Jointly published by Akadémiai Kiadó, Budapest and Springer, Dordrecht

Scientometrics, Vol. 79, No. 1 (2009) 7–25
DOI: 10.1007/s11192-009-0401-7

The lifespan of "informetrics" on the Web: An eight year study (1998–2006)

JUDIT BAR-ILAN,[a] BLUMA C. PERITZ[b]

[a] *Department of Information Science, Bar-Ilan University, Ramat Gan, 52900, Israel*
[b] *Hebrew University of Jerusalem, Jerusalem, 91904, Israel*

The World Wide Web is growing at an enormous speed, and has become an indispensable source for information and research. New pages are constantly added, but there are additional processes as well: pages are moved or removed and/or their content changes. We report here the results of an eight year long project started in 1998, when multiple search engines were used to identify a set of pages containing the term *informetrics*. Data collection was repeated once a year for the last eight years (with the exception of 2000 and 2001) using both search engines and revisiting previously identified pages. The results show that the number of pages grew from 866 in 1998 to 28,914 in 2006 – a 33-fold growth. Besides the obvious growth of the topic on the Web, we observed both decay (pages disappearing from the Web) and modification. Even though most of the pages from 1998 either disappeared or ceased to contain the term *informetrics*, 165 pages (19.1%) still exist in 2006 and contain the search term. We followed the "fate" of these 165 pages: characterized the publishers, the contents and the changes that occurred the whole period. In recent years e-print servers and publishers' sites became sources of large number of pages related to *informetrics*. Longitudinal studies following the evolution of a topic on the Web are very important, since they provide insights about content and the underlying Web processes.

Introduction

The World Wide Web is continuously growing at an incredible speed both in terms of its content and in terms of the number of users accessing it. The Web has become an indispensable source for information and research. Its growth patterns are of interest for theoretical, technical, social and economic reasons.

The present study examined the evolution of *informetrics* on the Web. To be more specific, we identified Web pages containing either the term *informetrics* or *informetric*. Two complimentary data collection techniques were utilized: retrieving data from multiple search engines and revisiting Web pages identified at previous data collection points. The combination of the two techniques allowed us to study several evolution patterns: creation of new pages, removal of previously existing ones and modifications.

Received December 5, 2007

Address for correspondence:
JUDIT BAR-ILAN
E-mail: barilaj@mail.biu.ac.il

0138–9130/US $ 20.00

This is the first study that we are aware of that tracks the evolution of a topic on the Web for such a long period of time using multiple collection methods for data collection.

Literature review and background

Longitudinal studies of the Web

Several previous studies examined sites and pages for shorter periods of time, usually for several weeks or months. For example, BAR-ILAN & PERITZ [1999] studied search results retrieved for the query *informetrics OR informetric* once a month for six months. BREWINGTON & CYBENKO [2000] observed about 100,000 pages daily for a period of seven months. These pages were accessed by users through a service called "Informant". CHO & GARCIA-MOLINA [2000] selected 270 sites and crawled 3,000 pages from each site strating from the root. The sites were crawled every day for a period of four months. FETTERLY & AL. [2004] crawled a set of more than 150 million Web pages once a week for a period of eleven weeks. NTOULAS & AL. [2004] crawled 154 "popular" Web sites once a week for a year. KIM & LEE [2005] monitored between 34,000 Korean Web sites at two-days intervals for 100 days. From each site a maximum of 3,000 pages were downloaded each time starting from the root URL of each site. Note that in these shorter term studies, the data sets were usually huge and the monitored pages were visited often (typically once a week).

There are only a few studies that report findings based on several years of data collection, but even these are for shorter length than the current study. One of the longest studies to this day was carried out by KOEHLER [2004]. He observed a fixed set of pages 361 Web pages for 325 weeks (over six years). The original set was assumed to be "a random representation of the Web as a whole". Although only 122 of the pages were still accessible after six years, the author concludes that pages tend to stabilize as they become older. GOMES & SILVA [2006] had data on the Portuguese national Web for a period of three years (8 data collection points) and their conclusion was that the lifetimes of URLs and their content can be modelled as logarithmic functions. BAEZA-YATES & POBLETE [2003] based their results on three data collection points over a period of three years of the Chilean Web. Their conclusion is that although the Web keeps growing, a significant part of it disappears. BAR-ILAN & PERITZ [2004] report the results of a five-year long study. TOYODA & KITSUREGAWA [2006] had access to the Japanese Web archive which collects data about once a year, and based their results on data from 2003–2004 (three data collection points). They wanted to identify whether newly discovered Web pages are actually new or they already existed on the Web, "waiting to be found". ORTEGA & AL. [2006] crawled about a thousand sites twice, once in 1997 and once in 2004; their results show considerable growth of different types of

Web elements (e.g., images and links) over time. ROUSSEAU [1999] studied the changes in the number of search results reported by AltaVista and Northern Light on three queries for a period of 84 days.

All previous studies that we were able to locate used a single data collection method. They either monitored a fixed data set (e.g., [FETTERLY & AL., 2004] or [KOEHLER, 2004]) or crawled in a pre-specified manner a fixed number of pages from given starting points (e.g. [CHO & GARCIA-MOLINA, 2000] or [KIM & LEE, 2005]), or attempts were made to download complete Websites (e.g. [NTOULAS & AL., 2004]) and/or entire national Webs (e.g., [BAEZA-YATES & POBLETE, 2003] or [TOYODA & KITSUREGAWA, 2006]). KE & AL. [2006] survey a large number of studies in the area of Web dynamics.

Persistence of Web references

Web sources are being referenced in scholarly publications, both Web pages and sites and scholarly publications freely available on the Web. Web references, unlike references to printed sources can disappear or change. A number of studies discussed this issue and evaluated its extent.

One of the largest studies was carried out by LAWRENCE & AL. [2001], in which they tried to locate 67,577 URLs referenced in articles indexed by the Citeseer database (publications dates between 1993 and 1999). On the one hand, they found considerable increase in the number of Web-references over the years, but on the other hand they emphasized the lack of persistence of the Web-references – 54% of the Web-references published in 1994 were not accessible by 2000. SPINELLIS [2003] studied the availability of URLs mentioned in two computer science journals: Computer and Communications of the ACM – 72% of the URLs were retrieved without problems. SELLITTO [2005] analyzed the references of conference papers from the AusWeb conference series; on the average 45.8% of the references were not locatable by November 2003; even for the papers published on July 2003, 9% of the Web references were already missing.

CASSERLEY & BIRD [2003] studied 1,425 LIS research articles published in 1999 and 2000. They were able to find at the original URLs, 56.4% of the references of their sample of references. The percentage of accessible Web references was increased through searching Google and using the Internet Archive – altogether 89.4% of the Web references were located. MARKWELL & BROOKS [2003] complain about "link rot" as a limiting factor of Web-based references. In two years 20% of the URLs disappeared, moved or changed their content. MCCOWN & AL. [2005] analyzed the references in the D-lib Magazine, and found that about 30% of the sample of Web-references published between 1995 and 2004 failed to resolve by February 2005. TYLER & MCNEILL [2003] studied the longevity of 2,729 URLs that appeared in College & Research Libraries

News Web bibliographies. They conclude that the half-life of the URLs in these lists is about 5 years. They also located "undead" URLs – URLs that seemed to be "dead" at the initial check, but became "alive and well" when they rechecked these URLs six weeks later. In a most recent article GOH & NG, [N.D.] examined the extent of "link rot" in three leading IS journals for a sample of articles published in 1997–2003. The authors were unable to access 31% of the 2,516 Web citations extracted from these publications.

WREN [2004] studied URL references in PubMed abstracts and found that about 63% of them were accessible. In a recent paper WREN & AL. [2006] studied URL decay in dermatology journals and found a relatively high availability (81.7% – varying by publication year). However, most authors agreed that the unavailable URL content was important for the publication. NELSON & ALLEN [2002] monitored the persistence of 1,000 digital library objects accessible through the Web between November 2000 and December 2001. These items came from digital libraries containing freely accessible collections of scientific materials. The observed objects were reports, e-prints or re-prints of scientific and technical information. Only 3% of the sample disappeared during the time span. Thus, compared with the stability of general Web pages, Web-references seem to be much more stable, possibly because these Web references refer to higher quality content produced by more authoritative sources than the "average" Web page. Presumably, pages containing the term *informetrics* (in the bibliometric-scientometric sense) will be higher quality pages as well (it was previously found that a large number of these pages contain bibliographic references – see [BAR-ILAN, 2003]), thus we can expect slower decay of these URLs as well.

Methods

Data collection

The experiment started in January 1998. In the first stage (until June 1998) data was collected from the then major search engines (AltaVista, Excite, Hotbot, InfoSeek, Lycos and Northern Light) by running the query *informetrics OR informetric*. Originally we intended to run the query *informetics* only, but because of Northern Light's automatic stemming the query had to be extended. Data was collected once a month and changes between the data collected in consecutive data collection points were observed. In June 1998, 866 URLs were identified through the collective effort of the above-mentioned search engines. The query was chosen because we were looking for information on the scientific field *informetrics* – quantitative analysis of documents in all forms. However, as can be expected, on the Web *informetrics* has additional meanings as well (e.g., names of companies).

Search results fluctuated considerably between the data collection points, thus when rerunning the experiment in June 1999, an additional data collection method was employed besides querying the search engines. The URLs that satisfied the query in June 1998 were revisited in 1999 even if they were not located by the search engines in 1999. No data was collected in 2000 and in 2001. However, in retrospect this has not been a shortcoming of the research, since the growth and modification patterns can be easily interpolated for the missing data collection points (see Figure 1).

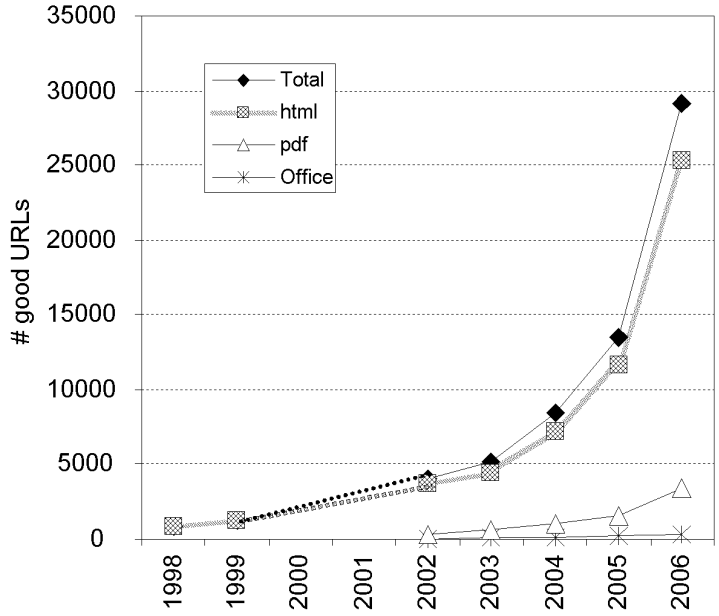

Figure 1. Growth curves for the different document types (growth curves interpolated for 2000 and 2001)

Thus in June 1999, 2002, 2003, 2004, 2005 and 2006 two separate data collection procedures were employed

1. Submitting the query *informetrics OR informetric* to the largest search engines at the time

   a) In 1999 the same search engines were used as in 1998, namely AltaVista, Excite, Hotbot, InfoSeek, Lycos and Northern Light

b) In 2002 and 2003 AllTheWeb, AltaVista, Google, HotBot, Teoma and Wisenut were employed. By 2002 search engines started to retrieve non-html pages as well (pdf, ps, doc, etc.)
c) In 2004, we queried AllTheWeb, AltaVista, Gigablast, Google, Hotbot, Teoma, Yahoo and Wisenut. Note that in June 2004, AllTheWeb and AltaVista still retrieved slightly different results from the then newly launched Yahoo search engine; and Hotbot served a different set of results as well.
d) In 2005 and 2006, Exalead, Google, MSN, Teoma (Ask) and Yahoo were queried.

Although the initial data set was rather small (less than 900 URLs), enormous growth was witnessed during the years, and in 2006 the search engines retrieved 24,272 different URLs (4,642 additional URLs were located through the "revisit" process in 2006).

Search engines limit the number of displayed result for a query (the limitations as of June 2006 were: 1000 for Google, Yahoo, 2000 for Exalead, 250 for MSN and 200 for Teoma). In order to try to overcome these limitations we used several techniques:

a) Including/excluding additional search terms (e.g. *informetrics-scientometrics* and *informetrics scientometrics* – these two queries together are supposed to cover all the pages indexed by the search engine that contain the word *informetrics*)
b) Limiting the query by site or filetype e.g. *informetrics site:.es –filetype:pdf* (non-pdf pages from Spain only) – including/excluding sites or filetypes.
c) Limiting the query by date using the betweendate feature of Ask, or using AltaVista's advanced search (now powered by Yahoo)

In one case we included/excluded 22 additional terms in order to break down the query results into small enough chunks.

The whole set of searches on all the search engines were run within 1–2 hours to minimize the effect of time on the results. Each year the searches were carried out in June. The URLs were extracted from the search results pages and duplicates (usually the same URL retrieved by several search engines) were eliminated. The URLs were compared as text strings, thus, for example, informetrics.com and www.informetrics.com were considered two different URLs.

All the documents residing at the identified URLs were downloaded to our local computer within 0–2 days of the searches, in order to minimize the effect of the time elapsed between the search time and download time on the possible changes that the documents undergo over time. A second attempt was made to download inaccessible URLs. Finally, the entire set of html documents was tested for the presence of the string *informetric*.

2) All pages that contained either the term *informetrics* or the term *informetric* (i.e., satisfied the query) at least at the first time that they were identified by the search process were revisited at each of the later data collection points.

The combination of the two methods allowed us both to follow the "fate" of previously identified pages and to enrich the collection of pages with newly retrieved ones from the search engines. Note that newly retrieved pages are not necessarily newly created pages. It is possible that the page existed before and was indexed by some of the search engines, but it did not contain the search term; or because of the incomplete coverage of the Web by the search engines, it is quite plausible that the page existed for a long time and was relevant to the search but was only discovered at one of the later data collection points. We are not aware of any other study that utilized multiple data collection methods for studying the evolution of a topic on the Web.

Data analysis

We analyzed the longitudinal patterns of the data set in general and the changes in the distribution of the domains over time. Our basic notion, *technical relevance*, defines whether a URL satisfies the query at a certain time. All html, text and office documents and a sample of the other document types (pdf and postscript) were checked for the presence of at least one of the search terms. We decided to use the terminology *technical relevance* instead of the more widely used term *relevance* in order to avoid the complex issues of defining relevance (see for example SARACEVIC, [1998] or MIZZARO, 1998]).

- All the URLs were checked at each *data check point* (in June of 1998, 1999, 2002–2006) for technical relevance (*trel* in short). Note that it is quite possible that the URL is accessible at a data check point, however its content was *modified* and the page ceases to be technically relevant to the query.
- Some of the documents were not accessible at the data check points. This inaccessibility could be temporary (*intermittent* URLs caused mainly by communication or server problems) or permanent. If from some point onwards the URL was never accessible then we conclude that the URL *disappeared*. Note that it is possible that a URL defined as disappeared based on the available data, will become intermittent if the data monitoring continues for a longer time.

Limitation of the data collection methods

It is well-known that search engines do not cover the whole Web (see for example [BHARAT & BRODER, 1998] or [LAWRENCE & GILES, 1998]). In order to increase coverage, several search engines were used, but even the combined coverage of search

engines is far from being complete [LAWRENCE & GILES, 1999]. In any case, search engines cannot access large parts of the so-called deep or invisible Web – mostly information that is retrieved upon request from databases with interfaces to the Web [BERGMAN, 2001]. Recently, Google started to index materials from the deep Web as well, but still its coverage is limited to the publishers' sites with whom Google has special agreements. Because of the shortcomings of the commercial search engines [BAR-ILAN, 2005] we decided to employ the second data collection method – revisiting previously discovered pages. Thus we made a huge effort to reach all relevant pages, but it is quite clear that we could not have been fully successful.

Another limitation of the study that the URLs were compared as stings, thus for example www.yahoo.com and yahoo.com were considered different URLs. In addition mirror sites and other forms of duplication were not detected. Inspection of the URLs showed that the extent of several highly similar URLs pointing to the same actual page, except for a few cases, is not high. No attempt was made to identify content duplication, i.e., same content residing at different URLs.

Results and discussion

Growth, disappearance and modification

During the whole period 36,282 different URLs were identified that satisfied the query at least at the first time they were located. Table 1 and Figure 1 decribe the overall growth of the topic over the years as reflected by the number of *technically relevant* URLs identified at each of the data collection points. The growth over the years is considerable, 80.2% of the total unique URLs identified during the whole period that satisfied the query were located at the last data check point (2006), while only 2.3% of the total were discovered by the search engines in 1998. When analyzing the data we have to take into account two processes: the growth of the Web as a whole, and changes in coverage of the search engines.

Growth is the definitely the strongest of all three processes: growth, decay and modification, however the other processes are considerable as well. Table 2 displays the changes that occurred to the original set of 866 URLs over time. We observe that in 2006, 165 (155 *trel* + 9 intermittent) documents out of the original set of 866 documents were still accessible and still satisfied the query. Table 3 displays the same data for the set of *trel* URLs located in 1999.

Table 1. Number of technically relevant (*trel*) URLs identified at the *data check points*

	Total *trel* URLs (% out of total)	*Trel* html or text documents	*Trel* pdf documents	*Trel* MS Office documents	*Trel* postscript documents	*Trel* xml documents
1998	866 (2.4%)	866	0	0	0	0
1999	1,249 (3.3%)	1,249	0	0	0	0
2002	4,034 (11.1%)	3,705	272	31	26	0
2003	5,176 (14.3%)	4,399	625	92	60	0
2004	8,454 (23.3%)	7,225	1,027	140	62	0
2005	13,454 (37.1%)	11,594	1,577	210	73	0
2006	28,914 (80.2%)	25,358	3,349	310	63	18
Total unique URLs during whole period	36,282	31,999	3,839	360	84	18

Table 2. The set of 866 URLs located in 1998 at the different data check points

	1999	2002	2003	2004	2005	2006
trel	648	291	242	216	176	156
intermittent			1	3	7	9
inaccessible/disappeared	183	495	551	575	615	629
term not in document	35	80	71	72	68	72

Table 3. The set of 1249 URLs located in 1999 at the different data check points

	2002	2003	2004	2005	2006
trel	509	408	362	304	266
intermittent		2	3	9	12
inaccessible/disappeared	620	742	784	849	882
term not in document	120	97	100	87	89

Tables 2 shows that after seven years (in 2005) only 20% of the original URLs from 1998 still existed and were relevant to the query; and for the 1999 dataset, after seven years (2006) the percentage of *trel* documents was almost the same (22%). This similarity is not surprising since the 1249 documents of 1999 include the 648 *trel* documents first discovered in 1998 (only those that were still *trel* by 1999). When considering the 601 documents first discovered in 1999 we find that the percentage of documents still *trel* in 2006 is lower, only 113 (18.8%) such documents were retrieved.

Page distribution on domains

In order to gain a better understanding of the document types, we tabulated the ten most "prolific" domains (i.e., domains with the largest numbers of *trel* pages that were located in each year) for each year for which data was collected. In this paper we define *domain* as a second level domain, where the full domain or host is the part of the URL immediately after "http://" and before the next "/", i.e. for the URL *http://www.cindoc.csic.es/cybermetrics/issues.html*, this would be *www.cindoc.csic.es*.

The second-level domain is the last two parts of the full domain name, i.e. *csic.es* in our example. This means that the previous URL and *http://internetlab.cindoc.csic.es/articulos.asp?art=148&offset=0* belong to the same second level domain – *csic.es* the Spanish National Research Council. The reason for considering second-level domains is that these usually represent the publishing organization. In some countries, e.g., the UK, Australia and Israel, the second level domains (e.g., ac.uk, edu.au or co.il) have the same role as the top-level domains in the USA (e.g., com, edu, org), and in these cases, we defined the domain as the third level domain, e.g., both *http://sistm.web.unsw.edu.au/conference/issi2001/index.html* and *http://birg.web.unsw.edu.au/text/about.htm* belong to the domain *unsw.edu.au* (the University of New South Wales).

Note that there are several different definitions of Internet domains; here we follow the above definition, which in our opinion provides the best information on the major publishers of information on informetrics on the Web. The results are presented in Tables 4–10.

Table 4. The most "prolific" domains in 1998 – number and percentage of html pages (N = 866)

	Site	No. pages	% pages
1	db.dk	54	6.2%
2	sfu.ca	44	5.1%
3	crrm.univ-mrs.fr	24	2.8%
4	bubl.ac.uk	22	2.5%
5	hu-berlin.de	22	2.5%
6	ust.hk	21	2.4%
7	informetric.com	20	2.3%
8	uni-trier.de	18	2.1%
9	rwth-aachen.de	18	2.1%
10	csic.es	15	1.7%
	Total	258	29.8%

The top ten domains covered 29.8% of the pages. Interesting to note that with the exception of informetric.com (a software company, not related to *informetrics* in the scientific sense) all the other domains are from outside the US, most of them in Europe. The Danish pages (from db.dk) are from the Royal School of Library and Information Science. The pages from Simon Fraser University (sfu.ca) are part of an early electronic library project (no longer active), providing lists of journal titles. The French server (crrm.univ-mrs.fr) is not functional anymore, but it had information on the ISSI Society and its conferences. The pages from BUBL Information Service – an Internet based information service for the UK higher education community (bubl.ac.uk) – served abstracts and tables of contents of journals. The pages from Humboldt University (hu-berlin.de) mostly contained information on the curriculum for the Library Science studies. The pages from ust.hk, uni-trier.de and rwth-aaschen.de are mirror pages of the DPLP server (Computer Science Bibliography) that contained the terms *informetric* or *informetrics*. The DBLP project provides bibliographic information

on major computer science journals and proceedings (http://www.informatik.uni-trier.de/~ley/db/welcome.html). Finally, csic.es hosts the journal Cybermetrics – International Journal of Scientometrics, Informetrics and Bibliometrics (http://www.cindoc.csic.es/cybermetrics/). In 1998 the journal had only published its first volume and issue (a single paper).

Table 5 displays the ten most frequently appearing domains in 1999. Seven domain in the list were among the top-ranking sites in 1998 as well. Note that csic.es was tenth in 1998, but by 1999 it is already ranked number 3. The "newcomers" are informetric.co.uk – a company providing instrumentation and software solutions for the measurement, capture, analysis and modelling of water and environmental data; leidenuniv.nl – the site of the Centre for Science and Technology Studies (CWTS) and uwo.ca – University of Western Ontario, where the huge majority of the pages originated from the Faculty of Media and Information Studies.

Table 5. The most "prolific" domains in 1999 – number and percentage of html pages (N = 1,249)

	Server	No. pages	% pages
1	db.dk	79	6.3%
2	sfu.ca	43	3.4%
3	csic.es	33	2.6%
4	bubl.ac.uk	30	2.4%
5	crrm.univ-mrs.fr	26	2.1%
6	informetric.com	25	2.0%
7	informetric.co.uk	24	1.9%
8	informatik.uni-trier.de	21	1.7%
9	leidenuniv.nl	20	1.6%
10	uwo.ca	20	1.6%
	Total	321	25.7%

Table 6. The most "prolific" domains in 2002 – number and percentage of html pages (N = 3,705)

	Server	No. pages	% pages
1	unsw.edu.au	136	3.7%
2	db.dk	128	3.5%
3	csic.es	74	2.0%
4	uni-trier.de	72	1.9%
5	collnet	71	1.9%
6	citeseer	62	1.7%
7	hu-berlin.de	54	1.5%
8	acm.org	46	1.2%
9	enssib.fr	46	1.2%
10	utk.edu	45	1.2%
	Total	734	19.2%

Comparing Tables 4, 5 and 6, there are some striking differences. The top domain, unsw.edu.au is a "newcomer" in the list (only 4 and 11 pages from this domain were located in 1998 and 1999 respectively). Most of the captured pages relate to the 2001 ISSI Conference in Sydney, hosted by the University of New South Wales. The Collnet site, a global interdisciplinary research network for the study of all aspects of

collaboration in science and technology (http://www.collnet.de/) provides information on the group members, most of them ISSI members. Collnet was established in 2000, thus it is not surprising that pages were not located in 1998 and 1999. Citeseer is an extensive digital library in computer science. Citeseer was developed in 1997; however no pages from it were located in 1998 and 1999. ACM is the Association for Computer Machinery, the acm.org domain also serves as a DBLP mirror site and hosts the SIGIR newsletter that sometimes relates to *informetrics*. The domain enssib.fr hosts the pages of SLISNET (Schools of Library and Information Science Network – http://www.enssib.fr/autres-sites/SLISNET/) that provides information on school curricula and on research interests of the research staff of the participating institutions. Discussions of the SIGMETRICS group are archived on listserv.utk.edu. This discussion list covers bibliometrics, scientometrics and informetrics and is a Virtual-SIG of ASIST (http://web.utk.edu/~gwhitney/sigmetrics.html). The list was established in June 1999, thus no pages from the listserv were retrieved in 1998 and 1999.

Table 7 lists the most frequently occurring domains in 2003. We see a significant increase in the number of pages from Citeseer and a decrease in the number of pages from the Danish server db.dk. The new domain in the list is upenn.edu that hosts Eugene Garfield's extensive site – most of the pages retrieved from www.garfield.library.upenn.edu came from the HistCite subsite.

Table 7. The most "prolific" domains in 2003 – number and percentage of html pages (N = 4,399)

	Server	No. pages	% pages
1	unsw.edu.au	133	3.0%
2	citeseer	107	2.4%
3	db.dk	84	1.9%
4	uni-trier.de	83	1.9%
5	csic.es	75	1.7%
6	acm.org	69	1.6%
7	upenn.edu	62	1.4%
8	hu-berlin.de	61	1.4%
9	utk.edu	52	1.2%
10	collnet	48	1.1%
	Total	774	17.6%

The "newcomers" in 2004 are the University of Arizona (arizona.edu), which hosts DLIST – the Digital Library of Information Science and Technology, an open access digital archive established in 2002; and ASIST (asis.org) with different announcements related to informetrics.

Table 8. The most "prolific" domains in 2004 – number and percentage of html pages (N=7,225)

	Server	No. pages	% pages
1	upenn.edu	340	4.7%
2	csic.es	269	3.7%
3	acm.org	258	3.6%
4	unsw.edu.au	158	2.2%
5	citeseer	135	1.9%
6	uni-trier.de	104	1.4%
7	db.dk	103	1.4%
8	arizona.edu	85	1.2%
9	hu-berlin.de	81	1.1%
10	asis.org	67	0.9%
	Total	1600	22.1%

In 2005, for the first time we see a blogger site (blogspot.com) among the top-ten domains. Blogspot.com appears among the top-ten domains as a result of the combined effort of 20 bloggers (http://www.blogger.com/start). The most prolific author is Francisco Javier Martínez Méndez (http://irsweb.blogspot.com/) who maintains a Spanish blog on Web information retrieval, and at the time of data collection had a link on the sidebar of his blog pages linking to the online copy of the Egghe & Rousseau book on Informetrics. As of September 2007 this link does not appear on the sidebar of the blog, but in June 2005, 112 technically relevant blog pages were located. The second most prolific author (with 90 pages located in June 2005) is Pedro Principe, who maintains the Portuguese blog Pedro Principe Rato de Biblioteca (http://ratodebiblioteca.blogspot.com/) and linked from the sidebar of his blog to the Cybermetrics journal (this link does not exist any more as of September 2007). It is interesting to note the major influence blogs have on the number of pages and links related to specific topics.

The two other "newcomers" are digital repositories: E-LIS (eprints.rclis.org) is an open access archive in librarianship, information science and technology, and doc.luc.ac.be hosts the repository of research papers of the Hasselt University (formerly Limburgs Universitair Centrum, thus the *luc* in the URL – by 2006 the repository was found both under doc.luc.ac.be and doc.uhasselt.be). The repository includes publications of Leo Egghe and Ronald Rousseau under the category "informetrics" (by September 2007 the name of the category changed to "bibliometrics"). The repository has interfaces in four languages: Dutch, English, French and German. The search engines often retrieved the same content displayed in the different languages, causing considerable content duplication.

Table 9. The most "prolific" domains in 2005 – number and percentage of html pages (N=11,594)

	Server	No. pages	% pages
1	upenn.edu	864	7.5%
2	csic.es	798	6.9%
3	acm.org	312	2.7%
4	rclis.org	254	2.2%
5	blogspot.com	230	2.0%
6	unsw.edu.au	152	1.3%
7	uni-trier.de	150	1.3%
8	arizona.edu	130	1.1%
9	informetric.com	122	1.1%
10	luc.ac.be	121	1.0%
	Total	3133	27.0%

Table 10. The most "prolific" domains in 2006 – number and percentage of html pages (N=25,358)

	Server	No. pages	% pages
1	csic.es	1,812	7.1%
2	uhasselt.be	1,351	5.3%
3	acm.org	1,102	4.3%
4	upenn.edu	1,038	4.1%
5	luc.ac.be	996	3.9%
6	elsevier.com	881	3.5%
7	utk.edu	758	3.0%
8	rclis.org	591	2.3%
9	blogspot.com	456	1.8%
10	citeseer	333	1.3%
	Total	9,318	36.7%

There is a huge increase in the number of pages from the csic.es site over time, which is mainly explained by the number of pages from the Cybermetrics journal located in 2006. Google lately changed its indexing policy – it currently indexes publishers' sites, even if access to the publication is fee or subscription based; if the publisher provides free access at least to the abstract of the article. This is the reason for the large number of pages from the ACM Digital Library (http://portal.acm.org/dl.cfm) and from Elsevier.

When we compare Table 4 with Table 10, we see considerably differences in terms of the content producers. In the top list of 1998 (Table 4), five domains provided mainly bibliographic information (Simon Fraser University's Electronic Library Project, the BUBL Information Service, and the three sites of the DBLP server), two domains were sites of departments/schools of information science (Institute of Library and Information Science at the Humboldt University and the Royal School of Library and Information Science), and one domain each of an e-journal (Cybermetrics), ISSI society information (crrm.univ-mrs.fr) and a commercial company named informetric.com, not related to informetrics in the scientific sense.

By 2006 (Table 10) the list contains four digital libraries/repositories (the repository at the University of Hasselt – under two different domain names, E-LIS and Citeseer), three publisher's domains (ACM, Elsevier and Cybermetrics, although the

Cybermetrics site also provides considerable bibliographical information). Additional contributing sites are Eugene Garfield's site containing considerable bibliographical and full text information, although the huge majority of the pages retrieved by the search engines are from HistCite, providing bibliographic data; the archives listserv of the ASIST SIG on metrics (SIGMETRICS – http://listserv.utk.edu/archives/sigmetrics.html) and the blogger site.

Thus we see a shift away from bibliographical data only towards providing full text. Note that although access to Elsevier and ACM requires subscription, the other repositories and Cybermetrics are open access.

Another interesting point is the concentration of information related to informetrics on a smaller number of servers over time. Tables 4–10 provided data on the domains, while now we provide data on the distribution of pages on hosts, where the host is the is the part of the URL immediately after "http://" and before the next "/". In Figure 2, the hosts identified (356 hostsin 1998 ; 1211 in 2002 and 4,850 in 2006) are arranged in decreasing order of the number of *trel* pages residing on them. In 1998, the top 10% of the hosts (35.6 hosts) covered 12.3% of the total pages located in that year, while in 2006, 0.1% of the hosts (4.85 hosts) covered 20.1% of the *trel* pages identified at the last data collection point.

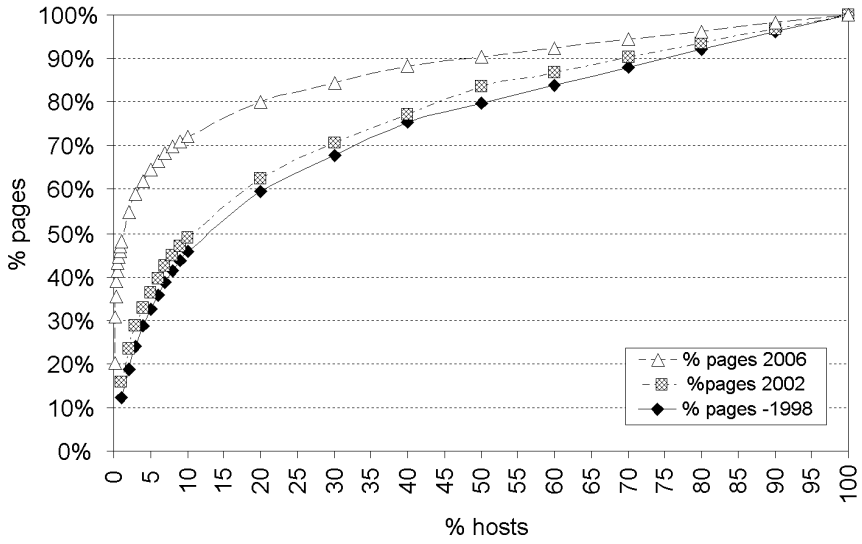

Figure 2. Concentration of pages on hosts in 1998, 2002 and in 2006

The persistent set of URLs from 1998

There were 165 URLs that were accessed and technically relevant at each of the data check points. We decided to analyze the contents of these pages (see [KRIPPENDORFF, 2003] or [NEUDORF, 2001]) and to characterize the modifications that occurred to these pages over time. However after a closer examination of these pages we discovered that a number of these pages were duplicates (from mirror sites and/or from alternative URLs). After the removal of the duplicates, the set was comprised of 97 URLs. We call this set the *persistent set* of URLs. For each URL we identified the publisher, the page type, the context in which *informetrics* was mentioned on the page and the type and frequency of modifications. In Table 11 the publishers with the largest number of pages in this set are displayed.

Table 11. The publishers with the largest number of pages in the *persistent set*

Publisher	No. pages (% of total)	Comments
Cybermetrics	12 (12.4%)	The electronic journal, Cybermetrics – International Journal for Scientometrics, Informetrics and Bibliometrics
DBLP	11 (11.3%)	Computer Science bibliography, publication lists of authors and TOCs of journals/proceedings
Hebrew University	8 (8.2%)	Hosted the 1997 ISSI Conference in Jerusalem
Humboldt University	8 (8.2%)	Study curriculum, pages of the Society for Science Studies, personal pages and list of events
Simon Fraser U., Rob Cameron	6 (6.2%)	An early project – electronic library in computer science
SLISNET	6 (6.2%)	Schools of Library and Information Science Networks, pages on participating institutions, study curriculum and research interests
Danmarks Biblioteksskole	5 (5.2%)	Newsletters and conference announcements from the Royal School of Library and Information Science, Denmark
Informetric company	4 (4.1%)	Pages of the Informetric System Inc. – a software company
SIGIRLIST	4 (4.1%)	Newsletters of the Special Interest Group on Information Retrieval
CSNA	3 (3.1%)	Newsletters of the Classification Society of North America
Total	67 (69.1%)	

Table 12 provides information on the different pages types, while Table 13 displays the distribution of contexts in which the term *informetrics* appeared on the pages of the *persistent set*. Note that the term may occur more than once on the page and it is quite possible that the different occurrences are categorized under different contexts. Thus the total for contexts is 127 and not 97 (the size of the *persistent set*).

For most of the pages in the *persistent set* (50 pages, 51.5%) no changes were observed at any of the data check points. Twenty seven pages (27.8%) were updated during the period: five pages were updated only once, six pages twice and the rest three or more times (out of the six data check points). For some pages (16 pages, 16.5%) only the formatting changed, such changes were mostly observed once or twice for this subset (for eleven out of the sixteen pages). Finally no content or format changes were

observed on four pages, only the "data of last update" was updated. Thus the set of *persistent* pages were not modified considerably during the period, some of these pages were seemingly "forgotten" or "abandoned".

Table 12. The distribution of page types in the *persistent set*

Page type	No. pages	% pages
Newsletter/news item	13	13.4%
Publication list	11	11.3%
Conference page	10	10.3%
Content page	8	8.2%
Journal list	7	7.2%
Course list	6	6.2%
Report/article	6	6.2%
TOC	5	5.2%
Company page	4	4.1%
Course page	4	4.1%
Faculty list/ list of researchers	4	4.1%
Abstract/list of abstracts	3	3.1%
Homepage	3	3.1%
Event list	2	2.1%
List of institutions	2	2.1%
Other list	8	8.2%
Other	1	1.0%

Table 13. The distribution of contexts in which of the term *informetrics* used in the *persistent set*

Use of *informetrics* on page	No. occurrences	% occurrences (out of 127)
Publication/presentation title	29	22.8%
Cybermetrics	21	16.5%
ISSI	15	11.8%
Topic discussed/expanded	15	11.8%
Research interest/area	12	9.4%
Conference topic	8	6.3%
ISSI proceedings	8	6.3%
Name of institute/company	8	6.3%
Other	5	3.9%
Course topic	4	3.1%
Affiliation	2	1.6%

Conclusion

In this study we followed the evolution of a topic on the Web for a period of eight years. This is the longest systematic longitudinal study that we are aware of. The use to two complimentary data collection methods: retrieval from search engines and revisiting of previously existing pages allowed us to study the growth, decay and modification processes of pages that contain the term *informetrics*. A structured and detailed analysis of the content and page distribution allowed us to obtain a more qualitative understanding of the evolution of pages on the Web, in our case for the term and field of *informetrics*. Although such studies are extremely labor intensive, it is

highly recommended to conduct additional ones. Since the Web has become during a relatively short period of time a major source not only for information but also for research and we rely more and more on this source, in depth longitudinal studies provide vital information on its coverage and changes over time. They can help in understanding the role of the Internet on the overall development of a topic or a field.

References

BAR-ILAN, J. (2000), The Web as information source on informetrics? – A content analysis. *Journal of the American Society for Information Science*, 51 (5) : 432–443.

BAR-ILAN, J., PERITZ, B. C. (1999), The life-span of a specific topic on the Web – The case of "informetrics": A quantitative analysis. *Scientometrics*, 46 : 371–382.

BAR-ILAN, J., PERITZ, B. C. (2004), Evolution, continuity and disappearance of documents on a specific topic on the Web – A longitudinal study of "informetrics". *Journal of the American Society for Information Science and Technology*, 56 : 980–990.

BAEZA-YATES, R., POBLETE, B. (2003), Evolution of the Chilean Web structure composition. In: *Proceedings of the First Latin American Web Congress (LA-WEB 2003),* Retrieved November 12, 2006 from: http://www.la-web.org/2003/stamped/02_baeza-yates-poblete.pdf

BERGMAN, M. K. (2001), The deep Web: Surfacing hidden value. *Journal of Electronic Publishing*, 7 (1), Retrieved September 12, 2007, from http://www.press.umich.edu/jep/07-01/bergman.html

BHARAT, K., BRODER, A. (1998), A technique for measuring the relative size and overlap of public Web search engines. *Computer Networks and ISDN Systems,* 30 : 379–388.

BREWINGTON, B. E., CYBENKO, G. (2000), How dynamic is the Web? *Computer Networks,* 33 : 257–276.

CASSERLY, M. F., BIRD, J. E. (2003), Web citation availability: analysis and implications for scholarship, *College and Research Libraries*, 64 (7) : 300–317.

CHO, J., GARCIA-MOLINA, H. (2000), The Evolution of the Web and implications for an incremental crawler. In: *Proceedings of 26th International Conference on Very Large Databases (VLDB)*, September 2000, (pp. 200–210).

FETTERLY, D., MANASSE, M., NAJORK, M., WIENER, J. L. (2004), A large scale study of the evolution of Web pages. *Software – Practice and Experience*, 34 : 213–237.

GOH, D. H., NG, P. K. (NO DATE), Link decay in leading information science journals. To appear in *JASIST*. Retrieved November 17, 2006 from: http://www3.interscience.wiley.com/cgi-bin/fulltext/113452914/HTMLSTART

GOMES, D., SILVA, M. J. (2006), Modeling information persistence on the Web. In: *Proceedings of the 6th International Conference on Web Engineering (ICWE06)*, (pp.193–200).

KE, Y., DENG, L., NG, W., LEE, D. L. (2006), Web dynamics and their ramifications for the development of Web search engines. *Computer Networks*, 50 : 1430–1447.

KIM, S. J., LEE, S. H.(2005), An empirical study on the change of Web pages. In: *Proceedings of APWeb 2005, LNCS 3399,* (pp. 632–642).

KOEHLER, W. (2004), A longitudinal study of Web pages continued: A report after six years. *Information Research*, 9 (2) paper 174. Retrieved November 12, 2006 from: http://InformationR.net/ir/9-2/paper174.html

KRIPPENDORFF, K. (2003), *Content Analysis: An Introduction to Its Methodology*. 2nd edition. Sage Publications.

LAWRENCE, S., GILES, C. L. (1998), Searching the World Wide Web. *Science*, 280 (5360) : 98–100.

LAWRENCE, S., GILES, C. L. (1999), Accessibility of information on the Web. *Nature*, 400 : 107–109.

LAWRENCE, S., PENNOCK, D. M., KROVETZ, R., COETZEE, F. M., GLOVER, E., NIELSEN, F. A., GILES, L. E. (2001), Persistence of Web references in scientific research. *Computer*, 34 (2) : 26–31.

MARKWELL, J., BROOKS, D. W. (2003), "Link rot" limits the usefulness of Web-based educational material in biochemistry and molecular biology. *Biochemistry and Molecular Biology Education*, 31 (1) : 69–72.

MCCOWN, F., CHAN, S., NELSON, M. L., BOLLEN, J. (2005), The availability and persistence of Web references in D-Lib Magazine. *5th International Web Archiving Workshop (IWAW05), Vienna, Austria*. Retrieved November 12, 2006 from: http://arxiv.org/ftp/cs/papers/0511/0511077.pdf

MIZZARO, S. (1998), How many relevances in information retrieval? *Interacting with Computers*, 10 (1998) : 305–322. Retrieved November 12, 2006 from: http://www.dimi.uniud.it/mizzaro/research/papers/IwC.pdf

NELSON, M. L., ALLEN, B. D. (2002), Object persistence and availability in digital libraries. *D-Lib Magazine*, 8 (1). November 12, 2006 from: http://www.dlib.org/dlib/january02/nelson/01nelson.html

NEUDORF, K. A. (2001), *The Content Analysis Guidebook*. Sage Publications.

NTOULAS, A., CHO, J., OLSTON, C.(2004), What's new on the Web? The evolution of the Web from a search engine perspective. In: *Proceedings of the World-Wide Web Conference (WWW)*, May 2004, (pp. 1–12).

ORTEGA, J. L., AGUILLO, I., PRIETO, J. (2006), A longitudinal study of content and elements in scientific Web environment. *Journal of Information Science*, 32 : 344–351.

ROUSSEAU, R. (1999), Daily time series of common single word searches in AltaVista and Northern Light. *Cybermetrics*, 2/3 (1), paper 2. Retrieved November 12, 2006 from: http://www.cindoc.csic.es/cybermetrics/articles/v2i1p2.html

SARACEVIC, T. (1998), Relevance reconsidered. In: *Proceedings of the Second Conference on Conceptions of Library and Information Science (CoLIS 2)*, Copenhagen, Denmark (pp. 201–218).

SELLITTO, C. (2005), The impact of impermanent Web-located citations: A study of 123 scholarly conference publications. *Journal of the American Society for Information Science and Technology,* 56 (7) : 695–703.

SPINELLIS, D. (2003), The decay and failures of URL references. *Communications of the ACM*, 46 (1) : 71–77.

TOYODA, M., KITSUREGAWA, M. (2006) What's really new on the Web? Identifying new pages from a series of unstable web snapshots. In: *Proceedings of WWW2006 (2006)*, (pp. 233–241).

TYLER, D. C., MCNEIL, B. (2003), Librarians and link rot: A comparative analysis with some methodological considerations. *Portal: Libraries and the Academy*, 3 (4) : 615–632.

WREN, J. D. (2004), 404 not found: The stability and persistence of URLs published in Medline. *Bioinformatics*, 20 (5) : 668–672.

WREN, J. D., JOHNSON, K. R., CROCKETT, D. M., HEILIG, L. F., SCHILLING, L. M., DELLAVALLE, R. P. (2006), Uniform Resource Locator decay in dermatology journals. *Archives of Dermatology*, 142 : 1147–1152.