Indicators for a Webometric Ranking of Open Access Repositories

Isidro F. Aguillo[1], José L. Ortega[2], Mario Fernández[1], Ana M. Utrilla[1]

[1]Cybermetrics Lab. CCHS - CSIC. Albasanz, 26-28. 28037 Madrid. Spain.
E-mail: {isidro.aguillo, mario.fernandez,ana.utrilla}@cchs.csic.es¸

[2] Scientific Programming Division. VICYT –CSIC. Serrano, 113 28006 Madrid Spain.
E-mail: jortega@orgc.csic.es

Abstract

The Ranking web of World Repositories (http://repositories.webometrics.info) is introduced. The objective is to promote Open Access Initiatives supporting the use of repositories for scientific evaluation purposes. A set of metrics based on web presence, impact and usage is discussed. The Ranking is built on indicators obtained from web search engines following a model close to the Impact Factor one. The activity accounts for a 50% of the index, including number of pages, pdf files and items in Google Scholar database, while the visibility takes into account the external inlinks received by the repository (the other 50%). The Ranking provides the Top 300 repositories from a total of 592 worldwide, with a strong presence of US, German and British institutional repositories and the leadership of the large subject repositories. Results suggest the need to take into consideration other file formats and the usage information, an option is not feasible today.

Keywords: repositories, open access, webometrics, indicators, visibility, usage, ranking

1 Introduction and objectives

The Web, digital libraries and repositories are changing the way scientific knowledge is distributed and accessed, providing new opportunities for a more exhaustive and balanced coverage of the scholar literature. Open access journals and large scale deposit of research outputs in repositories will not only favourably affect paper’s visibility (Antelman, 2004; Harnad & Brody, 2004; Kurtz et al., 2005; Hajjem, Harnad & Gingras, 2005; Eysenbach, 2006; Moed, 2007; Norris, Oppenheim & Rowland, 2008) but it opens research evaluation to a more user centred methodologies.

Traditionally usage metrics, linked to journal circulation measures, have not been as successful in bibliometrics or scientometrics as citation analysis, mostly due to the quality of the available data. In the electronic arena, usage figures are larger, composition richer, more discriminative but also far more inclusive (Organ, 2006; Mayr, 2006; Xia & Sun, 2006; Westell, 2006; Scholtz & Dobratz, 2006; McDonald, 2007; Zuccala et al., 2008; Kim & Kim, 2008).

Unfortunately, there are still no standards for usage indicators so although many repositories publish extensive reports about visits, visitors and downloads, the data cannot be compared and combined as concepts measured are not the same.

Current efforts for developing citation and usage based indicators are on the way (Citebase: Brody,2003; CiteSeer: Lawrence, Giles & Bollacker, 1999; RepEc Ideas: Zimmermann, 2007)., but in the meantime a first generation of measures of Open access repositories is needed. At least two reasons for such a move: first to reinforce the OA initiatives, promoting document deposit by researchers and second to provide empirical data about this new scholarly communication system for further and comparative analysis.

Our objective is to introduce the Ranking web of Repositories (http://www.webometrics.info/top200_rep.asp), the indicators developed for measuring their activity and visibility and the results obtained, with comments on current and future developments.

2 Methodology

Since mid nineties Cybermetrics/Webometrics is helping to describe in a quantitative way the scholarly communication processes, focusing not on traditional citation databases (WoK, Scopus) but on the information available in the publicly accessible Web. This includes all the contents available through crawlers or search engines robots (Visible Web), but also those published in the so called Invisible, Deep or Hidden Web (Wouters et al., 2006), usually databases or other resources that need a search gateway to recover them (no longer invisible as Google and other engines really crawl and index their contents). This includes web academic databases, scientific repositories and electronic journals (Figure 1).

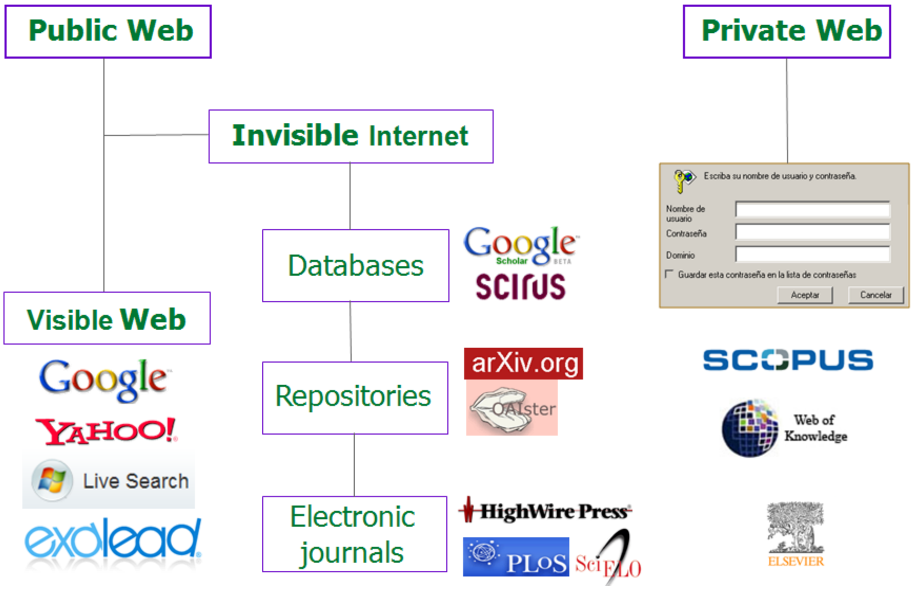


Figure 1. Classification of the main web sources for cybermetric analysis

Cybermetrics Lab has developed in the last years a series of indicators for measuring Web activity and visibility of academic and research institutions (Aguillo et al., 2005, 2006). Using search engines as intermediaries, it is possible to extract quantitative information about the web contents published by these institutions, including both formal papers but also informal material related to other missions such teaching, divulgation or technological transfer (Barjak, 2006).

Originally aimed to promote the Open access initiatives (OAI) the Ranking web of World Universities (Aguillo et al., 2006, 2008) has been proved as a useful tool also for institutions evaluation. Unfortunately it is not easy to split the components involved in the rank of a university, as it is an overall measurement combining different aspects and activities. Due to this but using a similar methodological approach, the Cybermetrics Lab decided to focus on the repositories themselves for describing and comparing the success of OAI in the academic sector.

Classification	*Units*
By provider	Personal homepages
	Institutional repositories
	Subject repositories
	Portal of e-journals
By content	Metadata (no full text)
	Preprints/postprints
	Thesis/MS Thesis

	Formal & Informal material
	Digitised Archives
Metarepositories	Directories
	Harversters

Table 1. Classification of the main repositories or repository related resources. In bold those categories covered in the Ranking Web of Repositories.

A first step was to classify the resources in order to make a homogeneous database of repositories. Due to feasibility reasons all the personal pages were excluded, although even considering the large numbers involved, an analysis of files in formats like pdf, doc or ps could be done for closed lists: departments, institutions, selected topics (Barjak, Li & Thelwall, 2007). The analysis of electronic journals is progressing through more traditional bibliometric and citation analysis, so individual journals and portals were also excluded.

Using OpenDOAR (www.opendoar.org) and ROAR (roar.eprints.org) directories of repositories, among other sources, a candidate list of repositories was assembled. After checking for thematic and institutional repositories, cleaning those whose primary contents does not consist of scientific papers, and excluding the mentioned portals of journals and harvesters, the final list includes only those with an autonomous domain or subdomain.

For each entry, the following statistics were recovered from search engines:

- **Size**, measured from the number of pages from Google, Live Search, Yahoo Search and Exalead. Several alternative measures were discarded such as the size in number of bytes because it is strongly linked to the size of graphics (a single large true colour graph could have a weight in order of several hundred bytes, more than dozens of only text pages).
- **Pdf files**, number of documents in pdf format from Google and Yahoo. There are other formats that can be used in the repositories, mainly Word (doc and rtf) and html and similar ones, but only a few repositories use them as unique or duplicate documents and in fact they are already counted in the previous indicator.
- **Scholar**, total number of entries in Google Scholar (still in beta version, with a the largest coverage of all the remaining public web academic databases, but very noisy), and
- **Visibility,** accounting number of external inlinks extracted from Yahoo and Exalead (Google only provides link numbers per page, not for full domains or subdomains). PageRank has been considered as a measure of link visibility but the figures offered by the Google Toolbar are integers from 0 to 10 in a logarithmic scale. Most of the values are 5, 6 or 7 and it is very difficult to climb to higher values that make this indicator unusable for ranking purposes.

The numbers were log-normalised (adding one unit to avoid log(0), max and min values were excluded (median, mean otherwise) and then converted into ordinals: first (highest value) is one, second is two and so on.

The four ranks were aggregated to obtain a unique value describing the overall presence and impact of the repository in the web. To better reflect the contribution of each variable a weight was assigned according to a subjective model. These values are "*a priori*" weights because they are decided prior to collection of data although minor adjustments could be done after examination of results.

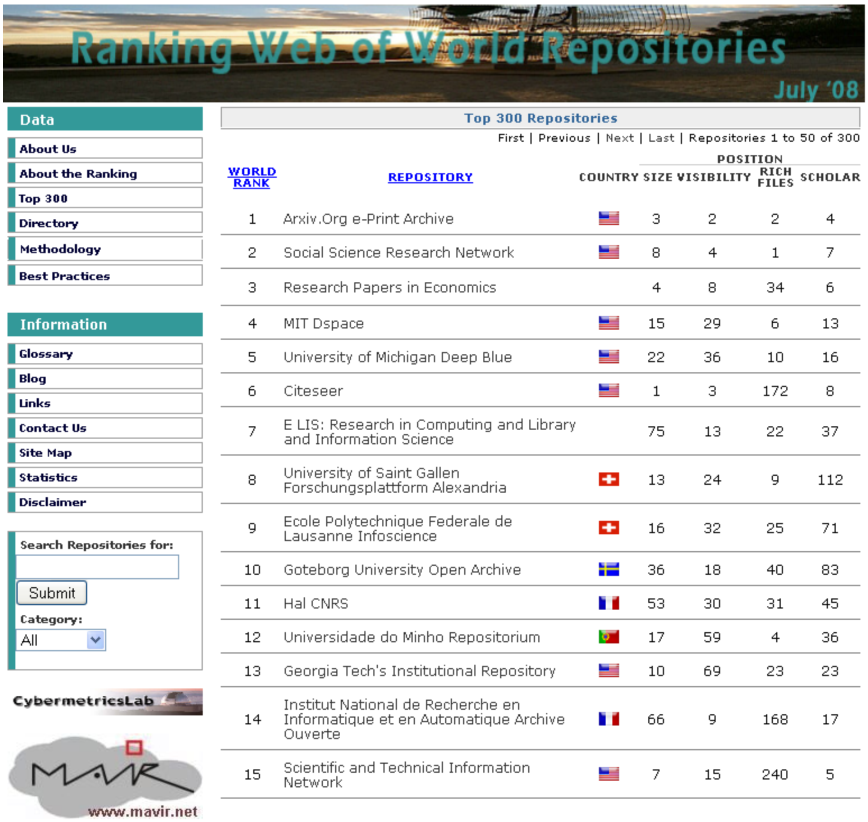
The model adopted is based in the Impact Factor indicator, that proposed a ratio 1:1 between the activity (50%) and impact (50%). Transferring the model from bibliometrics to webometrics means that number of publications is now number of web pages, while the number of citations is now the number of external inlinks. In order to reinforce the weight of more formal academic material, both the documents in pdf format and the items recovered from Google Scholar database are taking into account. To choose only the pdf files instead of a wider representation of Rich files (doc, rtf, ppt, ps and others) is for avoiding counting documents twice if they are available in two different formats. As the number of total files already count these other types, this indicator gives a closer approximation to actual number of deposited documents, excluding descriptive and navigation pages. In Table 2 the three components of activity: size (20%), pdf files (15%) and Scholar (15%) preserve the combined weight of 50% assigned in the model. The distribution of weights derived from an informed guess, but perhaps it should be reconsidered for future editions

Table 2. Weights applied to each indicator for building the Ranking

Variables	*Weight (percentage)*
Size (number of pages)	20%
PDF files	15%
Scholar	15%
Visibility (external inlinks)	50 %

A first beta version was published in February 2008. After receiving feedback regarding repositories missed in the first draft list and the variables involved, a second beta version were published during May 2008. The final version, corresponding to July 2008, is available from the ranking's webportal (http.//repositories.webometrics.info/), and it is the source for the current analysis. Further editions will be published two times (January and July) per year.

The main change is the extra weight given to recent records in Scholar database. The value for this indicator is now the mean between normalised total number of items and those published in the 2001-2008 period.



Top 300 Repositories

First | Previous | Next | Last | Repositories 1 to 50 of 300

WORLD RANK	REPOSITORY	COUNTRY	POSITION SIZE	VISIBILITY	RICH FILES	SCHOLAR
1	Arxiv.Org e-Print Archive	🇺🇸	3	2	2	4
2	Social Science Research Network	🇺🇸	8	4	1	7
3	Research Papers in Economics		4	8	34	6
4	MIT Dspace	🇺🇸	15	29	6	13
5	University of Michigan Deep Blue	🇺🇸	22	36	10	16
6	Citeseer	🇺🇸	1	3	172	8
7	E LIS: Research in Computing and Library and Information Science		75	13	22	37
8	University of Saint Gallen Forschungsplattform Alexandria	🇨🇭	13	24	9	112
9	Ecole Polytechnique Federale de Lausanne Infoscience	🇨🇭	16	32	25	71
10	Goteborg University Open Archive	🇸🇪	36	18	40	83
11	Hal CNRS	🇫🇷	53	30	31	45
12	Universidade do Minho Repositorium	🇵🇹	17	59	4	36
13	Georgia Tech's Institutional Repository	🇺🇸	10	69	23	23
14	Institut National de Recherche en Informatique et en Automatique Archive Ouverte	🇫🇷	66	9	168	17
15	Scientific and Technical Information Network	🇺🇸	7	15	240	5

Figure 2. Ranking Web of World Repositories (http://repositories.webometrics.info/)

2 Results

The July 2008 Ranking consists of a Directory of 592 repositories, from which the Top 300 are ranked. The Table 3 shows the distribution by country (53 countries plus one international category are represented), with special attention to the leaders

Table 3. Country distribution of the Repositories

Country	*Top 50*	*Top 100*	*Top 200*	***Total***
USA	17	39	60	**162**
Germany	6	11	23	**43**
United Kingdom	4	5	16	**78**
Switzerland	3	3	5	**6**
International	3	3	4	**4**
France	2	4	7	**16**
Australia	2	3	12	**27**

Japan	2	3	8	**34**
Spain	2	3	4	**14**
Canada	1	5	9	**22**
Netherlands	1	4	7	**15**
Sweden	1	4	7	**14**
India	1	2	4	**15**
Finland	1	2	3	**7**
Denmark	1	1	3	**6**
Portugal	1	1	2	**6**
Norway	1	1	1	**2**
Venezuela	1	1	1	**2**
Brazil		1	3	**8**
Belgium		1	2	**9**
Others		3	19	**102**

The three more important countries, according to the number of repositories ranked in the top positions, are the USA (30% of the Top 200 are US repositories), Germany (12%) and United Kingdom (8%), with other European countries, Japan and Canada also well represented.

Table 4 Top 20 repositories: Numbers under S (size), V (visibility), R (rich files), Sc (Google Scholar) are ordinals (1 is first, 2 is second, 3 is third, and so on)

Rank	*Repository*	*S*	*V*	*R*	*Sc*
1	Arxiv.Org e-Print Archive	3	2	2	4
2	Social Science Research Network	8	4	1	7
3	Research Papers in Economics	4	8	34	6
4	MIT Dspace	15	29	6	13
5	University of Michigan Deep Blue	22	36	10	16
6	Citeseer	1	3	172	8
7	E LIS: Research in Computing and Library and Information Science	75	13	22	37
8	University of Saint Gallen Forschungsplattform Alexandria	13	24	9	112
9	Ecole Polytechnique Federale de Lausanne Infoscience	16	32	25	71
10	Goteborg University Open Archive	36	18	40	83
11	Hal CNRS	53	30	31	45
12	Universidade do Minho Repositorium	17	59	4	36
13	Georgia Tech's Institutional Repository	10	69	23	23
14	Institut National de Recherche en Informatique et en Automatique Archive Ouverte	66	9	168	17
15	Scientific and Technical Information Network	7	15	240	5

16	National Library of Finland Dspace Services	43	53	65	21
17	University of Oregon Scholars' Bank	19	67	33	54
18	Humboldt Universitat zu Berlin Publikationsserver	14	65	37	66
19	Oregon State University Scholarsarchive	120	31	29	48
20	Munich Personal Repec Archive	59	62	11	46

Table 4 shows the main indicators of the twenty top ranked repositories (numbers are ordinals, with the lowest values representing the top positions). Most of the first ones are thematic, that usually are older (Arxiv since 1991), larger and prestigious at least in some disciplines. However the institutional ones are far more numerous and probably due to the copyright restrictions (institutions are the holders of the rights of their authors) will be those to persist.

Some comments on the results shown:

- CiteSeer is now changing to a new interface CiteSeerX that although still in beta is already ranked 121th worldwide
- HAL, the French Open archives system maintain not only a central repository (archives-overtes.fr) but also institutional ones like the INRIA subset that in fact receives more links than the parent one (9th vs 30th position in visibility).
- The Smithsonian/NASA Astrophysics Data System does not appear in the Top positions because they have two subdomains for the same repository (adsabs.harvard.edu and adswww.harvard.edu) that is a bad naming practice and penalise its web visibility.

In general the consistency among the web indicators are low, as leaders according to one indicator are not so strong in the other ones. A Spearman's rho test (Table 5) shows that the correlation between size, pdf files and Scholar items is significant (although it is very low).

Table 5. Correlation between web variables used in the Ranking Web (**significant at the 0.01 level (2-tailed))

Spearman's rho (n=200)	Correlation Coefficient			
	Google	PDFs	Scholar	Sch2001-08
Google	1	.198**	.368**	.308**
PDFs		1	.353**	.277**
Scholar			1	.778**
Sch2001-08				1

There are several possible explanations for this behaviour:

- Probably pdf is not ever the preferred format. Other formats are locally important like ps (arxiv, citeSeer). Although not included in the analysis, the Open Journal System (http://pkp.sfu.ca/?q=ojs) journals are providing papers in Adobe Acrobat format without the pdf suffix.
- Scholar (still in beta version) data are including references and other non full text items. Probably pre-2001 items are very important in repositories involved in large digitizing projects.
- Many institutions have several independent repositories, one for dissertations, another for papers and even other one for reports. In the extreme Caltech have 19 different entries in the Directory.
- Invisibility, the inability of search engines crawlers for collect data due to barriers in the design of the web databases, is probably a serious problem.

3 Discussion

Among the new aims of the OAI is not only to motivate institutions to develop scientific papers repositories but also to provide statistical information about these contents (Harnad, 2007, Armbruster, 2008). Ideally three groups of indicators should be deployed: activity related (documents deposited by author, institution, subject), usage related (visits, visitors, downloads), and visibility related (citations and web links). Today there is no universal source for many of these indicators, with some interesting initiatives on the way (http://projectcounter.org/; http://www.mesur.org/).

The usage data is an important indicator for the future, representing an innovation as its bibliometric counterpart (journal circulation) has not been frequently used. The diversity of statistics available is very large (visits, visitors, downloads, referrers, referrals …) but there is no open reports available for most of the repositories and even when the data is published the lack of standards avoid the comparative analysis.

The Ranking Web has been able to combine web data for producing a list of institutional and thematic repositories. The advantages of these preliminary efforts are evident as organizations are being classified according to their commitment to the Open access Initiatives. Current success of discipline focused repositories is not an indicator of future developments as more and more institutions are developing deposit mandates mostly to populate local repositories. This movement is especially interesting as deriving indicators for evaluation purposes for institutions could be easier to monitor and to compare with other bibliometric sources and previous analysis enhancing the value of the web data.

Even if the Ranking Web is providing a solid basis for such analysis, more indicators, especially those related to usage and citation, are badly needed:

After the demise of Live Academic the web bibliographic databases available for global analyses are few and the only one with citation data is Google Scholar. Scholar could play an important role (Kousha & Thelwall, 2006; Kousha & Thelwall, 2007; Meho & Yang, 2007), but a non beta version is long awaited and tags based searching is mandatory for obtaining clean and detailed results. Surprisingly the webmasters of many repositories are not paying enough attention to facilitate the citation of individual papers, as the URLs are very long, uninformative and without adequate file format suffix. If this situation is not resolved, then the number of citations does not increase adequately for performing further statistical analysis.

Regarding usage, the current generation of free tools represented by Google Analytics (http://www.google.com/analytics) have an important added value as they can represent a universal standard, but unfortunately it is not offering yet academic oriented capabilities. The usage information provided by individual repositories is very scant, not standardised, and easy to manipulate.

From a practical point of view, the Ranking is targeting institutional mandates in order not only to reinforce its use but suggesting a statistics rich environment. It is also important that new repositories take into consideration the indexing and positioning in commercial search engines like Google, Yahoo or Live: Suggestions here include the use of independent domains or subdomains, avoid the barriers to robots and adding semantic value to metadada.

The results show that more research should be done regarding the rich file formats, as the use of solely pdf files is not providing a complete picture. Some personal communication propose not take them into account but this is against the objective of measuring the contents. Another alternative that will be tested in January 2009 edition will be to combine all the main filetypes. Future editions will consider bibliographic citations and usage data when these indicators become available.

References

Aguillo, I.F., Granadino, B., Ortega, J.L., Prieto, J.A. (2005). What the Internet says about science. The Scientist, 19 (14): 10

Aguillo, I.F.; Granadino, B.; Ortega, J.L.; Prieto, J.A. (2006). Scientific research activity and communication measured with cybermetric indicators. Journal of the American Society of Information Science and Technology, 57(10): 1296 – 1302

Aguillo, I.F., Ortega, J.L., Fernández, M. (2008). Webometric ranking of world universities: Introduction, methodology, and future developments. Higher Education in Europe, 33(2-3): 233-244

Antelman, K. (2004). Do Open-Access Articles Have a Greater Research Impact?. College and Research Libraries, 65(5):pp. 372-382.

Armbruster, C. (2008). Access, Usage and Citation Metrics: What function for Digital Libraries and Repositories in Research evaluation. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1088453

Barjak, F. (2006). The role of the Internet in informal scholarly communication: Research Articles. Journal of the American Society for Information Science and Technology, 57(10):1350-1367.

Barjak, F., Li, X.; Thelwall, M. (2007). Which factors explain the Web impact of scientists' personal homepages?. Journal of the American Society for Information Science and Technology, 58(2): 200-211.

Brody, T. (2003) Citebase Search:Autonomous Citation Database for e-Print Archives. In:Third international technical workshop and conference of the project SINN, 17-19 September 2003, Oldenburg, Germany. http://eprints.ecs.soton.ac.uk/10677/1/brody_sinn03_paper.pdf

Eysenbach, G. (2006). Citation Advantage of Open Access Articles. PLOS Biology, 4(5):e157.

Hajjem, Ch.; Harnad, S. & Gingras, Y. (2005). Ten-Year Cross-Disciplinary Comparison of the Growth of Open Access and how it Increase Research Citation Impact. IEEE Data Engineering Bulletin, 28(4):39-47. http://sites.computer.org/debull/A05dec/hajjem.pdf

Harnad, S. (2007). Mandates and Metrics: How Open Repositories Enable Universities to Manage, Measure and Maximise their Research Assets. http://eprints.ecs.soton.ac.uk/14990/1/openaccess.pdf

Harnad, S. & Brody, T. (2004). Comparing the Impact of Open Access (OA) vs. Non-OA Articles in the Same Journals. D-Lib Magazine, 10(6). http://www.dlib.org/dlib/june04/harnad/06harnad.html

Kim, Y. H. & Kim, H.H. (2008). Development and validation of Evaluation Indicators for a Consortium of Institutional Repositories: A Case Study of dCollection. Journal of the American Society for Information Science and Technology, 59(8):1282-1294

Kousha, K. & Thelwall, M. (2006). Motivations for URL citations to open access library and information science articles. Scientometrics, 68(3):501-517.

Kousha, K. & Thelwall, M. (2007). Google Scholar Citations and Google Web/URL Citations: A Multi-Discipline Exploratory Analysis. Journal of the American Society for Information Science and Technology, 58(7):1055-1065

Kurtz, M. J.; Eichhorn, G.; Accomazzi, A.; Grant, C.; Denmleitner, M.; Henneken, E. & Murray, S.S. (2005). The effect of use and access on citations. Information Processing and Management, 41:1395-1402.

Lawrence, S.; Giles, C.L.; Bollacker, K. (1999). Digital Libraries and Autonomous Citation Indexing. IEEE Computer, 32(6): 67-71

McDonald, J. D. (2007). Understanding Journal Usage: A Statistical Analysis of Citation and Use. Journal of the American Society for Information Science and Technology, 58(1):39-50

Meho, L.I. & Yang, K. (2007). Impact of Data sources on Citation Counts and Rankings of LIS Faculty: Web of Science versus Scopus and Google Scholar. Journal of the American Society for Information Science and Technology, 58(3):2105-2125

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, 58(13):2047-2054

Mayr, P. (2006). Constructing experimental indicators for open access documents. Research Evaluation, 15(2):127-132.

Norris, M.; Oppenheim, Ch.; Rowland, F. (2008). The Citation advantage of Open Access articles. Journal of the American Society for Information Science and Technology, 59(12):1963-1972.

Organ, M.K. (2006). Download Statistics – What do they tell us?. The Example of Research Online, the Open Access Institutional Repository at the University of Wollongong, Australia. D-Lib Magazine, 12(11). http://www.dlib.org/dlib/november06/organ/organ11.html

Scholtz, F. & Dobratz, S. (2006). International Workshop on Institutional Repositories and Enhanced and alternative Metrics of Publication Impact, 20-21 February 2006, Humboldt University Berlin, Report. High Energy Physics Libraries Webzine, 13. Paper 2. http://library.cern.ch/HEPLW/13/papers/2/

Westell, M. (2006). Institutional Repositories: Proposed indicators of success. Library Hi Tech, 24(2):211-226.

Wouters, P., Reddy, C., Aguillo, I.F. (2006). On the visibility of information on the Web: an exploratory experimental approach. Research Evaluation, 15(2):107-115

Xia, J. & Sun, L. (2006). Factors to Assess Self-Archiving in Institutional Repositories. Serials Review, 33(2):73-80.

Zimmermann, C. (2007). Academic Rankings with RePEc. Working papers 2007-36, University of Connecticut, Department of Economics. http://www.econ.uconn.edu/working/2007-36.pdf

Zuccala, A.; Oppenheim, C. & Dhiensa, R. (2008). Managing and evaluating digital repositories. Information Research, 13(1): paper 333. http://InformationR.net/ir/13-1/paper333.html