Measuring the Institution's Footprint in the Web

Isidro Aguillo
Cybermetrics Lab, Centre of Social Sciences and Humanities (CCHS), Madrid, Spain

Abstract

Purpose:

Our purpose is to provide an alternative, although complementary, system for the evaluation of the scholarly activities of academic organizations, scholars and researchers, based on web indicators, in order to speed up the change of paradigm in scholarly communication towards a new fully electronic 21st century model.

Design/methodology/approach:

In order to achieve these goals, a new set of web indicators has been introduced, obtained mainly from data gathered from search engines, the new mediators of scholarly communication. We found that three large groups of indicators are feasible to obtain and relevant for evaluation purposes: activity (web publication); impact (visibility) and usage (visits and visitors).

Findings:

As a proof of concept, a *Ranking Web of Universities* has been built with Webometrics data. There are two relevant findings: ranking results are similar to those obtained by other bibliometric-based rankings; and there is a concerning digital divide between North American and European universities, which appear in lower positions when compared with their US & Canada counterparts.

Research limitations / implications:

Cybermetrics is still an emerging discipline so new developments should be expected when more empirical data become available.

Practical implications:

The proposed approach suggests the publication of truly electronic journals, rather than digital versions of printed articles. Additional materials such as raw data and multimedia files should be included along with other relevant information arising from more informal activities. These repositories should be Open Access, available as part of the public Web, indexed by the main commercial search engines. We anticipate that these actions could generate larger Web-based audiences, reduce the costs of publication and access and allow third parties to take advantage of the knowledge generated, without sacrificing peer review, which should be extended (pre- & post-) & expanded (closed & open).

Originality / value:

A full taxonomy of web indicators is introduced for describing and evaluating research activities, academic organizations and individual scholars and scientists. Previous attempts for building such classification were more incomplete and not taking into account feasibility and efficiency.

Paper type:

Conceptual paper

Keywords:

Scholarly communication, web indicators, Webometrics, link visibility, web usage

1. Introduction

The electronic publication of scientific papers has greatly increased the global audience for research activities (Evans, 2008) and also academic productivity (Barjak, 2006; Vakkari, 2008). Open access initiatives also have a great impact, and in the coming years will change scholarly communication. But most of these efforts are based on the old model of paper based journals with peer-review of the formal and almost-final version of the research results. There are several shortcomings linked to the traditional editorial process which can be overcome in the electronic (web) arena, but these have not yet been confronted.

Limitations of paper editions are clearly linked to their production and distribution costs. This means that only final results are published, in an economic format (short, not detailed, one language, without color photographs). A wide range of scholarly activities, including informal ones, are excluded, in particular the whole process leading to the results and access to the raw data used. As is shown by the current evolution in academic journals, a modern view should provide an extension of peer review (selected referees combined with open review (Beel and Gipp, 2008) and improved access to additional material, including new media (Van de Sompel, 2004). However the journal-centered model is no longer valid and current evaluation needs suggest focusing more on the user.

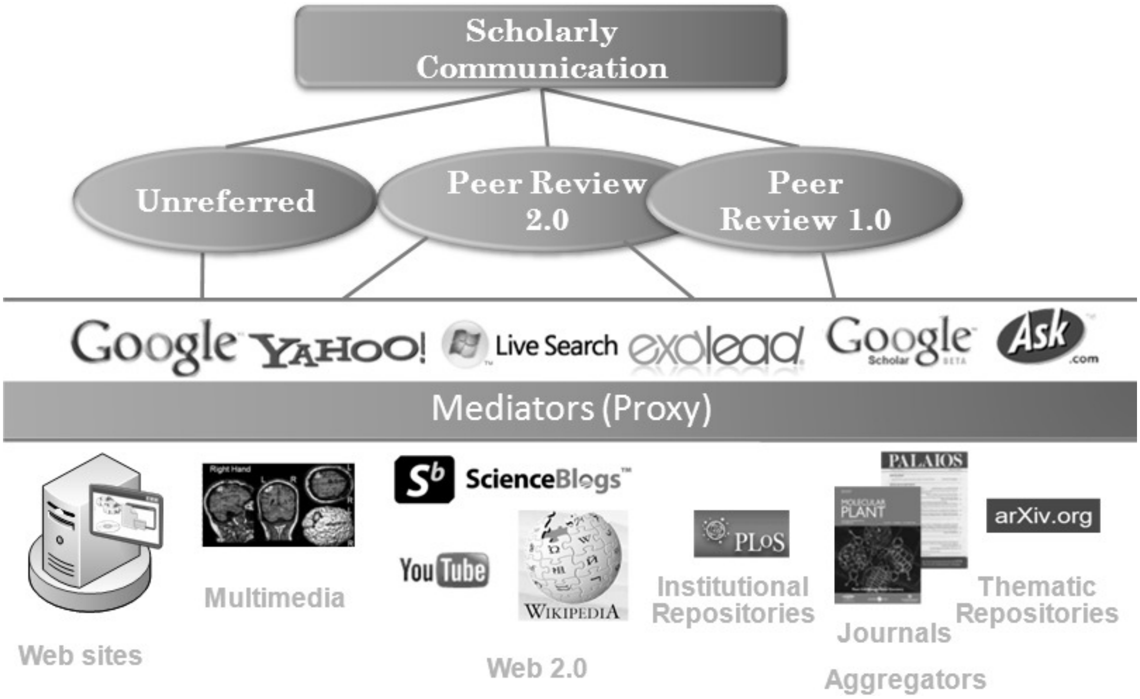


Figure 1. Proposal for a new model for scholarly communication, covering more activities, and proposing a central role for the Web search engines

The Web indicators are designed not only to monitor the presence and impact of an individual or an organization in webspace but to promote a more open, global, societal, and detailed knowledge of the scholars' organization, activities and results (Barjak, Li and Thelwall, 2007; Kousha and Thelwall, 2007). The proposal is to measure the Personal or Institutional Page 2.0 of an academic or research unit, including indicators of activity (number of webpages, documents or papers), impact (invocation, link visibility, page rank, “sitation” analysis) or usage (popularity, traffic).

2. Justification

There are several reasons for changing the way scholarly activities and research results are communicated.

The “serials crisis” (Swan, 2007) has shown that scholars have lost control of the system, which is in the hands of commercial publishers. Researchers freely give away their papers to publishers, who sell them back again to libraries, meaning that funders pay twice.

Most of the informal networks are far from democratic: Peer review processes are secretive and probably biased (Smith, 2006; Bornmann, Nast and Daniel, 2008), but it is extremely difficult to detect fraud. Most of these networks do not extend to developing countries and third (non-academic) parties are ignored.

The Open Access initiatives are very limited and their success is still under threat, in part because of reluctance towards institutional and self-archiving. Open peer review is an option, but only for a few disciplines and journals. Blogs and discussion boards are currently excluded.

In spite of the space constraints and high rejection rates from premium journals, which are slow and expensive to produce and distribute, there are still few truly electronic journals with multimedia support, digital access to primary resources or open forums.

e-Research is becoming more and more important (Rousay *et al.*, 2005), but the channels they use are unconventional. These contributions are mostly ignored in the evaluation processes in which there is low uptake of web-only materials.

A shift from bibliographic databases to web search engines has been observed, not only for information recovery purposes but also in citation studies: it is time for new ways of research assessment: "e-publish or perish".

3. Footprints in the Web

Academic and research organizations are perhaps the best reflected in webspace for several reasons. The web was born for scholarly communication purposes, the technical support needed for a good web presence is available in these institutions, academic freedom allows a large number of independent web editors and today it is cheaper to publish on the web than in traditional journals. But universities and research centers are very complex institutions, with a lot of different missions and a large number of academic and para-academic activities.

Today, universities have at least three core missions: teaching, involving not only traditional campus based learning but also distance and online education; research, done by faculty members or autonomous researchers but also by doctoral students; and the so-called "third mission" that consists of innovation and technological transfer to industrial and economic sectors and community engagement with local and regional social, cultural or political agents. Many universities host external events, support university hospitals, are in charge of museums, TV or radio stations or have important sports teams.

The web offers a feasible alternative for describing and evaluating all these missions and the activities involved. Moreover in many cases the web is not only a mediator .e-research (or e-science) activities show that the web is also an object of study. The web as an integrated communication tool is universal (global), democratic (very large audiences, rather than closed colleagues' clubs) and cheap (far cheaper than traditional paper-based journals and books). Web indicators complement the scenarios described by other scientometric statistics and provide new and unexpected relationships due to their larger coverage.

There are two sources of data for the web indicators. The websites can be crawled directly using specially designed robots that collect basic information through hypertextual navigation, or the statistics can be extracted from previously crawled databases obtained from commercial search engines. This indirect way is more flexible as access to search engines is universal and these robots are usually among the best ones available. There are technical and economic reasons for not using robots in large collections of websites, but perhaps the most important reason for using engines is that currently everybody uses them for information recovery. Despite coverage biases or other shortcomings, if a webpage is not indexed by them, then that page does not exist for any purpose. Web search engines are not only proxies but visibility mediators. Positioning strategies will become more and more important for scholarly communication in the future.

In summary, web indicators can be classified in three major categories: activity-related, measuring the volume (size) of information published; impact, according to the global network of links that connect webpages; and usage, counting visits and visitors and their behavior.

3.1 Activity

Web presence can be described fairly well from quantitative data obtained from search engines. Using special operators called delimiters, most of the large commercial engines provide figures (rounded or estimated) for the number of pages in a certain language, in a top level or institutional domain, from a country or in a specific file format. The syntax is not universal, but operators are more or less the same as shown in Table 1 (updated from Aguillo *et al.*, 2006).

OPERATOR	GOOGLE	YAHOO	LIVE	EXALEAD	ASK	GIGABLAST
Top Level Domain	site:aa.xx	Site Explorer http://xx	site:aa.xx	site:aa.xx	site:aa.xx inurl:aa.xx	site:aa.xx
Institutional Domain	site:aa.xx	Site Explorer http://aa.xx	site:aa.xx	site:aa.xx	site:aa.xx inurl:aa.xx	site:aa.xx
Directory	site:aa.xx/bb	site:aa.xx inurl:bb	site:aa.xx/bb	site:aa.xx/bb	inurl:aa.xx/bb	site:aa.xx suburl:bb
Term in URL	inurl:bb	inurl:bb	inurl:bb	inurl:bb	inurl:bb	suburl:bb
External Links	Only pages	linkdomain:aa.xx -site:aa.xx	NO	link:aa.xx -site:aa.xx	NO	NO
Country	Advanced Search	Advanced Search	loc:XX	country:XX	Advanced Search	NO
Language	Advanced Search	Advanced Search	language:zz	language:zz	Advanced Search	Advanced Search
File format	filetype:yy	originurlextension:yy	filetype:yy	filetype:yy	NO	type:yy

Table 1. Syntax of the main search engines

The size of a website or a web domain could be measured by the number of pages, usually in html or assimilated formats. Since the beginning of the century most of the engines also allow the counting of specific formats, which can be useful because they have document properties. They are referred to as rich files and include popular types as Adobe Acrobat (pdf), MS Office formats (doc and rtf for Word, ppt for Powerpoint) or PostScript (ps). These rich files are important because many of them are entire papers or other scientific documents, so they are a good indicator of academic information published. However, not every piece of information in these formats has that academic origin, so specialized search engines must also be considered. After the demise of Live academic, Google Scholar is by far the most relevant (Kousha and Thelwall, 2007) being compared even with Web of Science and Scopus, the giant subscription-based bibliographic databases. Other interesting options are the open version of the Elsevier databases (Scirus, www.scirus.com) and the increasing number of repositories, especially the large harvesters that offer a unique search interface to recover records simultaneously from different repositories. Unfortunately, the Webometrics capabilities of many of these databases are limited.

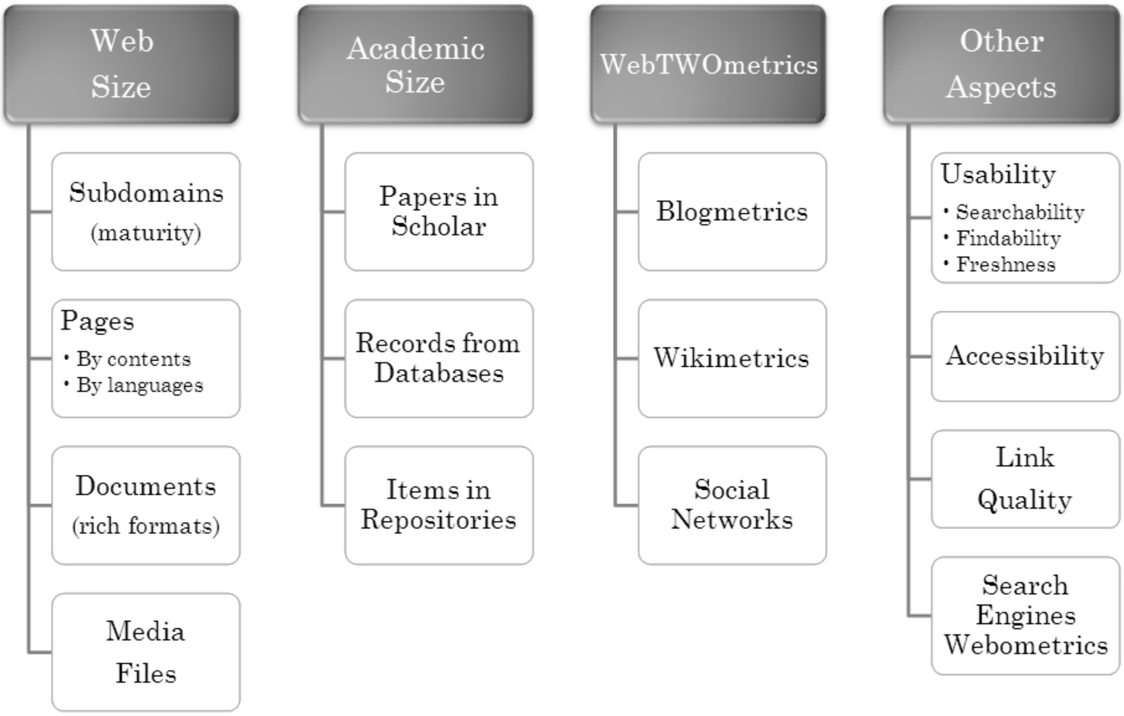
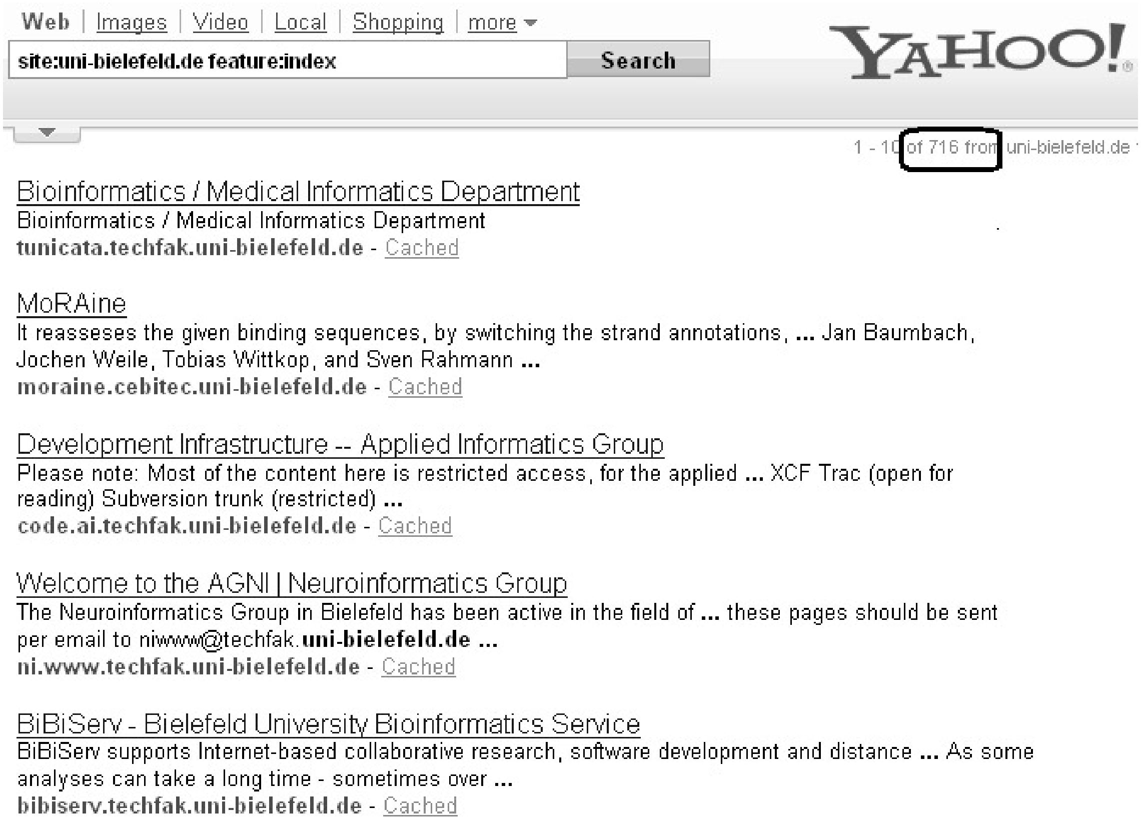


Figure 2. Activity (Web Presence)

Figure 2 summarizes other relevant presence-related indicators. Some of them are not currently very useful, but their future impact could be different. For example, media files (video, audio and other similar archives) could help to describe commitment to communicating science to the general public.

Nowadays the large universities no longer have central control of the contents of their web domain, so each department or research group has the possibility of establishing their own autonomous website with a specific subdomain. Counting the number of these subdomains could be seen as a measure of the maturity of the whole domain (syntax for the Yahoo engine offered in Figure 3).



site:uni-bielefeld.de feature:index

1 - 10 of 716 from uni-bielefeld.de

Bioinformatics / Medical Informatics Department
Bioinformatics / Medical Informatics Department
tunicata.techfak.uni-bielefeld.de - Cached

MoRAine
It reasseses the given binding sequences, by switching the strand annotations, ... Jan Baumbach, Jochen Weile, Tobias Wittkop, and Sven Rahmann ...
moraine.cebitec.uni-bielefeld.de - Cached

Development Infrastructure -- Applied Informatics Group
Please note: Most of the content here is restricted access, for the applied ... XCF Trac (open for reading) Subversion trunk (restricted) ...
code.ai.techfak.uni-bielefeld.de - Cached

Welcome to the AGNI | Neuroinformatics Group
The Neuroinformatics Group in Bielefeld has been active in the field of ... these pages should be sent per email to niwww@techfak.**uni-bielefeld.de** ...
ni.www.techfak.uni-bielefeld.de - Cached

BiBiServ - Bielefeld University Bioinformatics Service
BiBiServ supports Internet-based collaborative research, software development and distance ... As some analyses can take a long time - sometimes over ...
bibiserv.techfak.uni-bielefeld.de - Cached

Figure 3. Number of subdomains (716) of Bielefeld University according to Yahoo! Search engine and its operator feature:index (January 2009)

New emerging quantitative disciplines are related to Web 2.0 technologies, so blogmetrics or wikimetrics (Smith, 2007; Torres-Salinas *et al.*, 2008; Voss, 2005) could provide further evidence of the importance of informal channels for scholarly communication.

Finally there is strong interest in the quantitative analysis of the architecture of the information on the Web, including formal aspects such as usability, accessibility, searchability or findability. Literature on these topics is widely available (Palmer, 2002; Olsina *et al.*, 1999).

The search engines themselves are also objects of analysis, focusing on quantitative aspects of their databases, such as size, coverage or freshness. Ranking characteristics could also be included in this section.

3.2 Impact

Impact on the web (Figure 4) could be inferred from the number of times the contents of a webpage or websites are mentioned in or linked from third pages. This is a strong group of indicators because only a selected group of people could "site" (sitation=site citation) a webpage: those who are authors or editors of webpages. Webmasters can be expected to know about the target page, which means that it should be visible (from search engines), with legible contents and available 24/7. Of course, in order for a link to be created, the content of the target page must fulfill the quality criteria of the linker and in most of the cases this means that this webmaster is familiar with the topic or even an expert, increasing the possibility in an academic environment of producing true bibliographic citations (Brody, Harnad and Carr, 2006).

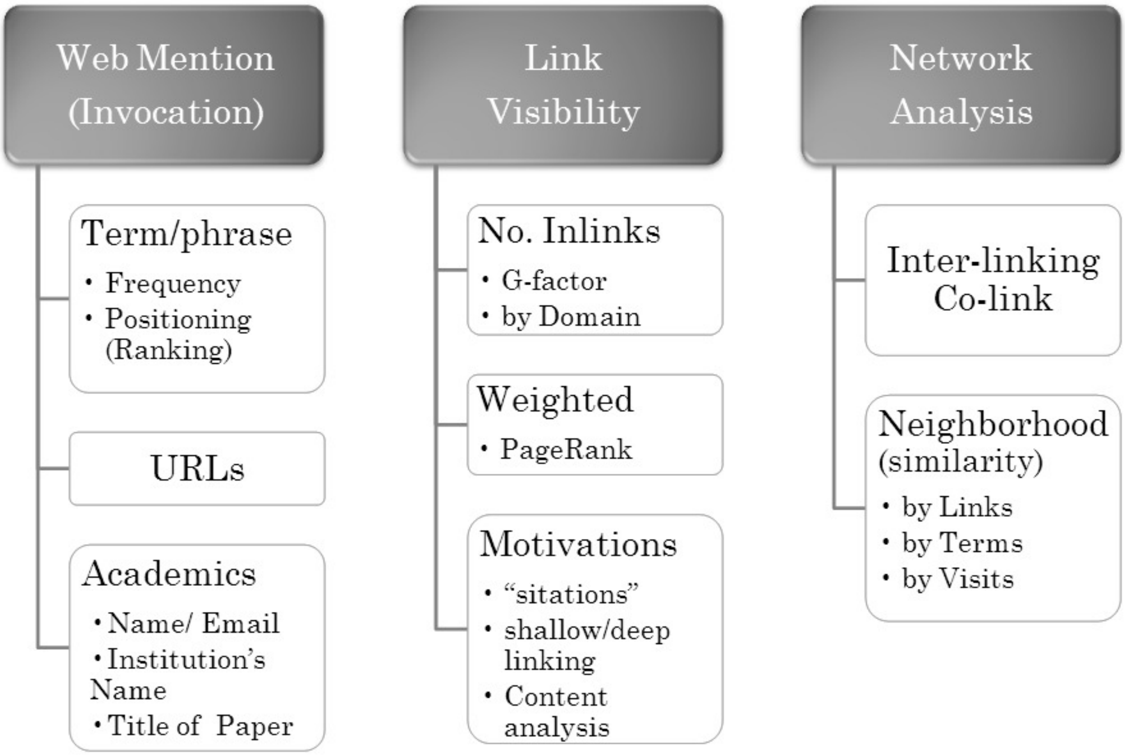
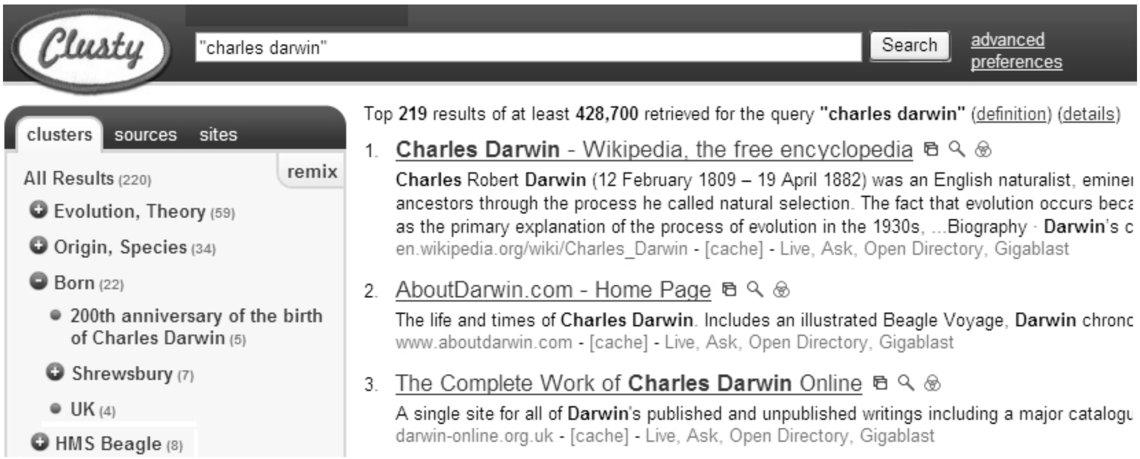


Figure 4. Impact

Mentioning is not as strong an option as providing a link, but nevertheless it could be used for projects requiring semantic delimitation, difficult to solve by link analysis. An interesting example is provided by clustering. A series of pages using one or several words can be grouped according to their contents and tagged with descriptive terms (based on frequency, for example). Examining these collections of tags we find an automatic build description of the targeted terms that works well in many cases (Figure 5)



"charles darwin"

All Results (220)
- Evolution, Theory (59)
- Origin, Species (34)
- Born (22)
  - 200th anniversary of the birth of Charles Darwin (5)
  - Shrewsbury (7)
  - UK (4)
- HMS Beagle (8)

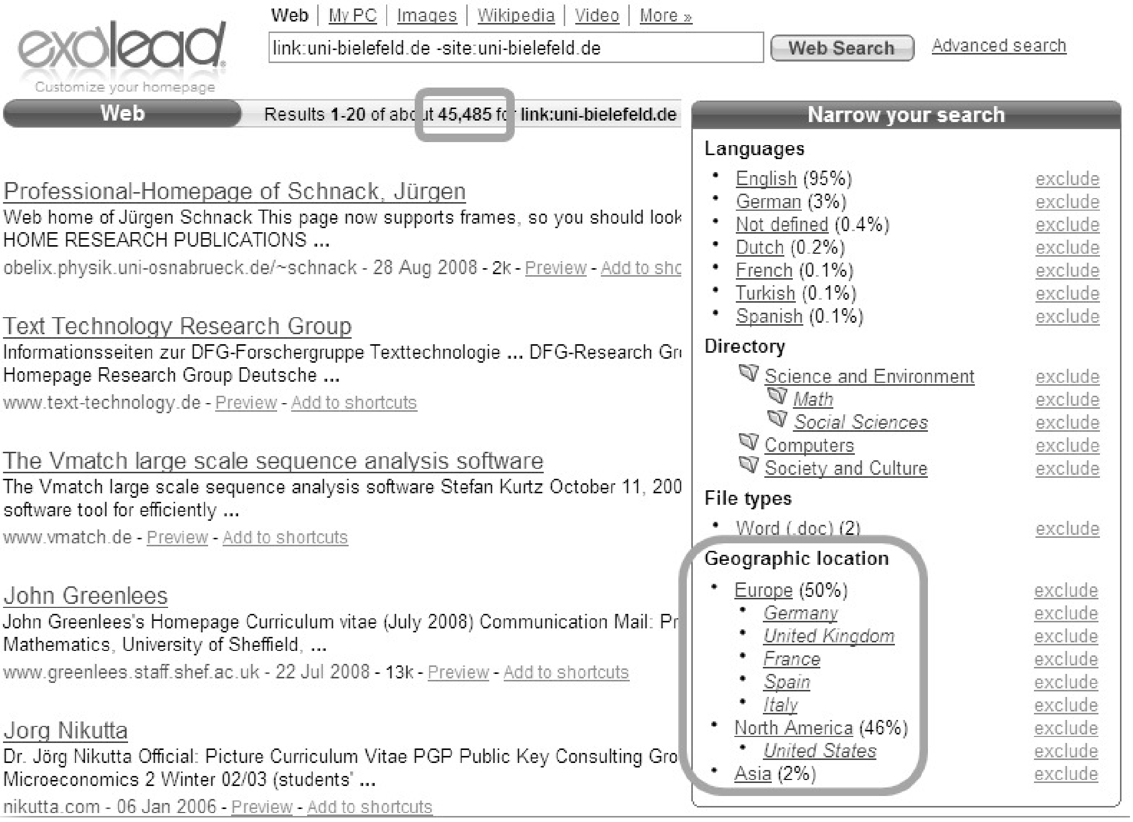
Top **219** results of at least **428,700** retrieved for the query **"charles darwin"** (definition) (details)

1. **Charles Darwin** - Wikipedia, the free encyclopedia
   Charles Robert **Darwin** (12 February 1809 – 19 April 1882) was an English naturalist, eminei ancestors through the process he called natural selection. The fact that evolution occurs beca as the primary explanation of the process of evolution in the 1930s, ...Biography · **Darwin**'s c
   en.wikipedia.org/wiki/Charles_Darwin - [cache] - Live, Ask, Open Directory, Gigablast
2. AboutDarwin.com - Home Page
   The life and times of **Charles Darwin**. Includes an illustrated Beagle Voyage, **Darwin** chronc
   www.aboutdarwin.com - [cache] - Live, Ask, Open Directory, Gigablast
3. The Complete Work of **Charles Darwin** Online
   A single site for all of **Darwin**'s published and unpublished writings including a major catalogu
   darwin-online.org.uk - [cache] - Live, Ask, Open Directory, Gigablast

Figure 5. Clustering of results for “Charles Darwin” provides a collection of terms (left side) that describes the main aspects of the renowned scientist (Clusty search engine, www.clusty.com, January 2009 edited)

Invocation of a word, name or sentence could either be noisy, as the search engines stem the terms, or silent, as variants, synonyms or translations are excluded. A good experimental design could help but it is important to remember that, for example, an academic institution could be included in several hundred variant forms if not more (Van Raan, 2005)

Link analysis is probably the most powerful tool of Webometrics (Thelwall, Vaughan and Björneborn, 2005). Although motivations for linking are diverse and complex, in the academic arena many of them can be assimilated to the bibliographic citations commonly used in bibliometric studies (Vaughan and Thelwall, 2005). The number of external inlinks received by a webpage can be easily obtained from several search engines (Figure 6).



link:uni-bielefeld.de -site:uni-bielefeld.de

Results 1-20 of about 45,485 for link:uni-bielefeld.de

Professional-Homepage of Schnack, Jürgen
Web home of Jürgen Schnack This page now supports frames, so you should look
HOME RESEARCH PUBLICATIONS ...
obelix.physik.uni-osnabrueck.de/~schnack - 28 Aug 2008 - 2k - Preview - Add to shc

Text Technology Research Group
Informationsseiten zur DFG-Forschergruppe Texttechnologie ... DFG-Research Gr
Homepage Research Group Deutsche ...
www.text-technology.de - Preview - Add to shortcuts

The Vmatch large scale sequence analysis software
The Vmatch large scale sequence analysis software Stefan Kurtz October 11, 200
software tool for efficiently ...
www.vmatch.de - Preview - Add to shortcuts

John Greenlees
John Greenlees's Homepage Curriculum vitae (July 2008) Communication Mail: Pr
Mathematics, University of Sheffield, ...
www.greenlees.staff.shef.ac.uk - 22 Jul 2008 - 13k - Preview - Add to shortcuts

Jorg Nikutta
Dr. Jörg Nikutta Official: Picture Curriculum Vitae PGP Public Key Consulting Gro
Microeconomics 2 Winter 02/03 (students' ...
nikutta.com - 06 Jan 2006 - Preview - Add to shortcuts

Languages
- English (95%) exclude
- German (3%) exclude
- Not defined (0.4%) exclude
- Dutch (0.2%) exclude
- French (0.1%) exclude
- Turkish (0.1%) exclude
- Spanish (0.1%) exclude

Directory
- Science and Environment exclude
  - Math exclude
  - Social Sciences exclude
- Computers exclude
- Society and Culture exclude

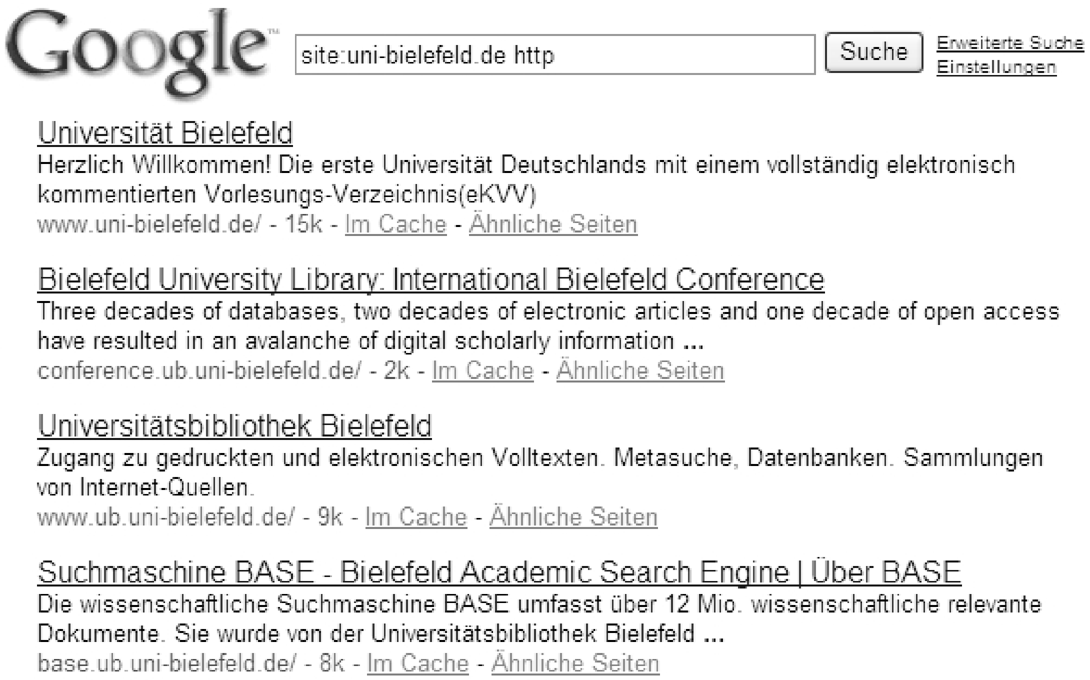
File types
- Word (.doc) (2) exclude

Geographic location
- Europe (50%) exclude
  - Germany exclude
  - United Kingdom exclude
  - France exclude
  - Spain exclude
  - Italy exclude
- North America (46%) exclude
  - United States exclude
- Asia (2%) exclude

Figure 6. Link visibility for Bielefeld University according to Exalead search engine (January 2009). This engine provides the total number of inlinks as well as the geographical origin of the linking pages.

As was mentioned earlier, the global number of inlinks could be a noisy measure, especially as a university or research center can receive links from non-academic websites. In these cases it is possible to use filtering by institutional domains (other university webdomains) or academic top level domains when available (edu, ac.uk, edu.au, ac.be). Mike Thelwall's research group has published extensively on these topics (Barjak and Thelwall, 2008; Li *et al.*, 2005a,b; Park and Thelwall, 2006; Payne and Thelwall, 2008; Stuart, Thelwall and Harries, 2007).

The raw number of links is a good indicator, but can be improved by adding different weights to the pages where the links originated. This is roughly the basis of the successful Google Pagerank algorithm (Thelwall, 2003). Pagerank (PR) allows the classification of webpages according to their importance in the hypertextual web network. Global calculation could be very complex but Google offers the possibility of obtaining lists of webpages organized by decreasing order of PR. Using a neutral term, the list of a delimited search (domain, subdomain, language, country) appears ordered by that algorithm (Figure 7).



1 Universität Bielefeld
Herzlich Willkommen! Die erste Universität Deutschlands mit einem vollständig elektronisch kommentierten Vorlesungs-Verzeichnis(eKVV)
www.uni-bielefeld.de/ - 15k - Im Cache - Ähnliche Seiten

2 Bielefeld University Library: International Bielefeld Conference
Three decades of databases, two decades of electronic articles and one decade of open access have resulted in an avalanche of digital scholarly information ...
conference.ub.uni-bielefeld.de/ - 2k - Im Cache - Ähnliche Seiten

3 Universitätsbibliothek Bielefeld
Zugang zu gedruckten und elektronischen Volltexten. Metasuche, Datenbanken. Sammlungen von Internet-Quellen.
www.ub.uni-bielefeld.de/ - 9k - Im Cache - Ähnliche Seiten

4 Suchmaschine BASE - Bielefeld Academic Search Engine | Über BASE
Die wissenschaftliche Suchmaschine BASE umfasst über 12 Mio. wissenschaftliche relevante Dokumente. Sie wurde von der Universitätsbibliothek Bielefeld ...
base.ub.uni-bielefeld.de/ - 8k - Im Cache - Ähnliche Seiten

Figure 7. List of Bielefeld University's webpages ranked by their Pagerank (Google, January 2009).

There are more sophisticated techniques based on the number of links connecting organizations or countries or using co-linking data. There are already several network analyses based on these results (Ortega and Aguillo, 2008a,b; Ortega *et al.*, 2008).

However it is possible easily to obtain similar results using some visualization services freely available from the Web. The three services introduced here allow the "neighborhood" of a web site to be shown according to three different criteria, but offer surprisingly similar results.

Touchgraph Google (www.touchgraph.com) is based in the "related" option (similar pages) of this search engine. Basically it is a graphic interface to the results provided by Google. Although the algorithm is not public, Google associates websites according to their link patterns, assuming two pages are closer if the overlap between their in- and out-links is high. This provides a hypertextual neighborhood, which, in the case of universities, mainly consists of other geographically close universities. Figure 8 shows that in the case of Bielefeld University, the closest websites are other German universities and information about the city.

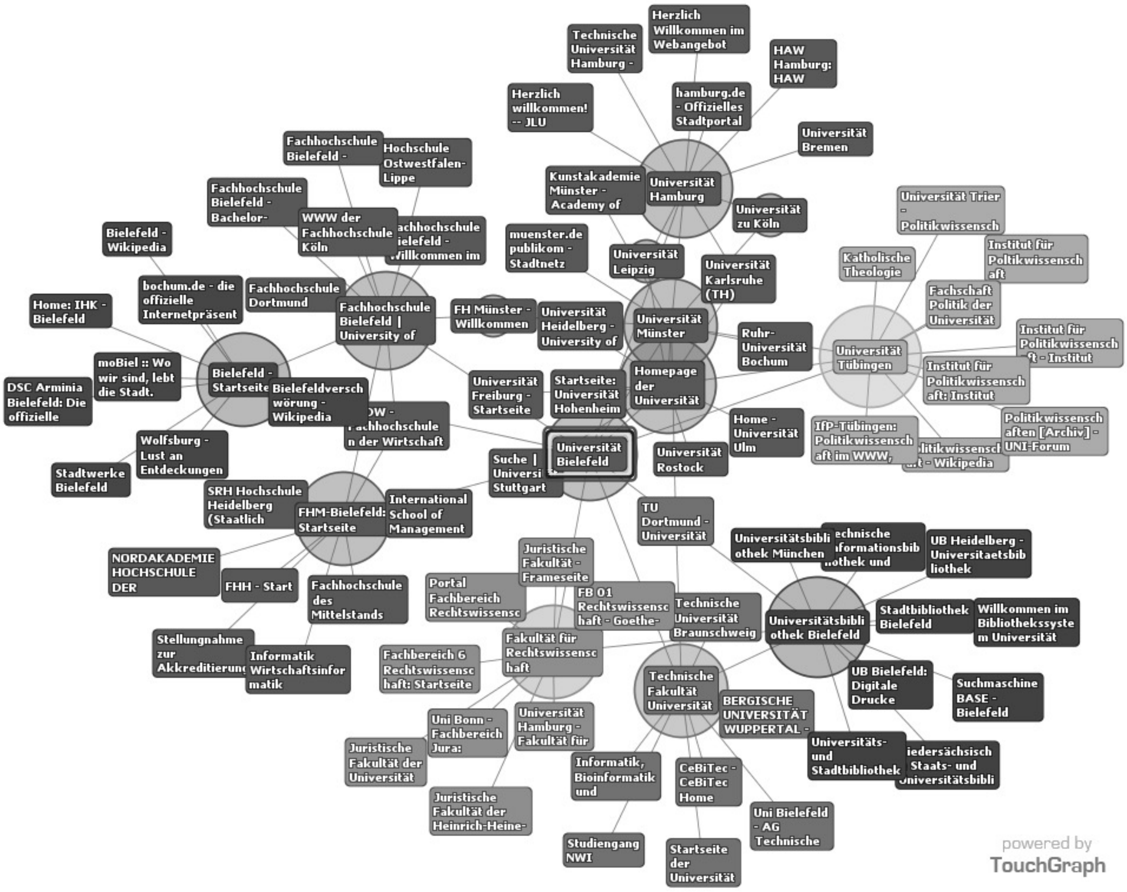
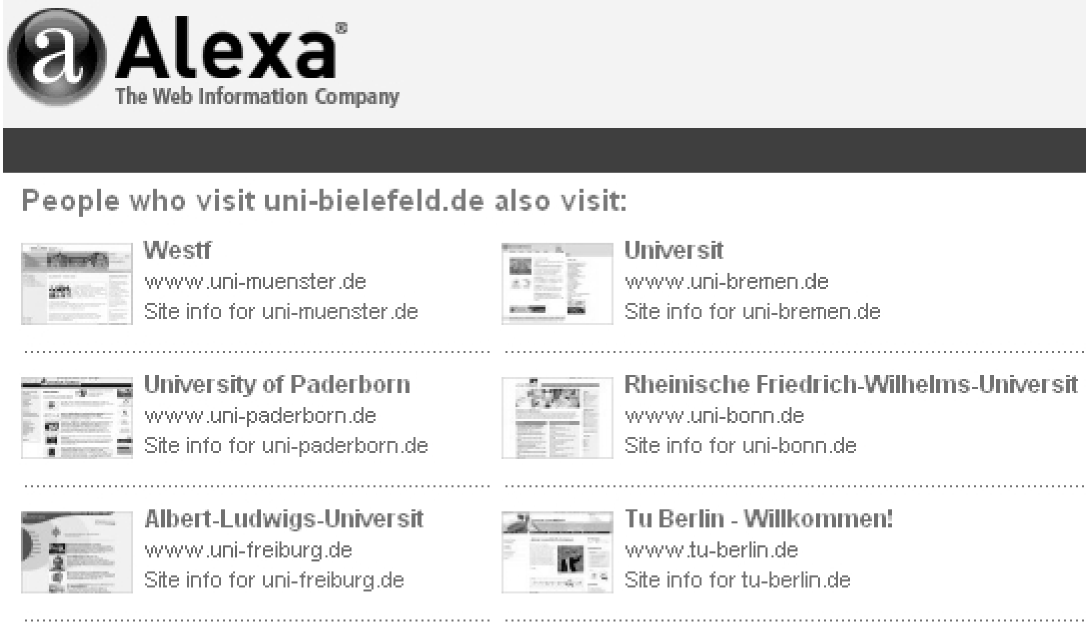

Herzlich Willkommen im Webangebot
Technische Universität Hamburg -
HAW Hamburg: HAW
Herzlich willkommen! -- JLU
hamburg.de - Offizielles Stadtportal
Universität Bremen
Fachhochschule Bielefeld -
Hochschule Ostwestfalen-Lippe
Kunstakademie Münster - Academy of
Universität Hamburg
Universität zu Köln
Universität Trier - Politikwissensch
Fachhochschule Bielefeld - Bachelor-
WWW der Fachhochschule Köln
achhochschule ielefeld - Willkommen im
muenster.de publikom - Stadtnetz
Universität Leipzig
Universität Karlsruhe (TH)
Katholische Theologie
Institut für Poltikwissensch aft
Bielefeld - Wikipedia
bochum.de - die offizielle Internetpräsent
Fachhochschule Dortmund
Fachhochschule Bielefeld | University of
FH Münster - Willkommen
Universität Heidelberg - University of
Universität Münster
Fachschaft Politik der Universität
Home: IHK - Bielefeld
Ruhr-Universität Bochum
Universität Tübingen
Institut für Politikwissensch ft - Institut
moBiel :: Wo wir sind, lebt die Stadt.
Bielefeld - Startseite
Bielefeldversch wörung - Wikipedia
Universität Freiburg - Startseite
Startseite: Universität Hohenheim
Homepage der Universität
Institut für Politikwissensch aft: Institut
DSC Arminia Bielefeld: Die offizielle
DW - Fachhochschule n der Wirtschaft
Home - Universität Ulm
IfP-Tübingen: Politikwissensch aft im WWW,
litikwissensch ft - Wikipedia
Politikwissensch aften [Archiv] - UNI-Forum
Wolfsburg - Lust an Entdeckungen
Universität Bielefeld
Suche | Universi Stuttgart
Universität Rostock
Stadtwerke Bielefeld
SRH Hochschule Heidelberg (Staatlich
FHM-Bielefeld: Startseite
International School of Management
TU Dortmund - Universität
Universitätsbibliothek München
echnische nformationsbib iothek und
UB Heidelberg - Universitaetsbibliothek
Juristische Fakultät - Frameseite
NORDAKADEMIE HOCHSCHULE DER
FHH - Start
Fachhochschule des Mittelstands
Portal Fachbereich Rechtswissensc
FB 01 Rechtswissensc haft - Goethe-
Technische Universität Braunschweig
Universitätsbibliothek Bielefeld
Stadtbibliothek Bielefeld
Willkommen im Bibliothekssystem Universität
Stellungnahme zur Akkreditierung
Informatik Wirtschaftsinformatik
Fachbereich 6 Rechtswissensc haft: Startseite
Fakultät für Rechtswissensc haft
Technische Fakultät Universität
UB Bielefeld: Digitale Drucke
Suchmaschine BASE - Bielefeld
BERGISCHE UNIVERSITÄT WUPPERTAL -
Uni Bonn - Fachbereich Jura:
Universität Hamburg - Fakultät für
Juristische Fakultät der Universität
Universitäts- und Stadtbibliothek
iedersächsisch Staats- und Universitätsbibli
Informatik, Bioinformatik und
CeBiTec - CeBiTec Home
Juristische Fakultät der Heinrich-Heine-
Uni Bielefeld - AG Technische
Studiengang NWI
Startseite der Universität
powered by TouchGraph

Figure 8. Neighborhood of Bielefeld University (www.uni-bielefeld.de) according to Touchgraph Google (January, 2009)

The semantic neighborhood can be seen using programs like Kartoo (www.kartoo.com) or Ujiko (www.ujiko.com) which show connections between webpages through common words. Unfortunately the information is collected from small databases and the results do not have high precision.

Alexa (www.alexa.com), provides a third possibility: grouping the websites according to visits received, so if several pages are visited during the same session they are considered neighbors. Of course the success of the system depends on a high volume of visits, so it is especially useful for very popular sites. Amazon, the parent company of Alexa, offers a similar service: "Customers Who Bought This Item Also Bought …" Figure 9 shows that German universities have a common base of customers.



People who visit uni-bielefeld.de also visit:

Westf
www.uni-muenster.de
Site info for uni-muenster.de

Universit
www.uni-bremen.de
Site info for uni-bremen.de

University of Paderborn
www.uni-paderborn.de
Site info for uni-paderborn.de

Rheinische Friedrich-Wilhelms-Universit
www.uni-bonn.de
Site info for uni-bonn.de

Albert-Ludwigs-Universit
www.uni-freiburg.de
Site info for uni-freiburg.de

Tu Berlin - Willkommen!
www.tu-berlin.de
Site info for tu-berlin.de

Figure 9. People searching for Bielefeld University also visit other German universities (Alexa, January 2009)

3.3 Usage

The evaluation of institutions using usage data is very new, as there are few papers dealing with journal or library circulation. The situation has changed abruptly with electronic publications, as there is no longer a lack of reliable and comparable statistics, but on the contrary a lot of new indicators have arisen from the log files of institutional and personal websites, journal portals and repositories.

Not all statistics collected from log files (the files that collect usage data in the webservers) are useful for academic purposes, and sometimes the server has to be customized in order to obtain figures for certain behaviors and according to specific criteria. The general pattern is very important, but so also are individual actions or actions related to specific files, such as downloading (Figure 10).

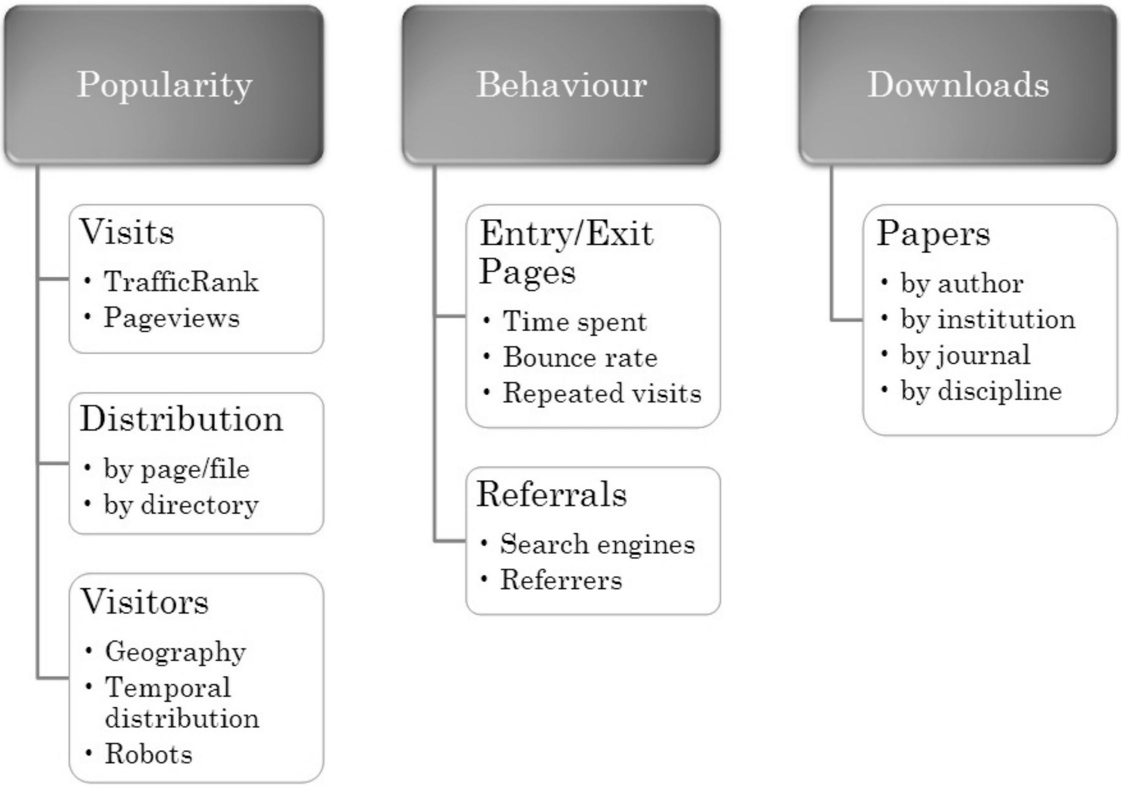
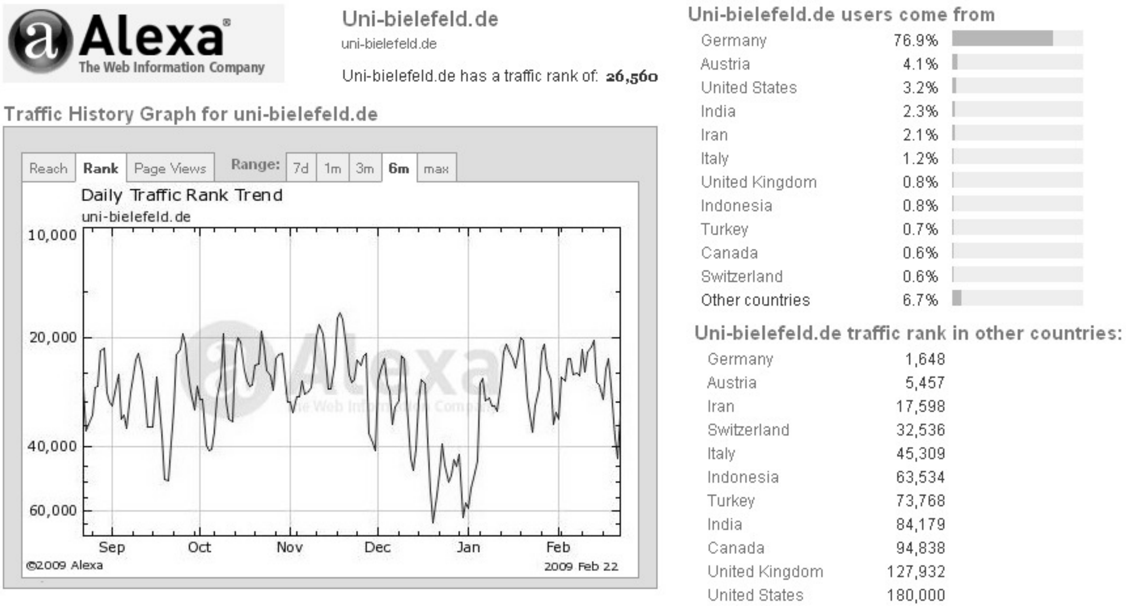


Figure 10. Usage

There are two possibilities for undertaking web metrics analysis (not to be confused with webometrics): using an intermediate database such as Alexa; or having access to a series of log files (each webserver generates one).

Alexa uses as a source a large group of users worldwide (although geographically biased) that inadvertently send information about the sites they visit to a central location. Alexa then ranks the webdomains (not individual websites) according to a three-month mean of visits. No raw data is provided, so it is difficult to know the actual differences between domains. Moreover, the ranking of academic institutions shows marked variations, especially during weekends and holiday periods.

As shown in Figure 11, the Traffic Rank indicates that the Bielefeld University webdomain is the 26560th most visited in the world (Alexa database), that three quarters of the visitors came from Germany, but that it is also a popular destination for Iranian and Indonesian people. This last conclusion is perhaps a result of the coverage bias mentioned above.



Uni-bielefeld.de
uni-bielefeld.de
Uni-bielefeld.de has a traffic rank of: 26,560

Traffic History Graph for uni-bielefeld.de

Uni-bielefeld.de users come from

Germany	76.9%
Austria	4.1%
United States	3.2%
India	2.3%
Iran	2.1%
Italy	1.2%
United Kingdom	0.8%
Indonesia	0.8%
Turkey	0.7%
Canada	0.6%
Switzerland	0.6%
Other countries	6.7%

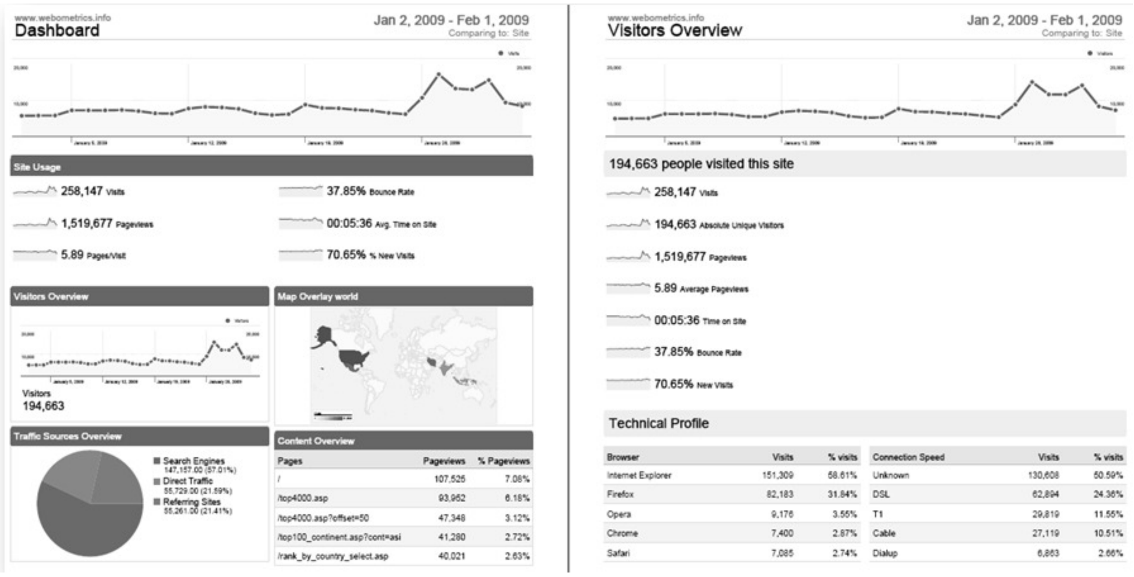
Uni-bielefeld.de traffic rank in other countries:

Germany	1,648
Austria	5,457
Iran	17,598
Switzerland	32,536
Italy	45,309
Indonesia	63,534
Turkey	73,768
India	84,179
Canada	94,838
United Kingdom	127,932
United States	180,000

Figure 11. Traffic Rank of Bielefeld University, with geographical distribution of the visitors' origin (Alexa, February 2009)

Extracting data from a log file is fairly easy, but there are a lot of shortcomings that should be taken into account. First of all, privacy issues make it inadvisable to use the personal information of the visitors. Cybergeography data might be acceptable, but demographic information, if available, would not. It is important to exclude visits by search engine robots and also to define time lags for defining different visits from the same visitor.

Google Analytics (Figure 12) has become a very popular option because it is free, it is based on a powerful system (Urchin) and of course it is supported by Google. However, the standard configuration is not very complete, it is difficult to customize and not very well designed for downloading analysis.


www.webometrics.info
Dashboard
Jan 2, 2009 - Feb 1, 2009
Comparing to: Site

Site Usage

258,147 Visits
1,519,677 Pageviews
5.89 Pages/Visit
37.85% Bounce Rate
00:05:36 Avg. Time on Site
70.65% % New Visits

Visitors Overview

Visitors
194,663

Map Overlay world

Traffic Sources Overview

Search Engines 147,157.00 (57.01%)
Direct Traffic 55,729.00 (21.59%)
Referring Sites 55,261.00 (21.41%)

Content Overview

Pages	Pageviews	% Pageviews
/	107,525	7.08%
/top4000.asp	93,962	6.18%
/top4000.asp?offset=50	47,348	3.12%
/top100_continent.asp?cont=asi	41,280	2.72%
/rank_by_country_select.asp	40,021	2.63%

www.webometrics.info
Visitors Overview
Jan 2, 2009 - Feb 1, 2009
Comparing to: Site

194,663 people visited this site

258,147 Visits
194,663 Absolute Unique Visitors
1,519,677 Pageviews
5.89 Average Pageviews
00:05:36 Time on Site
37.85% Bounce Rate
70.65% New Visits

Technical Profile

Browser	Visits	% visits
Internet Explorer	151,309	58.61%
Firefox	82,183	31.84%
Opera	9,176	3.55%
Chrome	7,400	2.87%
Safari	7,085	2.74%

Connection Speed	Visits	% visits
Unknown	130,608	50.59%
DSL	62,894	24.36%
T1	29,819	11.55%
Cable	27,119	10.51%
Dialup	6,863	2.66%

Figure 12. Dashboard of the visits to the Webometrics.info website according to Google Analytics (February 2009)

4. Ranking Web

In order to test the value of these indicators, the Cybermetrics Lab has started to collect them for a large group of universities and research centers. As recently as 2003 the Shanghai Jiaotong University published their famous Academic Ranking of World Universities (www.arwu.org) and it was decided to use a similar system for the Webometrics data.

Inspired by the Journal Impact Factor (a ratio between citations and papers), a Web Impact Factor (links/pages) was proposed, an indicator that does not work due to the power-law distribution of the statistics on which it is based. A new indicator (Webometrics Rank or WR) was proposed, inspired by former indicators, which maintains the ratio 1:1 between links (a kind of citation) and pages (web presence as a measure of activity). In order to reinforce the academic weight, the activity indicator was split into three subcomponents: number of webpages, number of documents (rich file formats) and number of papers (items in Google Scholar).

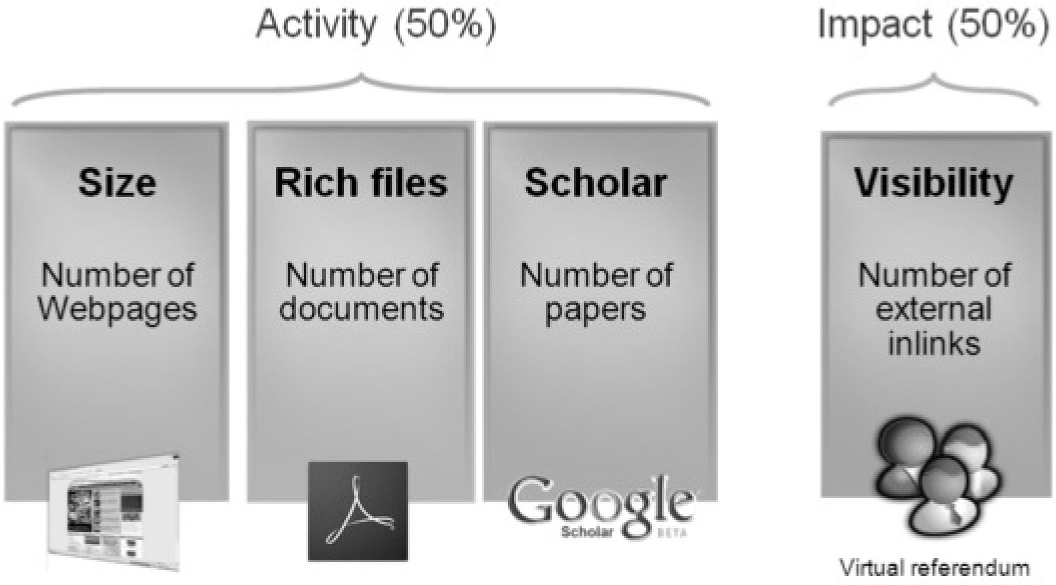
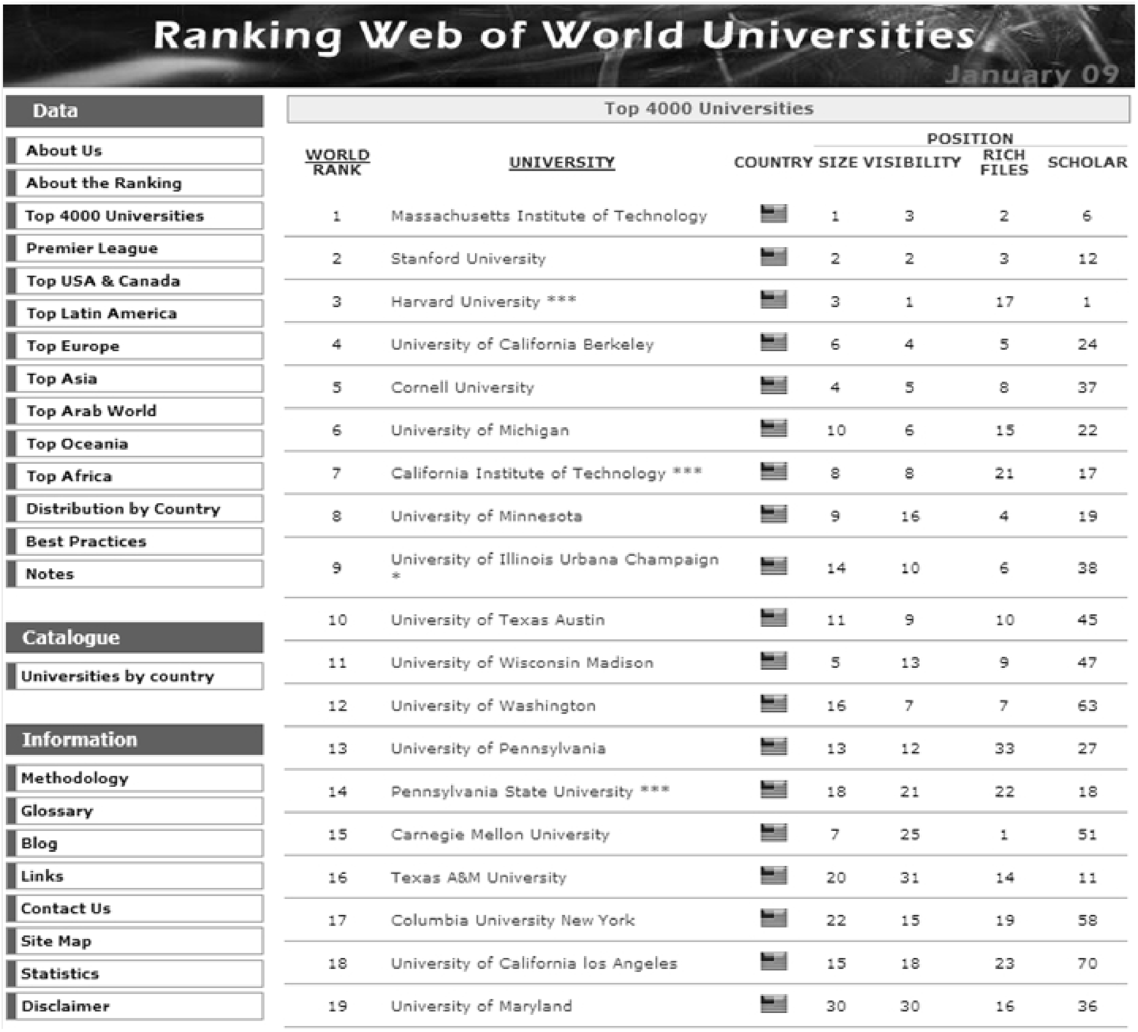


Figure 13. Academic Model of the WR indicator

The Ranking Web of Universities (www.webometrics.info) has been published, twice a year (January & July) since 2004 (Aguillo, Ortega, and Fernández, 2008). Based on a Directory of more than 16,000 Higher Education Institutions, it classifies more than 6000 universities worldwide according to web indicators (WR). There is also a Ranking Web of Research Centers (research.webometrics.info) listing the Top 2000 organizations from a Directory containing 7,000 entries.

The results (Figure 14) obtained are similar to those provided by other ranking systems, with several US prestigious universities in the top positions, Cambridge and ETH Zurich leading European institutions, and UNAM, Tokyo and the Australian National University being other top regional universities.



Ranking Web of World Universities

January 09

Data

About Us
About the Ranking
Top 4000 Universities
Premier League
Top USA & Canada
Top Latin America
Top Europe
Top Asia
Top Arab World
Top Oceania
Top Africa
Distribution by Country
Best Practices
Notes

Catalogue

Universities by country

Information

Methodology
Glossary
Blog
Links
Contact Us
Site Map
Statistics
Disclaimer

Top 4000 Universities

WORLD RANK	UNIVERSITY	COUNTRY	POSITION SIZE	 VISIBILITY	 RICH FILES	 SCHOLAR
1	Massachusetts Institute of Technology	🇺🇸	1	3	2	6
2	Stanford University	🇺🇸	2	2	3	12
3	Harvard University ***	🇺🇸	3	1	17	1
4	University of California Berkeley	🇺🇸	6	4	5	24
5	Cornell University	🇺🇸	4	5	8	37
6	University of Michigan	🇺🇸	10	6	15	22
7	California Institute of Technology ***	🇺🇸	8	8	21	17
8	University of Minnesota	🇺🇸	9	16	4	19
9	University of Illinois Urbana Champaign *	🇺🇸	14	10	6	38
10	University of Texas Austin	🇺🇸	11	9	10	45
11	University of Wisconsin Madison	🇺🇸	5	13	9	47
12	University of Washington	🇺🇸	16	7	7	63
13	University of Pennsylvania	🇺🇸	13	12	33	27
14	Pennsylvania State University ***	🇺🇸	18	21	22	18
15	Carnegie Mellon University	🇺🇸	7	25	1	51
16	Texas A&M University	🇺🇸	20	31	14	11
17	Columbia University New York	🇺🇸	22	15	19	58
18	University of California los Angeles	🇺🇸	15	18	23	70
19	University of Maryland	🇺🇸	30	30	16	36

Figure 14. Ranking Web of World Universities

The Ranking is a good tool, not only as a League Table but also for uncovering unexpected patterns. The most important contribution is the discovery of an academic digital divide that affects not developing countries but European ones. The Ranking shows that there are many more North American (US but also Canadian) universities in the Top 100 & 200 than their European counterparts, by a factor of two or three, as is shown in Figure 15.

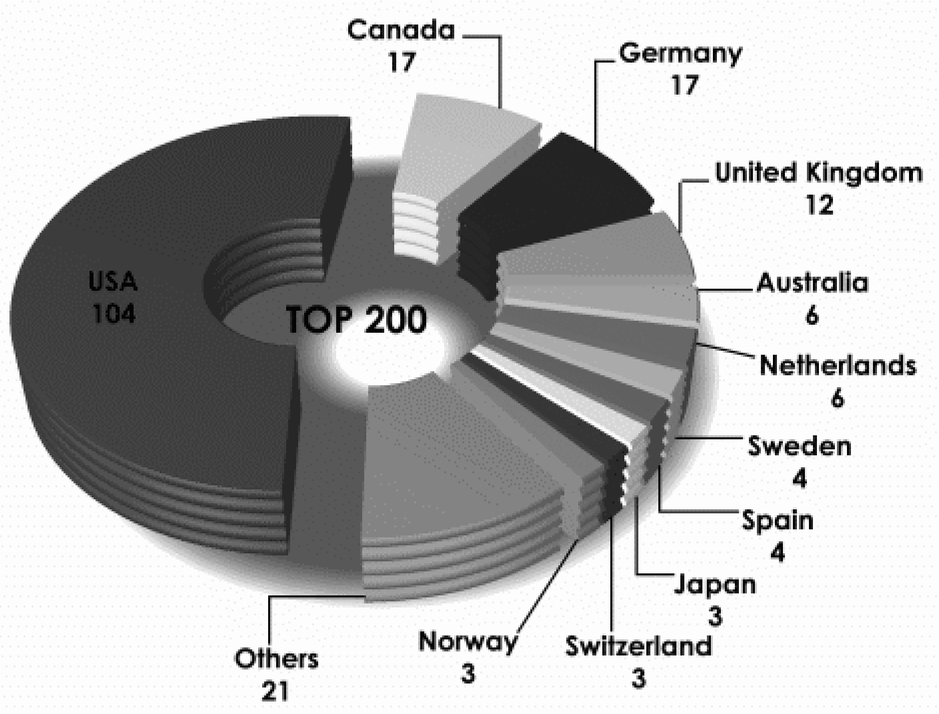


Figure 15. Digital Divide as shown by the country distribution of the Top 200 Universities in the Ranking web of World Universities (January 2009)

5. Conclusions

Scholarly communication is already digital, but the new characteristics of electronic publications are not yet being fully exploited. In fact a new revamped system is needed to take advantage of the possibilities offered by the Web, which is a more universal, democratic, powerful and a cheaper communication tool than the paper based ones.

The problem is that current evaluation techniques of scholarly activities, research and academic performance are still based on the previous paradigm. Those methods ignore the contributions of informal channels such as those related to Web 2.0, the contributions deposited in repositories, the advantages of open peer-review, the motivations for creating references which are not expressed as bibliographic citations and the enormous impact of the commercial search engines.

It is proposed that a new generation of Web indicators be used for wider, fairer and more feasible evaluation purposes. The aim is not only to improve evaluation, but also to support open access initiatives beyond the current definition to include all aspects of scholarly activity and access to data in addition to results.

We introduced a series of Web indicators classified in three large groups: activity, related to web presence and publication; impact, according to link visibility or the number of times a term is mentioned; and usage, counting visits and visitors and their behavior. Many of these indicators are collected from search engines, which assume an important role not only as intermediaries for data recovery but also as proxies for increasing visibility and impact. The current set of Webometrics indicators offers a wide range of opportunities for improving our knowledge of the academic system, how it is organized and works and to better monitor the persons and organizations involved.

References

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*, Vol. 57 No.10, pp. 1296-1302.

Aguillo, I.F., Ortega, J. L. and Fernández, M. (2008), "Webometric Ranking of World Universities: Introduction, Methodology, and Future Developments", *Higher Education in Europe*, Vol. 33 No 2/3, pp. 234-244.

Barjak, F. (2006), "Research productivity in the internet era", *Scientometrics*, Vol. 68 No. 3, pp. 343-360.

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.

Barjak, F. and Thelwall, M. (2008), "A statistical analysis of the web presences of European life sciences research teams", *Journal of the American Society for Information Science and Technology*, Vol. 59 No. 4, pp. 628-643.

Beel, J.and Gipp, B. (2008), "The potential of collaborative document evaluation for science", *Lecture Notes in Computer Science*, Vol. 5362, pp. 375-378.

Bornmann, L., Nast, I. and Daniel, H.-D. (2008), "Do editors and referees look for signs of scientific misconduct when reviewing manuscripts? A quantitative content analysis of studies that examined review criteria and reasons for accepting and rejecting manuscripts for publication", *Scientometrics*, Vol. 77 No. 3, pp. 415-432.

Brody, T., Harnad, S. and Carr, L. (2006), "Earlier web usage statistics as predictors of later citation impact", *Journal of the American Society for Information Science and Technology*, Vol. 57 No. 8, pp. 1060-1072.

Espadas,J., Calero, C. and Piattini, M. (2008), "Web site visibility evaluation", *Journal of the American Society for Information Science and Technology*, Vol. 59 No. 11, pp. 1727-1742.

Evans J.A. (2008), "Electronic publication and the narrowing of science and scholarship", *Science*, Vol. 321 No.5887, pp. 395-399.

Kaphingst, K., Zanfini, C.; Emmons, K. (2006), "Accessibility of Web Sites Containing Colorectal Cancer Information to Adults with Limited Literacy (United States)", *Cancer Causes and Control,* Vol. 17 No. 2, pp. 147-151.

Kousha, K. and Thelwall, M. (2007a), “The web impact of open access social science research”, *Library and Information Science Research*, Vol. 29 No. 4, pp. 495-507.

Kousha, K. and Thelwall, M. (2007b), “Google Scholar citations and Google Web/URL citations: A multi-discipline exploratory analysis”, *Journal of the American Society for Information Science and Technology*, Vol. 57 No. 6, pp. 1055-1065.

Li, X., Thelwall, M., Musgrove, P. and Wilkinson, D. (2005a), “National and international university departmental web site interlinking: Part 1, validation of departmental link analysis”, *Scientometrics*, Vol. 64 No. 2, pp. 151-185.

Li, X., Thelwall, M., Musgrove, P. and Wilkinson, D. (2005b), “National and international university departmental web site interlinking: Part 2, link patterns”, *Scientometrics*, Vol. 64 No. 2, pp. 187-208.

Mayr, P. (2006), “Constructing experimental indicators for open access documents”, *Research Evaluation*, Vol. 15 No. 2, pp. 127-132.

McInerney, C. and Bird, N. (2007), “Quantifying quality: Evolution of an instrument to assess website quality”, *Proceedings of the American Society for Information Science and Technology*, Vol. 43 No. 1, pp. 1-12.

Olsina, L., Godoy, D., Lafuente, G. and Rossi, G. (1999), “Assessing the quality of academic websites: a case study”, *New Review of Hypermedia and Multimedia*, Vol. 5, pp. 81-103.

Ortega, J.L. and Aguillo, I.F. (2008a), “Linking patterns in European Union countries: Geographical maps of the European academic web space”, *Journal of Information Science*, Vol. 34 No. 5, pp. 705-714.

Ortega, J. L. and Aguillo, I.F. (2008b), “Visualization of the Nordic Academic web: Link analysis using social network tools”, *Information Processing and Management*, Vol. 44 No 4, pp. 1624-1633.

Ortega, J.L., Aguillo, I.F., Cothey, V. and Scharnhorst, A. (2008), “Maps of the academic web in the European Higher Education Area - An exploration of visual web indicators”, *Scientometrics*, Vol. 74 No. 2, pp. 295-308.

Palmer, J.W. (2002), “Web site usability, design, and performance metrics”, *Information Systems Research*, Vol. 13 No. 2, pp. 151-167.

Park, H. and Thelwall, M. (2006), “Web science communication in the age of globalization: Links among universities’ websites in Asia and Europe”, *New Media & Society*, Vol. 8 No. 4, pp. 631-652.

Payne, N. and Thelwall, M. (2008), “Longitudinal trends in academic web links”, *Journal of Information Science*, Vol. 34 No. 1, pp. 3-14.

Petricek, V., Escher, T., Cox, I.J. and Margetts, H. (2006), “The Web Structure of E-Government - Developing a Methodology for Quantitative Evaluation”, paper presented at WWW2006, May

23–26, 2006, Edinburgh, Scotland, available at: http://www.adastral.ucl.ac.uk/~icox/papers/2006/WWW06.pdf (accessed 7 April 2009).

Ravid, G., Bar-Ilan, J., Baruchson-Arbib, S. and Rafaeli, S. (2007), “Popularity and findability through log analysis of search terms and queries: the case of a multilingual public service website”, *Journal of Information Science*, Vol. 33 No. 5, pp. 567-583.

Rousay, E., Fu, H., Robinson, J. M., Essex, J. W. and Frey, J. G. (2005), “Grid-based dynamic electronic publication: a case study using combined experiment and simulation studies of crown ethers at the air/water interface”, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, Vol. 363 No.1833, pp. 2075-2095.

Smith, A. (2007). “Issues in "blogmetrics" - case studies using BlogPulse to observe trends in weblogs”, in Torres-Salinas, D. and Moed, H.F. (Eds.), *Proceedings of ISSI 2007, 11th International Conference of the International Society for Scientometrics and Informetrics, CSIC, Madrid, Spain*, CINDOC-CSIC, Madrid, pp. 726-730.

Smith, R. (2006), “Peer review: A flawed process at the heart of science and journals”, *Journal of the Royal Society of Medicine*, Vol. 99 No. 4, pp. 178-182.

Stuart, D., Thelwall, M. and Harries, G. (2007), “UK academic web links and collaboration – an exploratory study”, *Journal of Information Science*, Vol. 33 No. 2, pp. 231-246.

Swan, A. (2007), “Open Access and the progress of science”, *American Scientist*, Vol. 95 No. 3, pp.198-200.

Tang, R. and Thelwall, M. (2008), “A Hyperlink Analysis of U.S. Public and Academic Libraries’ Web Sites”, *The Library Quarterly*, Vol. 78 No. 4, pp. 419–435.

Thelwall, M., Vaughan, L. and Björneborn, L. (2005), “Webometrics”, *Annual Review of Information Science and Technology*, Vol. 39, pp. 81-135.

Thelwall, M. (2003), “Can Google's PageRank be used to find the most important academic Web pages?”, *Journal of Documentation,* Vol. 59 No. 2, pp. 205-217.

Torres-Salinas, D., Cabezas-Clavijo, A. and Delgado-López-Cózar, E. (2008), “Análisis métrico de los blogs españoles de biblioteconomía y documentación (2006-2007)”, *El Profesional de la Información*, Vol. 17 No. 1, pp. 38-48.

Vakkari, P. (2008), “Perceived influence of the use of electronic information resources on scholarly work and publication productivity”, *Journal of the American Society for Information Science and Technology*, Vol. 59 No. 4, pp. 602-612.

Vaughan, L. and Thelwall, M. (2005), “A modeling approach to uncover hyperlink patterns: The case of Canadian universities”, *Information Processing & Management*, Vol. 41 No. 2, pp. 347-359.

Van de Sompel, H., Payette, S., Erickson, J., Lagoze, C. and Warner, S. (2004), "Rethinking Scholarly Communication. Building the System that Scholars Deserve", *D-Lib Magazine*, Vol. 10 No. 9.

Van Raan, A.F.J. (2005), “Fatal attraction: Conceptual and methodological problems in the ranking of universities by bibliometric methods”, *Scientometrics,* Vol. 62 No. 1, pp. 133-143.

Voss, J. (2005), “Measuring Wikipedia”, in Ingwersen, P. and Larsen, B. (Eds.), *Proceedings of ISSI 2005: the 10th International Conference of the International Society for Scientometrics and Informetrics, Stockholm, Sweden, July 24-28 2005*, Karolinska University Press, Stockholm, pp. 221-231, available at: http://eprints.rclis.org/archive/00003610/ (accessed 22 April 2009).

About the author

Isidro Aguillo is Head of Cybermetrics Lab, Centre of Social Sciences and Humanities (CCHS), Madrid, Spain. Isidro Aguillo can be contacted at: isidro.aguillo@cchs.csic.es