one is "Multidisciplinary"). The original plan was to check the top 1,000 scientists in each field (except Arts and Humanities), but we only checked the top 500 scientists because there were already large differences in the number of citations between the first and the 500th researcher. We hoped to add about 200-300 extra women to the population so that females could represent a quarter of the total but this process only yielded 91 extra women. Hence the final population consisted of 1,589 EHC researchers, which represent almost a quarter of total of the highly cited researchers identified by ISI/Thomson Reuters. Deceased researchers were removed on the assumption that their personal websites may not be maintained, a total of 64. The final population for this study thus consisted of 1,525 living EHC scientists, 1365 (90%) men and 160 (10%) women. Table 1 displays the distribution by country and discipline of this population. Data collection was completed in May 2012.

Table 1. Distribution by country and discipline of 1,525 EHC researchers.

Country	**Discipline**					
	Engineering	Hard sciences	Health sciences	Life sciences	Social sciences	**Total**
UK	59	104	153	123	49	**488**
Germany	44	78	65	71	0	**258**
France	38	36	43	43	4	**164**
Switzerland	18	37	29	25	4	**113**
Netherlands	15	19	35	31	3	**103**
Italy	15	29	29	11	3	**87**
Sweden	8	6	23	29	4	**70**
Israel	18	14	3	10	2	**47**
Belgium	5	4	19	8	3	**39**
Spain	3	9	4	20	2	**38**
Denmark	8	5	7	8	2	**30**
Finland	4	3	7	9	0	**23**
Austria	3	6	8	1	0	**18**
Norway	0	0	4	10	0	**14**
Ireland	1	0	1	9	0	**11**
Hungary	0	2	2	1	1	**6**
Russia	2	1	0	2	0	**5**
Greece	1	1	1	1	0	**4**
Poland	0	0	1	2	0	**3**
Romania	1	0	0	1	0	**2**
Cyprus	0	0	0	1	0	**1**
Portugal	0	0	1	0	0	**1**
Total	**243**	**354**	**435**	**416**	**77**	**1,525**

4.2 Selection of scientists' websites

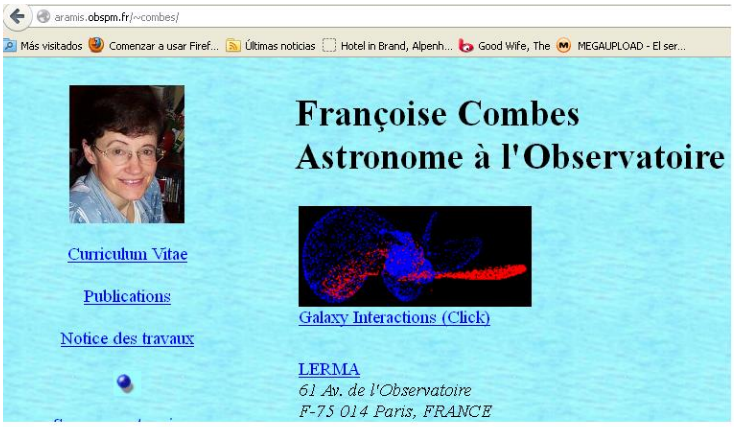
We limited the study to pages associated with EHC scientists within institutional websites. Some scientists used research group websites rather than personal pages and so these were also collected.

- An *institutional website* was understood to be a website hosted on a web domain of an academic institution (university, faculty, department, institute, laboratory, research group). We recognized two different types:
  - *Personal website*: an institutional website created by or for a researcher regardless of content. It typically includes biographical information, a Curriculum Vitae (CV), a list of publications, talks, and/or class materials.
  - *Research group website*: a website hosted on the web domain of an academic institution and focused on a research group or laboratory.

After identifying the population and their institutional websites (personal websites and research group websites), we identified if these researchers had a specific page for their Web CV and for their academic publications.

- A *Web CV* was understood to be a web page within an institutional website that provides only the researcher's CV. It can include a short CV (only biographical information) or a long CV (biographical information and publications).
- An *online list of publications* was understood to be a web page within either an institutional personal website or a research group website that provides only a list of the researcher's academic publications.

To illustrate this classification, an example is shown in Figure 1. Françoise Combes is a EHC researcher working at *Astronome à l'Observatoire de Paris*. Her personal website (http://aramis.obspm.fr/~combes/, see Figure 1), is hosted within the website of her institution (http://obspm.fr/). At the top left of her personal website (see Figure 1) there is a link to her CV and a link to her publications. The first link gives her Web CV (http://aramis.obspm.fr/~combes/fcombes/cv_fc.html) and the second link gives an online list of her publications (http://aramis.obspm.fr/~combes/pub_fc.html). Hence, the personal website is a set of web pages providing different types of contents, while the online list of publications is a unique web page (usually within a personal website) cataloging a researcher's publications.



Françoise Combes
Astronome à l'Observatoire

Curriculum Vitae

Publications

Notice des travaux

Galaxy Interactions (Click)

LERMA
61 Av. de l'Observatoire
F-75 014 Paris, FRANCE

Figure 1: Example of personal website.

4.3 Identifying outlinks from scientists' websites

We had initially intended to gather the outlinks from all the major common sources of pages of potentially academic-related links associated with individual researchers: personal websites, Web CVs and online publication lists but web CVs were subsequently excluded. Of the 247 Web CVs identified, only 13 listed publications in HTML format (providing links pointing to the abstract or full text of publications). This tiny proportion was excluded from the outlink analysis to keep the source document type uniform and this did not substantially affect the results.

Many researchers included in their personal websites either a list of recent publications or a list of all publications, though they did not have a specific web page for them. For this reason we carried out two different outlink analyses:

- An outlink analysis of researchers' online lists of publications. This analysis only covered those researchers that had a specific web page for their publications.

- An outlink analysis of researchers’ personal websites. This analysis included scientists that provided a list of their publications but did not have a specific page for them.

Many personal websites were multilingual, that is, they were in the local language(s) and English. In such cases only the English version was used because if both versions were used then the links would be counted twice.

4.4 Tools used for generating outlinks

Several tools have been created for link data collection, such as *SocSciBot* (Thelwall, 2001), *iWatch Web Crawler* (Jensen, 2007) and *LinkDiscover* (Yang and Qin 2008). In this paper, *SocSciBot4.1* (http://socscibot.wlv.ac.uk/) and *Webometric Analyst* 2.0 (http://lexiurl.wlv.ac.uk/) were used to automatically generate different types of outlink statistics from EHC scientists’ online lists of publications and personal websites, in order to determinate which web resources (repositories, databases, fee-based publishers and social networks sites) were being linked by them to provide (full text) access to their publications. *SocSciBot*, a web crawler for link analysis research, was used for crawling the websites and web pages, and *Webometric Analyst* was used to analyze links from websites and web pages crawled.

More specifically, *SocSciBot* was used to crawl the EHC researchers’ personal websites and online lists of publications in June 2012. As part of the web crawling, this program removed duplicate URLs (Thelwall, 2002b). However, sometimes it did not crawl web pages due to apparently temporary problems (e.g., server busy) and broken links.

The pages downloaded by the *SocSciBot* crawls were processed by *Webometric Analyst* to extract their hyperlinks and, for each extracted hyperlink, to identify and list the type of website and type of file linked to (see section 4.4.1).

Using *Webometric Analyst*, we analyzed outlinks from the EHC scientists’ web presences in two different ways. First, we analyzed outlinks to specific types of websites that were common hyperlink targets from researchers’ websites. For this, a list of websites was compiled by examining the most common websites linked to by a collection of researchers’ websites. The list included 25 OA repositories, 38 fee-based publishers’ websites, 13 other OA or non-OA databases, 5 book databases or book publishers and 12 social network sites, as summarized in Table 2. The results of this should indicate a range of specific types of linking in the websites. Second, the frequency of outlinks to 11 common file types was also analyzed by file name extension (pdf, doc, docx, ps, rtf, gz, zip, ppt, pptx, wmv, mp3). These were the most common non-HTML file name extensions found in a pilot test of links from researchers’ websites. Linking to a non-HTML resource suggests a more substantial reason for linking than for the average link to a web page and these links are expected to be mainly to a researcher’s own publications (pdf, doc, docx, ps, rtf, gz, zip, ppt, pptx) or resources supporting their research (wmv videos, mp3 sound recordings).

Table 2. Common websites identified in the study.

Resource type	**Specific resources**
Open access repositories	arxiv.org, cdsads.u-strasbg.fr,slac.stanford.edu, philpapers.org, biomedcentral.com, papers.ssrn.com, philsci-archive.pitt.edu, sammelpunkt.philo.at, cdsweb.cern.ch, eprints.ma.man.ac.uk, vixra.org, adsabs.harvard.edu, citeseer.ist.psu.edu, osti.gov, odysci.com, cogprints.org, xxx.lanl.gov, citeseerx.ist.psu.edu, dialnet.unirioja.es, scielo, redalyc.uaemex.mex, hal.archives-ouvertes.fr, repec, diva-portal.org.
Fee-based publishers' websites	springerlink.com, springer.com, springermedizin, .wiley.com, blackwell-synergy.com, reddit.com, journals.uchicago.edu, journals.cambridge.org, aip.org, sciencedirect.com, tandfonline, sagepub, oxfordjournals.org, jstor.org, proquest.com, ieeexplore.ieee.org, emeraldinsight.com, iop.org, ebscohost.com, ebsco.com, routledge.com, refdoc.fr, nature.com, lww.com, bmj.com, cambridge.org, acm.org, aps.org, ascelibrary.org, plosone.org, engineeringvillage2.org, bioone.org, ingentaconnect.com, steiner-verlag, agricola.nal.usda.gov, elsevier.com, scholarsportal.info, thieme-connect.com.
Scholarly databases	eric.ed.gov, mendeley.com, dx.doi.org, scholar.google, nlm.nih.gov,

(either OA or non-OA)	scirus.com, highwire.standford.edu, academic.research.microsoft.com, inspirehep.net, citeulike.org, oclc.org, base.ub.uni-bielefeld.de, handle.net.
Book databases or booksellers	bookshop.blackwell, bookstore, amazon, books.google, bookdepository.
Social websites	youtube.com, facebook.com, twitter.com, myspace.com, researchgate.net, linkedin.com, delicious.com, academia.edu, arnetminer.org, readrmeter.org, slideshare.net, wikipedia.

5. Results

As summarized in Table 3, two thirds of the scientists had an institutional website (i.e., a personal website or research group website), 16% had a specific page for a Web CV and about one third had an online list of publications. The online lists of publications were provided mainly by researchers with a personal website, while researchers with a research group website did not usually include a list of their publications separately from a list of the group's publications.

Regarding the format of the online lists of publications, 366 (76%) were in HTML, 62 (13%) were in PDF, and 17 (3.5%) in both formats, while 36 lists (7.5%) were within an institutional repository. Both HTML and PDF page types were examined, although outlinks were not extracted from PDF pages, if they contained any. Other page types, such as .txt or .doc, were ignored because most contained no links.

Table 3. Web presences identified for the 1,525 EHC scientists.

Web presence type	No. of scientists	
Institutional web presence of any of the types below	1,016	(67%)
Personal website	934	(61%)
Page within research group website	126	(8%)
Separate Web CV	247	(16%)
Separate online list of publications	481	(32%)

5.1 Types of analyses performed

Two outlink analyses were carried out: the first for the 481 scientists with an online list of publications and the second for the 937 scientists with an institutional personal website. Some scientists had more than one list of publications and more than one personal website, so 561 online lists of publications and 1074 personal websites were crawled in total.

The online publication lists generated 16,027 outlinks, while personal websites generated 23,326 matching outlinks. These figures include only links pointing to the resources in Table 2 and exclude all outlinks to other sites (e.g., links to institutional repositories). The total set of outlinks was classified into two groups: those that pointed to non-HTML files and those that pointed to HTML files. Non-HTML files were links with the non-HTML file extensions listed above (pdf, doc, docx, ps, rtf, gz, zip, ppt, pptx, wmv, mp3), regardless of the server in which they were hosted. HTML files were classified as those with any other extension (e.g., .htm, .html, .asp, .php). This includes a small number of rare non-HTML file types but is predominantly a collection of HTML files. Approximately 50% of the outlinks pointed to each of the two groups (non-HTML files and HTML files) from both the personal websites and the online lists of publications

The following should be considered when interpreting the results.

a) Within online lists of publications, outlinks to non-HTML files seemed to be always links to full text publications but this was not true for personal websites.
b) An outlink to a HTML file did not necessarily point to a full text publication, but might also point to an abstract, the journal where the paper was published, or a repository or database.
c) Although the initial sample included 481 researchers with lists of publications and 937 researchers with personal websites, sometimes the program did not work (for reasons discussed in the Methods section) so the analyses were performed for 425 EHC researchers with an online

list of publications and 873 EHC researchers with a personal website, and the outlinks found were created by 215 and 488 researchers respectively.

d) There were 892 scientists (58.5% of 1,525 EHC researchers) with either a personal website or an online list of publications for which the crawler worked, of which 537 had at least one outlink.

The data had a highly skewed distribution (most links were created by a few researchers) with a large number of zeros and so neither the median nor the mean are helpful to describe it. Instead, the proportion of EHC scientists with at least one outlink to the resource types in Table 2 was calculated, showing that in all disciplines more than half of the scientists with a web presence linked to these resources. Moreover, some disciplines seem to link disproportionately much to one file type (see Table 4). The proportion of EHC scientists with at least one link from their personal website was larger overall than for online publication lists, which justifies analysing personal websites as well as online lists of publications.

Table 4. The proportion of EHC scientists with at least one outlink for each discipline and for the two file types, from either online lists of publication or personal websites, for which the crawler worked.

Discipline	EHC scientists with either a personal website or an online list of publications	Proportion of EHC scientists with at least one outlink to:		
		Non-HTML files	HTML files	Any file type
Engineering	171 (70%)	45.6%	35.1%	54.4%
Hard science	258 (73%)	51.6%	48.4%	68.2%
Health sciences	200 (46%)	18.5%	47.5%	55.0%
Life sciences	204 (49%)	26.5%	43.6%	52.5%
Social sciences	59 (77%)	66.1%	66.1%	89.8%

Using the data of Table 4, differences in proportions tests (with a Bonferroni correction for n=10) were used to seek evidence of differences between disciplines. Both health sciences and life sciences had lower proportions of non-HTML files than the other three areas (p<0.01 in all cases). Social sciences had a higher proportion of HTML files than engineering (p<0.001) and life sciences (p <0.05). For all file types combined, there was significant evidence that social sciences had a higher proportion than all other areas (p < 0.01), and hard sciences had a higher proportion than health sciences (p<0.05) life sciences (p<0.01), and engineering (p<0.05). Overall, then, social scientists link the most with engineering and hard sciences tending to link to non-HTML files more than life sciences and health sciences.

Although social sciences was the group with the largest proportion of scientists providing at least one outlink, engineering and hard sciences researchers created the most outlinks overall (see Appendices A and B). So, for example, taking into account the total outlinks found from online lists of publications, engineers created on average 86 links and half of them created 41 or more outlinks; and hard sciences researchers created on average 83 links and half of them created 32 or more. Life scientists created the fewest outlinks.

5.2 Outlinked file types

The first type of analysis counted links to non-HTML files. Of the 892 scientists with either a personal web site or an online list of publications, 332(37%) linked at least once to any of these file types, with the proportions varying among disciplines (see Table 4). PDF files were by far the most common file type, being linked to by 94% of researchers, while PostScript files and Gzip files were linked to by less than a fifth of the researchers (18% and 12% respectively), and the other file types (doc, docx or rtf; ppt or pptx; wmv, mp3) were linked to by less than 10% of the scientists. Note that previous studies (Aguillo *et al.*, 2007; Thelwall and Kousha, 2008) have shown that many scholarly presentations that were initially created in PowerPoint are provided afterwards on the Web as PDF files.

As can be seen in Table 5, hard sciences and engineering scientists linked to a greater variety of files types and PostScript and Gzip files were mainly linked to by them. Social sciences researchers also linked to a great variety of files types (all except PostScript files and Gzip files) while scientists from other disciplines (health sciences and life sciences) mainly linked to PDF files.

Table 5. Proportion of EHC scientists with at least one link to common file types for each discipline.

Discipline	EHC scientists with at least one outlink to non-html files	pdf	doc, docx, rdf	ps	gz	zip	ppt, pptx	wmv, mp3
Engineering	78 (45.6%)	99%	9%	26%	26%	3%	14%	5%
Hard science	133 (51.6%)	94%	5%	31%	13%	4%	10%	2%
Health sciences	37 (18.5%)	97%	11%	0%	0%	0%	3%	3%
Life sciences	51 (26.5%)	98%	4%	2%	4%	2%	6%	0%
Social sciences	33 (77%)	100%	15%	0%	0%	24%	9%	6%

Perhaps surprisingly, the number of .doc, .docx and .rtf linked files was higher from the lists of publications than from personal websites. This occurred because from online lists of publications almost all links to this format type were created by a single investigator, and the crawl of that researcher's personal website did not include her or his publication list. This is an example of how the design of a web site can affect a link analysis. The program used was capable of crawling the sites at different depths – controlling how deep the crawler goes into a site. We selected a depth of 1 to include the main page and the main additional pages for any researcher, such as a page of their publications, assuming that all the main pages would be linked from the home page. However, in this specific case the page of publications was not linked from the home page. We did not select a depth greater than 1 because the extra depth brings extra risk of the crawl straying into irrelevant pages hosted by the researcher.

5.3 Outlinked website types

The second type of analysis counted links to a variety of common types of target websites. Of the 892 scientists with either a personal web site or an online list of publications, 408 (46%) linked at least once to these websites (see Table 4). As shown in Figure 2, all types of websites were linked to by all disciplines, but with different proportions; scholarly databases were most linked to by EHC scientists. The largest proportion of health sciences researchers (36.5%) and life sciences researchers (29%) linked to scholarly databases, while many engineers (18%) linked to both scholarly databases and publishers' websites. Hard science researchers linked to these resources, but most of them (26%) mainly linked to OA repositories. Social sciences researchers linked to scholarly databases (34%), social sites (32%) and publishers' websites (25%).

A larger proportion of researchers linked to each of these target websites from online lists of publications than from websites, except for the social websites, which were unsurprisingly linked to by a larger proportion of researchers' personal websites.

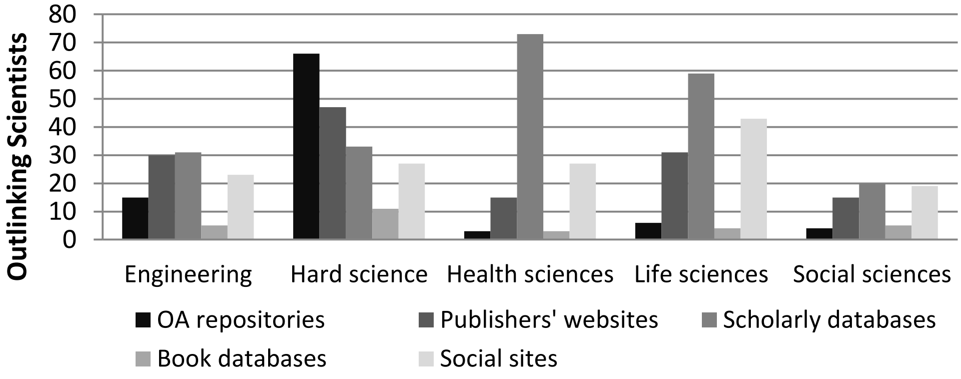


Figure 2: EHC scientists with at least one outlink to various types of scholarly website from either their personal website or their online list of publications.

The OA repositories/archives most linked to by EHC scientists were ArXiv (http://arxiv.org/), an open e-print archive with papers in physics, mathematics, non-linear sciences, computer science, quantitative biology, quantitative finance and statistics, and the SAO/NASA Astrophysics Data System (http://adsabs.harvard.edu/), a portal for researchers in astronomy and physics, operated by the Smithsonian Astrophysical Observatory (SAO) under a NASA grant (see Appendix C), suggesting that OA repositories were especially successful among researchers from hard sciences and engineering, except for chemistry researchers, since none of them linked to any OA repositories.

Most links pointing to ArXiv were created by Spanish and British researchers (52% of the outlinks from online lists of publications were created by Spanish researchers and 45% of the outlinks from personal websites were created by British researchers) and most links pointing to SAO/NASA were created by Israeli researchers (82% of the outlinks from online lists of publications and 61% from personal sites were created by them). ArXiv was originally named the LANL e-print archive because it was hosted by the Los Alamos National Laboratory (http://xxx.lanl.gov/) (McKiernan, 2000), so we also analyzed outlinks to this source, which was linked to by some EHC scientists. Other OA repositories linked to, although to a lesser extent, were MIMS EPrints (http://eprints.ma.man.ac.uk/), a repository for the papers, preprints and theses produced in the School of Mathematics at the University of Manchester and the HAL (Hyper Article en Ligne) repository (http://hal.archives-ouvertes.fr/) from the Centre pour la Communication Scientifique Directe (CCSD) at the Centre National de la Recherche Scientifique (CNRS) in France.

EHC scientists also linked to a wide range of publishers' websites (see Appendix D). Outlinks from online lists of publications mainly pointed to the American Physical Society (APS) website (http://www.aps.org/) and ScienceDirect (http://www.sciencedirect.com/), an electronic platform with full-text articles published in Elsevier journals. Outlinks from personal websites pointed to the Nature Publishing Group, Journal Storage (JSTOR), ScienceDirect and Wiley websites. Most outlinks (70%) pointing to the APS website were created by Austrian hard sciences researchers, while those pointing to ScienceDirect were mainly created by German hard sciences researchers (54% of the outlinks from online lists of publications) and French scientists (45% of the outlinks from personal websites). Most links pointing to the Nature Publishing Group (http://www.nature.com/) were created by life scientists (41% of outlinks from online lists of publications and 83% from personal websites were created by them). JSTOR (http://www.jstor.org/), a non-profit service that includes full-text content of academic journals, was especially linked to by social scientists (they created the 67% of the outlinks from online lists of publications and 92% of the outlinks from personal sites) and Wiley (http://wiley.com) was especially linked to by life scientists and hard sciences researchers (83% of the outlinks from online lists of publications and 68% from personal websites were created by them).

Regarding other scholarly databases (OA or non-OA), the vast majority of outlinks (87.5% from lists of publications to 80% from personal websites) pointed to the DOI (Digital Object Identifier) site. A DOI is a unique alphanumeric string assigned by the International DOI Foundation to identify content and to help provide a persistent link to its location on the Internet. The DOI can provide access to the full text of a publication, although the article may be accessible to subscribers or purchasers only. EHC scientists seem to be aware of the importance of the DOI, as one third of the outlinks extracted from their online lists of publications pointed to a DOI, although there were large differences between countries. For example, 87% of the outlinks created by Belgian scientists and 63% of outlinks created by British researchers pointed to DOIs, while links pointing to this source constituted less than 10% of the outlinks in many other countries. Another database linked to by EHC researchers was the National Library of Medicine at the National Institutes of Health.

In general, book databases or booksellers were rarely used, with very few links pointing to Amazon, Google Books and Bookstore. Social network sites were also rarely linked to by EHC researchers, with Facebook, Twitter and YouTube being the most common (see Appendix E). These links were especially created by British scientists. About 14% of the outlinks from personal websites linked to Wikipedia, although it was never linked to from online lists of publications. Normally, scientists that linked to Facebook also linked to Twitter and vice versa.

A result that appears to be inconsistent is that many of links pointing to the National Library of Medicine (NLM) at the National Institutes of Health were created by social sciences researchers.

However, this may be because some of these researchers belong to psychology (within social sciences) and social health (within both social and health sciences), but are included in social sciences in the database used to identify the highly cited researchers.

6. Discussion

The results above reveal new information about the link creating behaviour of EHC researchers based upon their online lists of publications and their institutional personal home pages, as gathered by a web crawler and as matching a pre-defined collection of popular academic websites and file types. The method makes some simplifying assumptions, such as a differentiation between links on the basis of whether the target is (probably) a HTML file type or not. This is a complex issue because outlinks to HTML files may point to metadata associated with publications when the full text is unavailable, although a recent study of research in OA repositories (Pérez *et al*, 2012) has shown that most pages only link to metadata records in OA repositories rather than full text documents. On the other hand, another study about researchers' personal websites (Barjak *et al*, 2007) showed that the full text was the most linked-to content. The current study was restricted to pages associated with EHC scientists within institutional websites; non-institutional personal websites were not taken into account. A limitation is that some scholars may have personal websites hosted elsewhere, such as in academia.edu, Facebook, an enhanced blog or a website with a personal domain name. The study is also limited by disciplinary biases inherited from the ISI data source, gender biases inherited from the use of citations as an indicator and a European bias due to the nature of the sample. This study found that only a third of the EHC scientists had an online list of publications, although more than half of them had created a lot of outlinks to the resources identified. This suggests that EHC scientists who use their personal homepages to disseminate information often use it extensively to provide access to their researches.

Both the mean and median number of outlinks from online lists of publications were higher than from personal websites. Thus scientists with an online list of publications seem not only concerned to announce their research outputs, but also to provide the (full text) access to them. In contrast, scientists that include their publications within their personal homepages, but do not have a specific web page for them could be more interested in publicity rather than concerned about access. Nevertheless, this may be influenced by institutional policies for some authors that have little control over their personal websites and some of these may host links to their research online but outside of their institutional website, or in another part of the institutional website, such as a research group site.

The proportion of scientists providing at least one outlink varied among disciplines, with the social sciences being the group with the largest proportion, and engineering and hard sciences researchers being the groups with the highest number of outlinks. This large proportion of social scientists may be a statistical artifact because of the small proportion of them in the sample or it could be because the social scientists are more elite because they form a smaller proportion of the total population of social scientists. A previous study claimed that "economists and computer scientists are more reliant than scientists from other disciplines on the WWW for obtaining and disseminating information" (Barjak, 2006, p. 1362), so another reason for the high social science results might be the need to increase the visibility of their outputs due to the poor coverage of traditional bibliographical databases (such as the ISI Web of Science) in these disciplines. Other work supports the finding regarding the high number of outlinks for engineering and hard sciences researchers: "Web sites in computer science, mathematics, and other physical science and engineering disciplines make more use of hyperlinks than do other scientific disciplines" (Barjak *et al*, 2007, p. 202).

Regardless of discipline, the most linked-to file type was PDF, with many more links than for the other file types. However, hard sciences and engineering researchers were also quite likely to use PostScript and Gzip files for disseminating their research. Previous studies (Goodrum *et al*., 2001) have already confirmed the widespread use of the PostScript format in computer science.

The most linked to type of target website was the scholarly database, especially the DOI website. Linking to a DOI ensures the permanent availability of a publication and almost a third of the outlinks from the EHC scientists' online lists of publications pointed to a DOI, suggesting that this group of researchers are becoming aware of its importance. Note that when the crawler used identified outlinks pointing to a DOI it did not resolve the DOI to identify which publisher (e.g., Springer, Wiley, ScienceDirect) gave access to the publications. Hence the outlink analysis cannot reveal which publishers gave most access to scientists' publications through their DOIs.

Researchers' behaviors seemed to differ between disciplines; health science and life science researchers mainly linked to scholarly databases while scientists from engineering, hard sciences and social sciences linked to a wider range of target websites. OA repositories were especially popular for EHC scientists from hard sciences and engineering. These disciplines or significant fields within them might therefore have an established pattern of using OA repositories as a channel for disseminating research, which explains the success of some repositories such as the SAO/NASA Astrophysics Data System, Arxiv or CiteSeerX. This success of OA repositories among hard sciences and engineering may be due to their use of grey literature, building upon pre-existing pre-print traditions (Pinfield, 2001; Goodrum *et al*., 2001). Since scientists in these disciplines also link to other web sources, EHC scientists (in these disciplines) could link to OA repositories to offer access to unpublished or unrevised works (preprints, working papers, etc.) and link to other sites, like publishers' websites or scholarly databases, to provide access to published versions; however more studies are needed to confirm this. This pattern cannot be applied to chemistry, however, since none of the 50 chemistry researchers linked to any OA repositories. This aligns with a previous finding (Björk *et al*., 2010) of low OA use by chemists, so traditional journal articles seems to continue to dominate communication in chemistry, as Lawal (2002) noticed a decade ago.

Publishers' websites were not used much by highly cited researchers, with values between 7.5% for health sciences researchers and 25% for social scientists. Our data show that EHC scientists from health sciences and life sciences preferred to use scholarly databases; however a study related to OA in the biomedical field (Matsubayashi *et al*., 2009) showed that of the 27% of OA articles published in 2005 in the biomedical field, the majority (72.1%) were provided by the journal publisher's website. Both methods have limitations, so more studies are needed to explain this difference. Social sites seem to be more important for life science and social scientists, whilst book databases or booksellers were rarely linked to by EHC scientists.

In this paper we did not analyze all links created by EHC scientists, but only those outlinks pointing to specific types of websites that were common hyperlink targets from researchers' websites (OA repositories, publishers' websites, scholarly databases, book databases or booksellers and social websites). However, in recent years much effort has been devoted to promoting institutional repositories and perhaps many scientists linked to them to provide an abstract of the full text of some publications. Thus, an interesting future research would be analyze all outlinks and compare the percentages of outlinks pointing to each of them. Besides comparing whether the links point to institutional repository or to the specific types of websites identified, it would also be interesting to analyze if the outlinks created by EHC scientists point to an abstract or full text of their publication as well as the type of publication (journal article, conference article, book, book chapter, technical report, working paper, etc.) and its version (preprint, postprint, final version of article).

7. Conclusions

The overall goal of this study was to seek substantial differences in the way that scientists use the web to publicise research so that future indicators can be sensitive to different but valid strategies for web research publicity. For instance, if direct links to PDF files from institutional personal websites were universally used by researchers to publicize and give access to their research then a logical indicator for online publicity for researchers would be a simple count of these PDF file links, but the reality is more complex than this, as the results of this study confirm.

It seems that EHC scientists are using the Web to point to online copies of their papers, since a high proportion of them with an institutional web presence (more than half in each all disciplines) created at least one outlink, and they had a high number of outlinks to different files types, although there are some disciplinary differences in how much they do this. Hence it seems that successful researchers have largely taken on board the message of the OA community (e.g., Harnad, 2005) and are taking advantage of the web to publicise their research.

The findings clearly demonstrate that successful scientists use a wide variety of strategies to publicize their research and that there are disciplinary differences in how this is done and even a variety of different strategies within disciplines (e.g., links to PDF files, links to repositories or publishers' web sites). Although disciplinary differences in internet use are to be expected (Nentwich, 2003), this confirms that any future web indicators for effective use of the web must take into account these major differences. As a result of the variety found, an important implication of the results of this study is that any reasonable web research publicity indicator would have to combine a variety of methods to identify ways for a scholar to point to their articles, including direct links to PDF files and indirect links to subject-specific archives, the DOI site and publishers' web sites and probably repositories too. The task of building an effective indicator will therefore be complex and will require gathering much information about the various different ways in which researchers point to their research online, including extensive lists of archives and repositories, so that links to these can be identified and differentiated from links to non-scholarly websites. For example, a simple indicator would be a count of the number of links from a researcher's web presence to rich files or any recognized scholarly website. Comparing this figure for academics from different countries and disciplines on a large scale would help to identify areas of good practice in research dissemination and areas where more online publishing advocacy is needed.

Since the advent of the Web, many studies have focused on the new informal communication channels used by academics, such as personal websites, repositories or social websites. There has also been great interest in OA initiatives to share scientific knowledge and the OA availability of research articles. During this period, many investigations have attempted to measure the online impact of web publishing and link analyses have also been used to assess different aspects of research communication, as discussed above, although these have investigated mainly interlinking between universities or departments. So far, no study has analyzed outlinking patterns, at an individual level, for identifying which web resources, such as OA repositories, databases, fee-based publishers’ sites and social network sites, are used by scholars to provide (full text) access to their publications. This study is a first attempt to fill this gap and the results should guide individual academics about how then could and perhaps should be using the web to publicize and disseminate their research in a manner appropriate for their own discipline.

Acknowledgements

This paper is supported by ACUMEN (Academic Careers Understood through Measurement and Norms) project, grant agreement number 266632, under the Seventh Framework Program of the European Union.

Appendices

Appendix A. Mean and median outlinks from online lists of publications, for EHC scientists with at least one outlink.

Discipline	Scientists with at least one outlink	Outlinks to non-HTML files		Outlinks to HTML files		Outlinks of any type	
		Mean	Median	Mean	Median	Mean	Median
Engineering	47	85	34	42	19	86	41
Hard sciences	73	72	32	53	6	83	32
Health sciences	44	20	3	71	17	76	19
Life sciences	36	42	17	27	14	45	22
Social sciences	15	41	29	50	2	60	28

Appendix B. Mean and median outlinks from personal websites, for EHC scientists with at least one outlink.

Discipline	Scientists with at least one outlink	Outlinks to non-HTML files		Outlinks to HTML files		Outlinks of any type	
		Mean	Median	Mean	Median	Mean	Median
Engineering	86	18	7	57	17	58	23
Hard sciences	162	15	3	51	11	48	10
Health sciences	97	58	8	17	2	54	7

Life sciences	96	22	4	15	2	24	4
Social sciences	47	36	3	47	15	61	16

Appendix C. The OA repositories/archives most linked to by EHC scientists.

OA repository	From online lists of publications	From personal websites
ArXiv	30.2%	23.9%
SAO/NASA ADS	23.3%	41.2%
Los Alamos National Lab.	19.0%	21.3%
Others	27.5%	13.6%
Total	100.0%	100.0%

Appendix D. The publishers' websites most linked to by EHC scientists.

Publishers’ website	From online lists of publications	From personal websites
APS Physics	33.4%	4.6%
ScienceDirect	17.9%	11.9%
Wiley	7.9%	10.1%
Nature Publishing Group	6.6%	14.0%
JSTOR	4.4%	12.1%
Others	29.8%	47.2%
Total	100.0%	100.0%

Appendix E. The social websites most linked to by EHC scientists.

Social website	From online lists of publications	From personal websites
Facebook	34.3%	26.8%
Twitter	31.4%	26.8%
Youtube	21.0%	20.4%
Linkedin	8.6%	7.5%
Wikipedia	0%	13.6%
Others	4.8%	4.8%
Total	100.0%	100.0%

References

Aguillo, I., Ortega, J.L., Prieto, J.A. and Granadino, B. (2007), “Indicadores Web de actividad científica formal e informal en Latinoamérica”, *Revista Española de Documentación Científica*, Vol. 30 No. 1, pp. 49-60.

ALLEA - ALL European Academies (2012), “Open Science for the 21st century. A declaration of ALL European Academies”, Rome, available at: http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/allea-declaration-1.pdf (accessed 12 April 2012).

Ajiferuke, I., and Wolfram, D. (2004), “Modelling the characteristics of web page outlinks”, *Scientometrics*, Vol. 59 No. 1, pp. 43-62.

Aksnes, D. W., Rorstad, K., Piro, F. and Sivertsen, G. (2011), “Are female researchers less cited? A large-scale study of Norwegian scientists”, *Journal of the American Society for Information Science and Technology*, Vol. 62 No 4, pp. 628-636.

Barjak, F. (2006), “The role of the Internet in informal scholarly communication”, *Journal of the American Society for Information Science and Technology*, Vol. 57 No 10, pp. 1350-1367.

Barjak, F., Li., X. and Thelwall, M. (2007), “Which factors explain the web impact of scientists’ personal homepages?”, *Journal of the American Society for Information Science and Technology*, Vol. 58 No 2, pp. 200-211.

Bar-Ilan, J. (2000), “The Web as an information source on informetrics? A content analysis”, *Journal of the American Society for Information Science*, Vol. 51 No 5, pp. 432-443.

Bar-Ilan, J. (2001), “Data collection on the Web for informetric purposes - a review and analysis”, *Scientometrics*, Vol. 50 No 1, pp. 7-32.

Bar-Ilan, J. (2004), “A microscopic link analysis of academic institutions within a country – The case of Israel”, *Scientometrics*, Vol. 59 No 3, pp. 391-403.

Basu, A. (2006), “Using ISI’s ‘Highly Cited Researchers’ to obtain a country level indicator of citation excellence”, *Scientometrics*, Vol. 68 No 3, pp. 361-375.

Batty, M. (2003a), “Citation geography: it’s about location”, *The Scientist*, Vol. 17 No 16, available at: http://jmichaelbatty.files.wordpress.com/2011/06/batty-scientist-2003.pdf (accessed 9 April 2012).

Batty, M. (2003b), “The geography of scientific citation”. *Environment and Planning A*, Vol. 35, No 5, pp. 761-765.

Björk, B-C., Welling P., Laakso, M., Majlender P., Hedlund T., and Gudnasson, G. (2010), “Open access to the scientific journal literature: Situation 2009”, *PLoS ONE*, Vol. 5 No 6, available at: http://www.plosone.org/article/info:doi/10.1371/journal.pone.0011273 (accessed 9 January 2013).

BOAI - Budapest Open Access Initiative (2002), available at: http://www.soros.org/openaccess/read (accessed 14 September 2012).

BOAI - Budapest Open Access Initiative (2012), “Ten years on from the Budapest Open Access Initiative: setting the default to open”, available at: http://www.soros.org/openaccess/boai-10-recommendations (accessed 14 September 2012).

Carmichael, F. and Charles, S. (2003), “The opportunity costs of informal care: does gender matter?”, *Journal of Health Economics*, Vol. 22 No 5, pp. 781-803.

Chen, C., Sun, K., Wu, G., Tang, Q., Qin, J., Chiu, K., and Liu, J. (2009), “The impact of internet resources on scholarly communication: a citation analysis”, *Scientometrics*, Vol. 81 No 2, pp. 459-474.

Chun, T. Y. (1999), “World Wide Web Robots: an overview”. *Online Information Review*, Vol. 23 No 3, pp. 135-142.

Chung, Y.M., Yu, S.Y., Kim, Y.K., and Kim, S.Y. (2009), “Characteristics and link structure of a national scholarly Web space: The case of South Korea”, *Scientometrics*, Vol. 80 No 3, pp. 595-612.

Danell, R. and Hjerm, M. (2013), “Career prospects for female university researchers have not improved”, *Scientometrics*, 2013, Vol. 94 No 3, pp. 999-1006.

García-Santiago, L., and Moya-Anegón, F. de (2009), “Using co-outlinks to mine heterogeneous networks”, *Scientometrics*, Vol. 79 No 3, pp. 681-702.

Gargouri, Y., Lariviere, V., Gingras, Y., Carr, L, and Harnad, S. (2012), “Green and Gold Open Access Percentages and Growth, by Discipline”, in *17th International Conference on Science and Technology Indicators (STI)*, 5-8 September 2012, Montreal, Canada.

Giles, L., Sun, Y., and Councill, I.G. (2012), “Measuring the web crawler ethics”, in *Conference World Wide Web Conference Series*, Raleigh, United States: ACM, pp. 1101-1102, available at: https://clgiles.ist.psu.edu/pubs/WWW2010-web-crawler-ethics.pdf (accessed 10 July 2012).

Goodrum, A.A., McCain, K. W., Lawrence, S., and Giles, L. (2001), “Scholarly publishing in the Internet age: a citation analysis of computer science literature”, *Information Processing & Management*, Vol. 37, No 5, pp. 661-675.

Harnad, S. (2005), “The Implementation of the Berlin Declaration on Open Access”, *D-lib Magazine*, Vol. 11 No 3, available at: http://www.dlib.org/dlib/march05/harnad/03harnad.html (accessed 7 August 2012).

Harnad, S. (2011), “Open Access to Research: Changing Researcher Behavior Through University and Funder Mandates”, *JEDEM Journal of Democracy and Open Government*, Vol. 3 No 1, pp. 33-41.

Holmberg, K. (2011), “Discovering shared interests through co-outlinking in a municipal web space”, in *Proceedings of the 13th ISSI conference*, Durban, Republic of South Africa.

Ingwersen, P. (1998), “The calculation of Web Impact Factors”, *Journal of Documentation*, Vol. 54 No 2, pp. 236-243.

Ioannidis, J.P.A. (2010), “Is there a glass ceiling for highly cited scientists at the top of research universities?”, *The FASEB Journal*, Vol. 24 No 12, pp. 4635-4638.

Jalal, S.K., Biswas, S.C., and Mukhopadhyay, P. (2010), “Web impact factor and link analysis of selected Indian universities”. *Annals of Library and Information Studies*, Vol. 57, pp. 109-121.

Jensen, C., Sarkar, C., Jensen, C., and Potts, C. (2007), "Tracking Website Data-Collection and Privacy Practices with the iWatch Web Crawler", in *Proceedings of Symposium On Usable Privacy and Security – SOUP*, Pittsburgh, pp. 29-40.

Koster, M. (1993), "Guidelines for robot writers", available at: http://www.robotstxt.org/guidelines.html (accessed 10 July 2012).

Kousha, K. (2009), "Characteristics of open access scholarly publishing", *ASLIB Proceedings*, Vol. 61 No 4, pp. 394-406.

Kousha, K., and Thelwall, M. (2007), "The Web impact of open access social science research", *Library and Information Science Research*, Vol. 29, pp. 495-507.

Kretschmer, H., Pudovkin, A. and Stegmann, J. (2012), "Research evaluation. Part II: gender effects of evaluation: are men more productive and more cited than women?", *Scientometrics*, Vol. 93 No 1, pp. 17-30.

Lawal, I. (2002), "Scholarly Communication: The Use and Non-Use of E-Print Archives for the Dissemination of Scientific Information", available at: http://www.istl.org/02-fall/article3.html (accessed 20 February 2013).

Leeuwen, T.N. van, Moed, H.F., Tijssen, R.J.W., Visser, M.S., and Raan, A.F.J., van (2001), "Language biases in the coverage of the Science Citation Index and its consequences for international comparisons of national research performance", *Scientometrics*, Vol. 51 No 1, pp. 335-346.

Matsubayashi, M., Kurata, K., Sakai, Y., Morioka, T., Kato, S., Mine, S. and Ueda, S. (2009), "Status of open access in the biomedical field in 2005", *Journal of the Medical Library Association*, Vol. 96, No 1, pp. 4-11.

Mauleón, E. and Bordons M. (2006), "Productivity, impact and publication habits by gender in the area of Material Science", *Scientometrics*, Vol. 66 No 1, pp. 199-218.

McAuley, L., Pham, B., Tugwell, P., and Moher, D. (2000). "Does the inclusion of grey literature influence estimates of intervention effectiveness reported in meta-analyses?", *Lancet*, Vol. 356 No 9237, pp. 1228-1231.

McKiernan, G. (2000), "arXiv.org: The Los Alamos National Laboratory E-Print Server", *The International Journal on Grey Literature*, Vol. 1 No 3, pp. 127–138.

Mas-Bleda, A., and Aguillo, I.F. (in press), "Can a personal website be useful as an information source to assess individual scientists? The case of European highly cited researchers", *Scientometrics*.

Nentwich, M., (2003), *Cyberscience. Research in the Age of the Internet*. Austrian Academy of Sciences Press, Vienna.

Ortega, J. L., and Aguillo, I. F. (2007), "Interdisciplinary relationships in the Spanish academic web space: A Webometric study through networks visualization", *Cybermetrics*, Vol. 11, available at: http://cybermetrics.cindoc.csic.es/articles/v11i1p4.html (accessed 22 March 2012).

Parker, J.N., Lortie, C. and Allesina, S. (2010), "Characterizing a scientific elite: the social characteristics of the most highly cited scientists in environmental science and ecology", *Scientometrics*, Vol. 85 No 1, pp. 129-143.

Payne, N., and Thelwall, M. (2004), "A Statistical Analysis of UK Academic Web", *Cybermetrics*, Vol. 8, available at: http://cybermetrics.cindoc.csic.es/articles/v8i1p2.html (accessed 22 March 2012).

Payne, N., and Thelwall, M. (2007), "A longitudinal study of academic webs: Growth and stabilization", *Scientometrics*, Vol. 71 No 3, pp. 523-539.

Payne, N., and Thelwall, M. (2009), "A longitudinal analysis of Alternative Document Models", *Aslib Proceedings*, Vol. 16 No 1, pp. 101-116.

Pérez Álvarez, S., Álvarez, F.P., and Aguillo, I. (2012), "EU FP7 in Open Access Repositories", in *Proceedings of 17th International Conference on Science and Technology Indicators*, Science-Metrix and OST, Montréal, pp. 58-70.

Pinfield, S. (2001), "How Do How Do Physicists Use an E-Print Archive? Implications for Institutional E-Print Services", *D-Lib Magazine*, Vol 7. No 12, available at: http://www.dlib.org/dlib/december01/pinfield/12pinfield.html (accessed 17 February 2013).

Pitzek, S. (2002), "Impact of Online-availability of Science Literature", available at: http://www.vmars.tuwien.ac.at/courses/proseminar/doc/paperserver.pdf (accessed 13 July 2011).

Prpic, K. (2002), "Gender and productivity differentials in science", *Scientometrics*, Vol. 55 No 1, pp. 27-58.

Pudovkin, A., Kretschmer, H., Stegmann and Garfield, E. (2012), "Research evaluation. Part I: productivity and citedness of a German medical research institution", *Scientometrics*, Vol. 93 No 1, pp. 3-16.

Qiu, J., Chen, J., and Wang, Z. (2004), "An analysis of backlink counts and Web Impact factors for Chinese university websites", *Scientometrics*, Vol. 60 No 3, pp. 463-473.

Reskin, B.F., and Bielby, D. D. (2005), "A Sociological Perspective on Gender and Career Outcomes", *Journal of Economic Perspectives*, Vol. 19 No 1, pp. 71-86.

Rodríguez Gairín, J. M. (1997), "Valorando el impacto de la información en Internet: AltaVista, el "Citation Index" de la Red", *Revista Española de Documentación Científica*, Vol. 20 No 2, pp. 175-181.

Rossini, C. (2007), "The open access movement: opportunities and challenges for developing countries. Let them live in interesting times. Diplo Foundation Internet Governance Program", available at: http://campus.diplomacy.edu/env/scripts/Pool/GetBin.asp?IDPool=3737 (accessed 10 July 2012).

Seeber, M., Lepori, B., Lomi, M., Aguillo, I. and Barberio V. (2012), "Factors affecting web links between European higher education institutions", *Journal of Informetrics*, Vol. 6 No 3, pp. 435-447.

Seiler, C. and Wohlrabe, K. (2011), "Ranking Economists and Economic Institutions Using RePEc: Some Remarks", working paper [n. 16], Ifo Institute for Economic Research, University of Munich, available at: http://www.cesifo-group.de/portal/pls/portal/docs/1/1201260.PDF (accessed 13 February 2013).

Suber, P. (2004), "Creating an Intellectual Commons through Open Access", available at: http://dlc.dlib.indiana.edu/dlc/bitstream/handle/10535/4445/Suber_Creating_041004.pdf (accessed 14 September 2012).

Sun, Y. (2008), "A comprehensive study of the regulation and behavior of web crawler", available at: http://books.google.com (accessed 1o July 2012).

Swan, A. (2007), "Open Access and the Progress of Science", *American Scientist*, Vol. 95, pp. 198-200.

Tang, R., and Thelwall, M. (2003), "Disciplinary differences in US academic departmental Web site interlinking", *Library and Information Science Research*, Vol. 25 No 4, pp. 437-458.

Thelwall, M (2001), "A web crawler design for data mining", *Journal of Information Science*, Vol. 27 No 5, pp. 319-325.

Thelwall, M. (2002a), "A Comparison of Sources of Links for Academic Web Impact Factor Calculations", *Journal of Documentation*, Vol. 58 No 1, pp. 60-72.

Thelwall, M. (2002b), "Conceptualizing Documentation on the Web: An Evaluation of Different Heuristic-based Models for Counting Links between University Web Sites", *Journal of the American Society for Information Science and Technology*, Vol. 53 No 12, pp. 995-1005.

Thelwall, M. (2002c), "Evidence for the existence of geographic trends in university Web site interlinking", *Journal of Documentation*, Vol. 58 No 5, pp. 563-574.

Thelwall, M. (2006), "Interpreting Social Science Link analysis Research: A theoretical framework", *Journal of the American Society for Information Science and Technology*, Vol. 57 No 1, pp. 60-68.

Thelwall, M., and Kousha, K. (2008), "Online presentations as a Source of Scientific Impact? An analysis of PowerPoint files citing academic journals", *Journal of the American Society for Information Science and Technology*, Vol. 59 No 5, pp. 805-815.

Thelwall, M., and Sud, P. (2011), "A comparison of methods for collecting web citation data for academic organizations", *Journal of the American Society for Information Science and Technology*, Vol. 62 No 8, 1488-1497.

Thelwall, M., and Sud, P. (2012), "Webometric Research with the Bing Search API 2.0", *Journal of Informetrics*, Vol. 6 No 1, pp. 44-52.

Thelwall, M., Sud, P., and Wilkinson, D. (2012), "Link and Co-link Network Diagrams with URL citations or Title Mentions", *Journal of the American Society for Information Science and Technology*, Vol. 63 No 4, pp. 805-816.

Thelwall, M., and Stuart, D. (2006), “Web crawling ethic revisited: cost, privacy, and denial of service”, *Journal of the American Society for Information Science and Technology*, Vol. 57 No 13, pp. 1771-1779.

Thelwall, M., Tang, R., and Price, L. (2003), “Linguistic patterns of academic Web use in Western Europe”, *Scientometrics*, Vol. 56 No 3, pp. 417-432.

Thelwall, M., Vaughan, L, and Björneborn, L. (2005), “Webometrics”, *Annual Review of Information Science and Technology*, Vol. 39 No 1, pp. 81-135.

Thomas, O., and Willet, P. (2000), “Webometric analysis of departments of Librarianship and Information Science”, *Journal of Information Science*, Vol. 26 No 6, pp. 421-428.

Torres-Salinas, D., Muñoz-Muñoz, A. M. and Jiménez-Contreras, E. (2011), “Análisis bibliométrico de la situación de las mujeres investigadoras de Ciencias Sociales y Jurídicas en España”, *Revista Española de Documentación Científica*, Vol. 34 No 1, pp. 11-28.

Vaughan, L., and Thelwall, M. (2005), “A modeling approach to uncover hyperlink patterns: The case of Canadian universities”, *Information Processing and Management*, Vol. 41 No 2, pp. 447-359.

Vaughan, L., and Wu, G. (2004), “Links to commercial websites as a source of business information”, *Scientometrics*, Vol. 60 No 3, pp. 487-496.

Wilkinson, D., Harries, G., Thelwall, M., and Price, L. (2003), “Motivations for academic web site interlinking: evidence for the Web as a novel source of information on informal scholarly communication”, *Journal of Information Science*, Vol. 29 No 1, pp. 49-56.

Yang, B. and Qin, J. (2008), “Data Collection System for Link Analysis”, in *Third International Conference on Digital Information Management*, London, pp. 247-252, available at: http://ir.las.ac.cn/bitstream/12502/3015/1/253.pdf (accessed 20 May 2012).