Chapter 15

METHODOLOGICAL ISSUES OF WEBOMETRIC STUDIES

Peter Ingwersen and Lennart Björneborn

Department of Information Studies, Royal School of Library and Information Science, Copenhagen, Denmark. E-mail: pi@db.dk

Abstract: The contribution defines webometrics within the framework of informetric studies, bibliometrics, and scientometrics as belonging to library and information science, and associated with cybermetrics as a generic sub-field. It outlines a consistent and detailed link typology and terminology and makes explicit the distinction between the web node levels when using the proposed terminological structures. Secondly, the contribution presents the meaning, methodology and problematic issues of the central webometric analysis types, i.e., Web engine and crawler coverage, quality and sampling issues. It discusses briefly Web Impact Factor and other link analyses. The contribution finally looks into log studies of humanWeb interaction.

1. INTRODUCTION: COLLECTION METHODOLOGY FOR WEBOMETRIC DATA

Since the mid-1990s escalating efforts have been made to study the nature of the World Wide Web, named the Web in this article, by applying modern informetric methodologies to its space of contents, link structures, and search engines. Studies of the Web have been named 'webometrics' by Almind and Ingwersen (1997) or 'cybermetrics', as in the electronic journal of that name[1]. This contribution points to research methods applied to

[1] http://www.cindoc.csic.es/cybermetrics/

H.F. Moed et al. (eds.), Handbook of Quantitative Science and Technology Research, 339-369.

selected areas of webometric investigations. These areas are search engine and Web crawler coverage, quality and sampling issues; link analyses, including Web impact analysis, and log studies of Web interaction behavior. The contribution is not an exhaustive review, but rather a view of the specialty.

Webometrics displays several similarities to informetric and scientometric studies as well as the application of common bibliometric and informetric methods. For instance, simplistic counts and content analysis of web pages can indeed be seen as analogous to traditional publication analysis; counts and analyses of outgoing links from web pages, here named outlinks, and of links pointing to web pages, called inlinks, can be seen as somehow similar to citation analyses. Outlinks and inlinks are then regarded like references and citations, respectively, in scientific articles. However, since the Web consists of contributions from anyone who wishes to contribute, its quality of information and knowledge is opaque owing to the lack of peer reviewing. Hence the Web most frequently demonstrates web pages of non-scientific nature or contents. An additional difference from traditional scientific databases and archives is the dynamics of the Web, i.e., web pages and entire sites may frequently alter contents, link structure, or completely disappear. Further, the links are *not necessarily normative*, such as credit granting or recognition providing devices, but rather *functional*, say, navigational in nature. There exists no convention of linking as in the scientific world. Further, *time* plays a different role on the Web, e.g., links can be deleted, and simultaneous reciprocal linking is a rare case in the conventional citation world and not possible in the paper based scientific communication. The analogy between links and references or citations is hence of the superficial kind and should definitively not be taken too far. On the other hand, the same analogy may indeed provide interesting hypotheses about the characteristics of links and their meaning. Also, the coverage of search engines of the total Web can in principle be investigated in the same way as the coverage of domain and citation databases in the total document landscape and possible overlaps between engines can be detected. Patterns of Web search behavior can be investigated as in traditional information seeking studies. Issue tracking and mining on the Web is feasible and knowledge discovery can be carried out, similarly to common data or text mining in administrative or textual (bibliographic) databases.

Because the Web is a highly complex distribution of all types of information carriers produced and searched by all kinds of people it is central to investigate as a social phenomenon; and informetrics indeed offers some methodologies to start from. However, one must be aware that *data collection* on the Web depends on the retrieval features of the various search engines and web crawlers or robots. Although their consistency has

improved from the mid-1990s, as demonstrated by Rousseau (1997; 1999), the various Web engines do not index the entire Web, their overlaps are not substantial (Lawrence and Giles, 1998), and their retrieval features are often too simplistic for extensive webometric analyses online. Sampling becomes thus an important issue, but is difficult to perform in a controlled manner.

The contribution first defines webometrics within the framework of informetric studies, as belonging to library and information science, and associated with cybermetrics as a generic sub-field. It outlines a link typology and terminology and makes a distinction between the web node levels when carrying out link analyses. Secondly, methods and methodological problems for Web engine coverage and quality studies, including Web crawling and sampling are discussed. This is followed by link analysis issues, including Web Impact Factor (WIF) analysis, and studies of Web interaction. The contribution owes substantially to recent works by Björneborn and Ingwersen (2001, 2004); Thelwall, Vaughan and Björneborn (2005); and Björneborn (2004).

2. THE FRAMEWORK OF WEBOMETRICS AND LINK TERMINOLOGY

Webometrics and cybermetrics are currently the two most widely adopted terms in library and information science (LIS) for this emerging research field. They are generically related, see Figure 15.1, but often used as synonyms. As originally in Almind and Ingwersen (1997), the present contribution distinguishes between studies of the Web and studies of *all* Internet applications. In this novel framework by Björneborn (2004) and Björneborn and Ingwersen (2004), *webometrics* is defined as:

"The study of the quantitative aspects of the construction and use of information resources, structures and technologies on the Web drawing on bibliometric and informetric approaches."

Hence, this definition covers quantitative aspects of both the construction and the usage sides of the Web taking on four main areas of current webometric research: (1) web page content analysis; (2) web link structure analysis; (3) web usage analysis (including log files of users' searching and browsing behavior); (4) web technology analysis (including search engine performance). All four central research areas include longitudinal studies of changes on the Web of, for example, search engine coverage, page contents, link structures, and usage patterns. In this webometric context the concept of

[2] A recent textbook emphasizing this point view is Hand, Mannila & Smyth (2001).

web archaeology (Björneborn and Ingwersen, 2001) is regarded as important for recovering historical web developments, for instance, by means of the Internet Archive (www.archive.org).

The above definition places webometrics as a LIS specific term in line with bibliometrics and informetrics, such as done by Cronin (2001). The terms 'drawing on' in the definition denote a heritage without limiting further methodological developments of web–specific approaches.

In the present framework, cf., Figure 15.1, *cybermetrics* is proposed as a generic term for:

"The study of the quantitative aspects of the construction and use of information resources, structures, and technologies on the *whole* Internet drawing on bibliometric and informetric approaches" (Björneborn, 2004; Björneborn and Ingwersen, 2004).

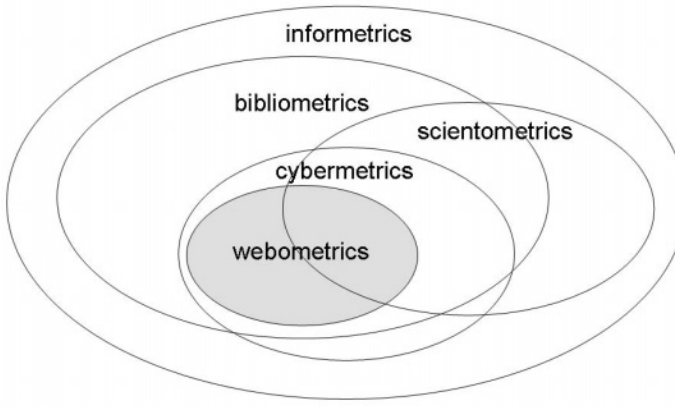

Figure 15.1. The relationships between the LIS fields of infor-/biblio-/sciento-/cyber-/webo-/metrics. Sizes of the overlapping ellipses are made for sake of clarity only (Björneborn,2004).

Cybermetrics thus encompasses statistical studies of discussion groups, mailing lists, and other computer mediated communication on the Internet (e.g., Bar-Ilan, 1997; Herring, 2002), including the Web. This definition of cybermetrics also covers quantitative measures of the Internet backbone technology, topology and traffic (cf., Molyneux and Williams, 1999). The breadth of coverage of cybermetrics and webometrics implies overlaps with approaches based on propagating computer science in Web analyses of various kinds. Björneborn (2004) and Thelwall, Vaughan and Björneborn (2005) provide comprehensive details on such analysis facets.

There are different conceptions of informetrics, bibliometrics and scientometrics. The diagram in Figure 15.1 shows the field of informetrics incorporating the overlapping fields of bibliometrics and scientometrics following the widely adopted definitions by, e.g., Brookes (1990), Egghe and Rousseau (1990) and Tague-Sutcliffe (1992). Tague-Sutcliffe states that *informetrics* is "the study of the quantitative aspects of information in any form, not just records or bibliographies, and in any social group, not just scientists" (1992). *Bibliometrics* is defined as "the study of the quantitative aspects of the production, dissemination, and use of recorded information" and *scientometrics* as "the study of the quantitative aspects of science as a discipline or economic activity" (ibid.). In Figure 15.1 politico–economical aspects of scientometrics are covered by the part of the scientometric ellipse lying outside the bibliometric one. Further, the figure shows the field of webometrics entirely covered by bibliometrics. This is because web documents, whether text or multimedia, are *recorded* information stored on web servers, cf., Tague-Sutcliffe's definition of bibliometrics (1992). The recording may be temporary only, just as not all paper documents are properly archived. Webometrics is partially covered by scientometrics because many scholar activities today are web–based whilst other such activities are even beyond bibliometrics, i.e., non-recorded, such as person–to–person conversation. Webometric studies clearly also circumscribe other social domains than the scientific one.

In the diagram the field of cybermetrics exceeds the boundaries of bibliometrics, because some activities in cyberspace commonly are not recorded, but communicated synchronously, as in chat rooms. Cybermetric studies of such activities still fit in the generic field of informetrics as the study of the quantitative aspects of information 'in any form' and 'in any social group', as stated above by Tague-Sutcliffe (1992). The inclusion of webometrics opens up the fields of bibliometrics, scientometrics, and informetrics, as webometrics inevitably will contribute with further methodological developments of web–specific approaches to the development of these embracing fields.

2.1 Link Terminology and Analysis Levels

Emerging fields like webometrics inevitably produce a variety in the terminology used. For instance, a link received by a web node has been named, e.g., incoming link, inbound link, inward link, back link, and 'sitation'; the latter term coined by Rousseau (1997) amongst others, with clear connotations to bibliometric citation analysis. The term 'external link' is an example of a more problematic terminology owing to its two opposite

meanings: 1) as a link pointing out of a web site or 2) a link pointing into a site.

Figure 15.2 presents an attempt to create a consistent basic webometric terminology for link relations between web nodes (Björneborn, 2004; Björneborn and Ingwersen, 2004; Thelwall, Vaughan and Björneborn, 2005). The figure implies that the Web can be viewed as a directed graph, using a graph–theoretic term (e.g., Kleinberg et al., 1999). In such a web graph web nodes are connected by directed links. The proposed basic webometric terminology in the legend of Figure 15.2 originates, hence, from graph theory but adheres also to social network analysis and bibliometrics (Otte and Rousseau, 2002; Park and Thelwall, 2003).

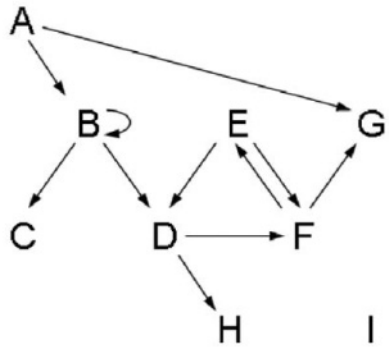

Figure 15.2. Basic webometric link terminology (Björneborn, 2004). The letters may represent different web node levels, for example, web pages, web directories, web sites, or top level domains of countries or generic sectors

- B has an *inlink* from A; B is *inlinked*; A is *inlinking*; A is an *in-neighbor* of B.
- B has an *outlink* to C; B is *outlinking*; C is an *out-neighbor* of B.
- B has a *selflink*; B is *selflinking.*
- A has no inlinks; A is *non-linked.*
- C has no outlinks; C is *non-linking.*
- I has neither in- nor outlinks; I is *isolated.*
- E and F have *reciprocal links*; E and F are *reciprocally* linked.
- D, E and F all have in- or outlinks connecting each other; they are *triadically interlinked.*
- A has a *transversal* outlink to G: functioning as a shortcut.
- H is *reachable* from A by a directed *link path*.
- C and D are *co-linked* by B; C and D have *co-inlinks.*
- B and E are *co-linking* to D; B and E have *co-outlinks.*

– Co-inlinks and co-outlinks are both cases of *co-links.*

The central terms, *outlink* and *inlink*, are commonly used in computer science–based Web studies (e.g., Pirolli et al., 1996; Chen et al., 1998; Broder et al., 2000). The term *outlink* signifies that a directed link and its two adjacent nodes are viewed from the source node providing the link, analogously to the use of the term *reference* in bibliometrics. A corresponding analogy thus exists between the terms *inlink* and *citation* with the target node as the spectator's perspective. The important conception of 'external node inlinks' in Web Impact Factor (WIF) analyses hence signifies those inlinks alone that derive from sources outside the node, i.e., excluding node selflinks. The two *co-linked* web nodes C and D in Figure 15.2 with co-inlinks from the same source node are analogous to the bibliometric concept of co-citation (Small, 1973). Correspondingly, the two *co-linking* nodes B and E having co-outlinks to the same target node are analogous to a *bibliographic coupling* (Kessler, 1963). The term 'co-links' is proposed as a generic term covering both concepts of co-inlinks and co-outlinks. The underlying assumption for the use of both the bibliometric and webometric concepts is that two documents (or two authors/link creators) are more similar, i.e., more semantically related, the higher the frequency of shared 'outlinks' (references) or shared 'inlinks' (citations).

A further discussion of this terminology can be found in Björneborn (2004) and Björneborn and Ingwersen (2004).

2.2 Levels of Link Analysis

The Web can be studied at different granularities employing what might be called micro, meso, and macro level perspectives (Björneborn, 2004; Björneborn and Ingwersen, 2004). The level depends on which of the four basic web node levels is investigated: web pages, web directories, web sites and country or generic top level domains, TLDs. Sublevels within each of the four basic node levels can exist. For example, a sub-TLD is often a central unit of analysis since many countries have assigned a level to educational, commercial, governmental and other sectors of society, for instance, *ac.uk, .co.uk, .ac.jp, .edu.au.*

Micro level webometric analyses are studies of the construction and use of web pages, web directories, and small sub-sub-sites, etc., for example, constituting individual web territories. Meso level webometrics is correspondingly concerned with quantitative aspects of larger sub-sites and sites. Macro level webometrics comprises studies of clusters of many sites, or focuses on sub-TLDs or TLDs. Several webometric studies, including classic ones by Larson (1996) and Almind and Ingwersen (1997), have used

meso level approaches concerned with site–to–site interconnectivity as well as macro level TLD to TLD analyses; primarily applying *page level* link counts to all analyses. However, in order to extract useful information, links may also be *aggregated* on different node levels as in the recently developed *Alternative Document Model* (ADM) (Thelwall, 2002; Thelwall and Harries, 2003). In contrast to the classic studies, the ADM may operate, say, at a sub-site level as analysis unit (representing, for instance, university departments), and with link counts also at sub-site levels, i.e., aggregating the page level link counts. It should be noted that a site level link *always* connects a source site with a target site. Correspondingly, a page level link always connects a source page with a target page. However, a target URL for a web page may misleadingly look like an URL for a web site, since it is common web practice to stem the target URL of top entry pages of a web site. For instance, instead of writing the full URL 'www.db.dk/default.htm' in a target link pointing to the top entry page of the Royal School of LIS, it is more expedient to stem the URL to 'www.db.dk' since web servers automatically look for default pages for stemmed URLs. However, this stemmed URL still denotes a web page and not a web site.

An adequate terminology for aggregated link relations should capture both the link level under investigation and the reach of each link. Such a terminology must reflect at least three elements: (1) the investigated link level; (2) the highest level web node border crossed by the link; and (3) the spectator's perspective, i.e., do we talk about inlink or outlink analyses. As a consequence, selflinks are used for a wider range of purposes on the Web than self–citations in the scientific literature. Page selflinks point from one section to another within the same page. Site selflinks (also known as internal links) are typically navigational pointers from one page to another within the same web site. Within the same TLD individual links connecting sub-TLDs as in- and outlinks may thus be aggregated into TLD selflinks. The unit of analysis is hence a central issue in webometrics.

3. WEB ENGINE COVERAGE, CRAWLER LIMITATIONS, AND SAMPLING ISSUES

Fundamentally, data collection made by commercial search engines (or indeed also by personal crawlers or robots) for webometric analyses takes four forms defined by two dimensions. The first dimension focuses on the Web *data types* used as the starting point for retrieval: searching for known Web locations by means of URLs (such as 'known item' searching in information retrieval); or searching for some topic(s), or other content data that define the subject area for which the Web space is to be investigated.

The second dimension deals with the *retrieval strategy* applied by the search engine (or crawler): 'content crawling' to retrieve all unique Web documents; and 'link crawling', i.e., to follow the link associations between web pages. This strategy will also retrieve duplicates (cf., Thelwall, Vaughan and Björneborn, 2005). The objective of the Web analysis determines the mode of data collection. Common webometric analysis objectives are studies of:

- *Selected Web spaces*, defined by, e.g., specific institutions, subject areas, web document/page genre, or (sub-) TLDs and/or geo- locations, or specific personal names or single web sites. Analysis units can be Web pages, (sub-)sites, sub-TLDs and/or link structures or types. The studies are often descriptive analyses of Web characteristics. Data collection is either generated by sets of URLs associated with institutions or other known entities (Thelwall, Vaughan and Björneborn, 2005) or made via searching on defined search keys, like terms and keywords, personal names or other metadata (Bar-Ilan, 2001, 2002; Jepsen et al., 2004). In both cases sampling may be mandatory owing to the size of the space investigated;
- *Web indicators*, calculated by a number of Web parameters, e.g., number of inlinks to single or sets of web pages, (sub-) sites, (sub-) TLDs divided by number of web pages receiving them (i.e., a kind of WIF); outlinks, selflinks and other types of linking are potential parameters to be considered, as are genre, subject matter, locations, scientific citations received or references made, terms applied, institutions mentioned, numbers of faculty staff, etc. data collection is carried out as for selected Web spaces above, combined with means to obtain well defined numeric data on links and other parameters, like the use of specific search engine commands or Web crawlers;
- *Human actor – Web interaction*, that is, studies of generation of Web contents, architecture and structure or link motivation, searching the Web by various populations, in diversities of subject matter and domains and for a multitude of purposes. Data collection is made from Web engine logs and/or *in situ* observations of interactive activities in real-time – also over longer periods. This kind of studies is closely associated with interactive information seeking and retrieval studies in context (Ingwersen and Järvelin, forthcoming).

For all three kinds of Web studies the investigations may take place as longitudinal studies.

Obviously, if commercial search engines are used the coverage and the qualities of the retrieved and downloadable material at search time are

crucial parameters for the resulting analysis. Hence coverage studies are central to webometric research.

3.1 Commercial Web Engine Coverage

Lawrence and Giles (1998) provided a substantial contribution with respect to the commercial search engine coverage of the Web space by introducing the concept of the publicly 'indexable Web'. The concept signifies the portion of the Web, which can be indexed by engines, excluding documents from commercial Web databases, such as, Dialog and the closed archives of publishers. That part of the information space is commonly called the 'hidden Web'. Lawrence and Giles (1999) also demonstrated that the coverage of any one engine is significantly limited by indexing only up to 17 % of the indexable Web. Central reasons behind this phenomenon are, for instance, the depth (exhaustiveness) of indexing at the local servers visited by the engine crawlers, which depends on the site structure and organization, and the link construction. Some search engines may also have indexing strategies that depends on 'pay for inclusion'. Lawrence and Giles (1999) applied randomly sampled Internet Protocol (IP) addresses. In that way it was possible to obtain a random selection of all web sites. However, this is not advisable owing to the introduction of the virtual server capability that allows one IP address to host many domain names and one 'chief' name only.

Also Clarke and Willett (1997) addressed the evaluation methodology of Web engines. They compared AltaVista, Excite, and Lycos. In addition, that paper provides a critical assessment of earlier research and produces a realistic methodology, including relative recall measures taken from IR research. It was found that AltaVista performed significantly better than Lycos and Excite. Oppenheim et al. (2000) produced a detailed review of the evaluation of Web search engines, including a discussion of test methodologies.

Whilst many coverage and evaluation studies looked into the relevance and number of web pages (recall) at a given point in time, other critical analyses covered link page retrieval by the engines (Snyder and Rosenbaum, 1999) or carried out Web structure investigations based on *time series*. As did Ingwersen (1998), Snyder and Rosenbaum observed large variations and inconsistency, in particular concerning the AltaVista engine's link page recovery at that time. Rousseau also observed the irregularity of that engine in two longitudinal studies (1999; 2001). In (1999) he compared AltaVista with NorthernLight on a daily basis over 21 weeks during 1999. This study used the same three common single words as test queries during the evaluation period. In line with the Web growth NorthernLight, as expected,

showed a steady increase of hits. However, AltaVista demonstrated large variations over time until the particular date (October, 25, 1999) when it became re-launched in a renewed and quite stable form. At that date the number of retrieved web pages increased dramatically — with this *nova*–like effect depending on the query (Rousseau, 1999) — later to drop slightly supposedly owing to the deletion of dead link pages. Rousseau used the same techniques, including (median) filtering (2001), to tracki an event on the web (the introduction of the euro). Even though that and other engines nowadays seem much more reliable (Thelwall, 2001b; Vaughan and Thelwall, 2003), and give good coverage of academic web sites (Thelwall, 2001a), their harvesting and updating algorithms, which are commercial secrets, are rarely performed at search time – but at intervals. The algorithms and commands are subject to change without notice and their advanced features are not always fully documented. Further, from a webometric research point of view, authors sending their web pages to be included in the engine's index distort Web engine coverage. In the future, pay for inclusion for commercial or simple visibility reasons may hence also bias analyses. As for the ISI citation databases the commercial Web engines display national biases in site coverage. Vaughan and Thelwall (2004) demonstrated recently that three major search engines much better covered U.S. sites than sites from China, Taiwan, and Singapore.

3.2 Commercial Web Search Engine Download Capacity

A typical method to apply to assess the coverage of Web search engines is to enter some terms, concepts, or indeed entire query profiles, as long as they are well defined, and compare to a substantial number of known Web sites that should be retrieved by the engine(s). As mentioned above Rousseau did use common words as a starting point (1999) and Bar-Ilan, for instance, used scientific domain concepts like 'informetrics' and associated terms to investigate longitudinal coverage, links and Web contents on that subject matter (1999, 2000, 2002). Similarly, Jepsen et al. (2004) applied three plant biological terms (including synonyms and spelling variations) as controlled search keys: Plant hormones; Photosynthesis; and Herbicide resistance. As with Allen et al. (1999) below the methodological idea was to observe what happens on the Web with strictly scientific topics (Plant hormones) compared more publicly and politically discussed issues (Herbicide resistance) or commonly known concepts like Photosynthesis. The goals of the study were several, amongst which three are of interest here: 1) defining the depth and overlaps of the coverage in popular Web engines (Google; AllTheweb; AltaVista); 2) their accessibility level ready

for download; 3) observing the scientific quality of the accessible material. For the latter goal see Section 4.1 below.

The individual engines retrieved different proportions of the Web in identical searches. For example, the number of hits by each engine for the key query term 'Photosynthesis' was 238,000 (Google), 119,300 (AllTheWeb), and 79,400 (AltaVista). Although the average overlap of the accessible URLs was quite substantial, 21%, the variation was also very high, from 13% to 58%. Search engine overlaps might prove to be well suited as a quality indicator, but evidently, a union of engine results may improve drastically the recall, i.e., the amount of Web materials conceivably to be investigated further.

Furthermore, the level of accessibility varied from engine to engine: only AllTheWeb provided access to several thousands of the indexed and retrieved Web publications (4,100). Google's cut-off was close to 1000 URLs and AltaVista only allowed access, presentation, and download of 200 URLs. Bar-Ilan emphasised (2001, p. 22) how this might also be a problem to the informetrician who is interested in the whole set of results for a given query, whereas it might be less problematic to the average user who only needs a few 'most relevant' URLs. Unfortunately, owing to the *skewness* found in accessible URLs between engines caused by the different page ranking algorithms applied by the engines, the data material collected may not easily legitimise a correlation analysis between search engine overlaps and quality assessed by experts. The overlaps are defined by the accessibility *and* the various ranking algorithms applied by the engines. Extended potential overlaps may hence exist between the engines outside the ranked list of URLs that can be downloaded and analysed. This facet of data collection of the Web poses problems for sampling, Section 3.4.

Notwithstanding, the reason why a webometric focus commonly is put on AltaVista is that the engine has a rather large Web coverage and hitherto has provided advanced search features fitting informetric studies of the Web. For example, AltaVista allows long and complex search strings consisting of Boolean operators combined with truncation options and specific search codes for various HTML elements. Time series paired with testing for the retrieval of the query search keys and known item (site) searching (applying AltaVista's `host:` command), conceivably also comparing to other engines' search results, seem thus very useful as tools when monitoring Web engine performance.

3.3 Web Crawler Issues

All of the considerations demonstrated above are incentives for researchers to develop data collection techniques that do not rely upon

AltaVista or any other search engine. This implies developing dedicated 'personal' Web crawlers or robots, i.e., software that automatically and iteratively downloads web pages and/or may mine and store their links and content (Thelwall, 2001a). The issue here is whether the software reaches all potential web pages, inlinks and outlinks and their associated remote web pages, for a given site or sites, or entire domains or genres (Björneborn, 2004) in defined locations.

Commonly a crawler does not extract 100 percent of the intended data. Fundamentally, the same problems concerning data collection and sampling methods for commercial Web engines also concern 'personal' crawlers – and *vice versa*. Problematic issues associated to data collections, leading to omissions of or invalid data are, for instance, (Thelwall, Vaughan and Björneborn, 2005):

- *Starting point comprehensiveness*, i.e., URL(s), contents search keys, metadata;
- *Crawler strategy*, i.e., content or link crawling, including 'random walks' used in graph theory (cf. Björneborn and Ingwersen, 2001);
- *Omission of web pages* because the crawler does not comply with their format, pages are protected by security measures that do not allow mining or crawling or by passwords, or servers are momentarily shot down;
- *Omission of isolated web pages*, see Figure 15.2, owing to lack of inlinks;
- *Non-integration of personal home pages* in institutional link structures within a site;
- *Crawl depth limitations* at web sites;
- *Page number limitation* per site visited;
- *Different domain names used* for the same entities under study, e.g., (inter)national corporations under .com, .net, .dk.

If an engine or a 'personal' crawler has omitted portions of a Web space, entire clusters of sites, each with links to and from the local Web space sought for as well as to other sites, do not become analysed. The result is that the original space to be investigated evidently becomes *deformed* by that omission and the resulting findings distorted. However, since the analyst may not know this phenomenon to have happened, it is commonly not taken into account.

What are not really problematic for 'personal' Web crawlers are Web-economic aspects, such as, update frequency or ranking of retrieved material. Also, after the collected material is downloaded a multitude of comparative analyses may take place, for instance, by means of combinations of Boolean sets. This is not possible in commercial engines.

However, commercial Web engines may make positive use of pages and URLs from previously indexed Web sites, including author submitted sites for visibility reasons mentioned above. An additional way in to data collection may be centered on local evaluation exercises. The Web servers under assessment may permit a total crawl of their directories, as done in academic bibliometric research evaluations of universities or defined sectors.

It seems evident that future Web analyses applying crawlers ought to explain the characteristics of the software, its well tested limitations and consequently the problematic issues encountered in the harvest of data for analysis.

3.4 Sampling Issues

The limitations of the search engines and web crawlers as well as the vastness of the web make comprehensive analyses of given Web spaces quite difficult. Sampling of links and web pages is hence necessary, either as randomised samples or in stratified systematic ways. The Thelwall (2001a) collection of downloaded UK inter-university Web links generated by means of a crawler and all known UK university URLs is an example of a very large data set from which random sampling could be made in order to make a variety of analyses. A way of obtaining web pages is to make use of the links between them. For instance, so called 'random walks' can be performed by means of a crawler starting from definite points on the Web (Henzinger et al., 2000; Björneborn and Ingwersen, 2001; Thelwall, Vaughan and Björneborn, 2005) and harvesting pages algorithmically during the walk to be part of the sample and downloaded. For instance, Hou and Zhang (2003) did experiments on two kinds of retrieval algorithms, starting from a given URL and based on either co-inlink analysis of associated web pages, or by application of linear algebra theories to find deeper relationships amongst web pages. Such modes of collection are link–dependent and is computing intensive.

Another way of generating samples of web pages via web sites is to use a commercial search engine and well-defined URLs from which a random sample has been drawn covering the space under investigation. Searching on a number of engines is required in order to obtain a list of relevant URLs as comprehensive as possible by means of comparing and pooling the results into a set of (home pages from) relevant sites. If the analysis unit is sub-sites or sub-sub-sites the sample should mirror their proportions, i.e., also take the stratification into account. If this pooled retrieval is performed by means of search keys, the different engines' ranking algorithms, as in the case below, have predisposed the accessible data. However, since several engines are applied the distortion may be decreased or neutralised.

In the Jepsen et al. study (2004) the search terms associated with the three search profiles on Plant biology were deliberately searched separately and comprehensively. This was owing to the variation in cut-offs of accessible URLs displayed by each search engine, shown above section 3.2, so that enough material was available for download, sampling and further analysis in a local software program. Since the study intended also to include other parameters of the web page contents in the analyses associated with the search keys, the accessible URLs were extracted, overlaps detected, duplicates removed and isolated and search engine distributions were investigated. This provided a pooled set of URLs for each profile that was stratified according to the distributions over engines and other parameters. From that set randomised and stratified samples were drawn; for instance, 200 web pages for each profile.

Ideally, the basic search results (number of hits) from the engines ought to have provided a qualified starting point for the creation of a stratified sample. Owing to the cut-off variation of accessibility between the engines, however, any stratification must involve the *accessible* portions of the web pages that are available. Since the real population is difficult to estimate the significance of the samples is uncertain and the results probably only valid as indications, not for generalization purposes. Rusmevichientong et al. (2001) and Thelwall, Vaughan and Björneborn (2005) have made a comprehensive review and discussion of Web sampling methods. Further Web characteristics underlying the data collection and hence influencing the analysis results are discussed below.

4. WEB PAGE ANALYSES PERTAINING TO DATA COLLECTION AND ANALYSIS

Fundamentally, two kinds of web page properties are at play influencing data isolation and processing. The first kind deals with the quality of the web pages actually retrieved and accessed. This involves the trustworthiness of the contents and the page ranking algorithms of the Web search engines. The second kind signifies properties of web pages, such as genre, number of outlinks and life span of pages and links.

4.1 Web Page Quality Analyses

Obviously the breakthrough for everybody to express themselves, practically without control from authorities, to become visible world wide, also by linking to which pages one wants to link to, to assume credibility by

being 'there', and to obtain access to data, information, values and knowledge in many shapes and degrees of truth, has generated a reality of freedom of information – also in regions and countries otherwise poor of infrastructure. Although some social Web etiquettes are developing, they might not be followed. The other side of the coin is that the Web increasingly becomes a *web of uncertainty* to its users; the borderline between opaqueness, shading truth, misinformation, beliefs, opinions, visions or speculation *and* reliability, validity, quality, relevance or truth becomes increasingly thinner. Picking information becomes a matter of trust. Hence Web archaeology will in future go hand in hand with webometric analyses and methods.

Evidently the individual Web engine's ranking algorithms determine *which* web pages are the highest ranked (prioritised) and thus are accessible. Lower ranked publications cannot be accessed. In the case of Google's page–ranking algorithm (Brin and Page, 1998), for instance, the 1000 top ranked and accessible pages might disproportionately belong to particular Web genres or topics that are characterised by many inlinks from authoritative Web sources, like commercial web pages and sites. The ranking algorithms are thus central to webometric analyses, because it becomes then an issue of whether the accessible publications actually are superior in a qualitative sense, e.g., are scholar or at least not opinionated or not completely untrustworthy. In particular, Google's ranking principle, building on number of inlinks to web entities and from which kind of web pages they derive. Kleinberg's (1999) recursively defined conceptions of 'hubs' (web pages with many outlinks to authorities) and authoritative web entities (with many inlinks from 'hubs') are consequently of central interest. Does a high number of inlinks always signify (cognitive) authority or rather many other hidden characteristics, e.g., that the entity with a certain probability is a commercial site? The underlying analogy with scientific citations is conceivably at play here. As noted by Otte and Rousseau (2002), the Kleinberg approach of hubs and authorities is related to the influence weight citation measure proposed by Pinski and Narin (1976) and mimics the idea of 'highly cited documents' (authorities) and reviews (hubs) in scholarly literatures.

Quality watch and assessments are currently in high demand. In particular, the health and medical domains are important areas to investigate for such issues. For instance, Courtois and Berry (1999) observed the quality among the top-20 or top-100 ranked documents retrieved by major engines. Relevant pages were such that formally contained *all* query words. This mode of 'algorithmic (or logical) relevance = quality' is similar to the most simplistic system–driven performance measures in information retrieval research (Cosijn and Ingwersen, 2000). No expert assessments were used.

Aside from the findings the paper discusses the more or less publicly available knowledge about the different indexing/retrieval features used by any one engine.

Cui (1999) made use of citation analysis methods on the Web to detect the overlapping and high frequency inlinked (cited) sites on medical information and Allen et al. (1999) looked into the reliability (and pertinence) of bio–related web pages. The former paper refers to several other studies of health and art issues treated on the Web for which Web citation analyses have been applied as a rudimentary quality indicator. The Bradford distribution of the thousands of outlinks from the 25 top medical US Schools was used as strong ties by Cui to determine the central sites concerned with specific health topics. Allen et al.'s (1999) contribution is a survey assessed by experts of the reliability of scientific web sites. As was the case for Rousseau's longitudinal study (1999) and the Jepsen et al. study (2004) mentioned above, the survey is based on the retrieval of sites according to three exemplary queries to the NorthernLight engine on 1) 'evolution', 2) 'genetically modified organisms', and 3) 'endangered species'. For each query two experts examined the top 500 web sites sequentially until each had independently reviewed approximately 60 sites containing information pertinent to the *topic*. This assessment mode is close to the methodology used in the current worldwide TREC IR evaluation experiments, applying topicality relevance measures. From 12 to 46% of the examined top pages were deemed pertinent, dependent on the topic. The 60 pertinent sites per query were scored as 'inaccurate' if they contained factually incorrect information (av. 22%), 'misleading' if they misinterpreted science or blatantly omitted facts supporting an opposing position (av. 28%), and 'un-referenced' if they presented information without any peer reviewed references (av. 71%). The latter score is purely objective. The overall agreement values for the referees' scores for the categories of 'inaccurate' and 'misleading' were 87.8% for the 'evolution' sites, 82.8% for the 'genetically modified organism' sites, and 73.6% for the 'endangered species' web sites. Un-referenced sites accounted for more that 48% for each query (1999, p. 722).

These results lead Jepsen et al. (2004) to look for filtering mechanisms in order to be able to distinguish between academic Web material and other kinds of Web information, as initially outlined above. A sample of 200 web pages from each of the three plant biological topics was drawn and assessed by one human expert into five categories plus a class of unavailable pages (11 to 22%). The first category 'scientific' was assigned to content that was deemed to be of scientific quality, for instance, preprints, conference reports, abstracts, scientific articles; 5 to 6% of the pages belonged to that category. The second category, 'scientifically related', was assigned to materials of

potential relevance for a scientific query, such as directories, CVs, institutional reports. This kind of information appeared to be abundant on the Web (17 to 25% in the analysis) and may interfere with a search for specific information on a given scientific subject. The third category of 'teaching' contained content related to teaching, e.g., textbooks, fact pages, tutorials, student papers, or course descriptions. The category accounted for 11 to 20% over the three searched topics, but was more present on web pages in Scandinavian languages (16 to 37%). The student papers did sometimes blur the distinction from formal and presumably accurate scientific papers. The fourth category 'low–grade' was reserved to content that failed to meet the criteria of the three previous groups, but still was on the topic. The category typically contained content of either commercial interest or deemed inaccurate or misleading. This proportion was definitively dependent on the nature of the topic, since 'herbicide resistance' accounted for 45% belonging to 'low grade' whilst the other two topics showed values of 15 to 27%. Content not meeting the criteria for the mentioned classes, and hence not pertinent to the topic, was assigned to the fifth category labelled 'noise' (3 to 17%).

The results of such quality assessments show that since the retrieval engines' ranking algorithms play a central role, as discussed previously, top-down sampling may produce distorted or disproportioned results that, albeit, may demonstrate something about the quality and nature of the *publicly visible Web* (Allen et al., 1999). Further, the findings indicate that even within academic Web spaces the proportion of scientifically reliable publications may be small compared to other kinds of academically associated web contents and may be blurred by other social groups interfering in the production processes on the Web, such as by scientists *in spe* (students at all levels) but also by commercial interests or lay men. Some domains may simply be more inclined to produce or receive more links than other domains or genres on the Web whereby compatibility between domains and genres becomes difficult — just as some scientific disciplines produce many more references turning into citations than others do.

4.2 Web Page Property Analysis

Aside from quality assessments web page analyses mainly deal with their contents and message providing characteristics. According to Cronin and McKim (1996, p. 170) "the Web is reshaping the ways in which scholars communicate with one another. New kinds of scholar and proto–scholar publishing are emerging. Thanks to the Web, work in progress, broadsides, early drafts and refereed articles are now almost immediately sharable … with authors able to choose between narrowcasting and broadcasting. And

peer review has emerged from the closet to reveal a spectrum of possibilities…". This belief and vision is indeed reality. Webometric analyses of the nature, such as, genres (or types) and their relationships, structures and content properties of web sites and pages, as well as link structures are important in order to understand the virtual highways and their interconnections. Larson (1996) was one of the first information scientists to perform an exploratory analysis of the intellectual structure of cyberspace. Shortly after, Almind and Ingwersen (1997) applied a variety of bibliometric–like methods to the Nordic portion of the Web in order to observe the kinds of page connections and define the typology of web pages actually found at national Nordic level. The methodology involved stratified sampling of web pages and download for local analysis purposes. The findings revealed that each web page, capable of outlinking, on average provided 9 outlinks – a proportion which nowadays still holds as approximation in the exponentially growing Web space (Björneborn, 2004). The contribution also attempted a comparison between the estimated share of scientific web pages and the distribution found in the citation indexes between the Nordic countries. Clearly, the visibility on the Web was quite different from that displayed in the citation databases. Norway, for instance, was much more visible on the Web than in the printed world at the time of analysis.

Bar-Ilan (2000) and Bar-Ilan and Peritz (2000) studied how a topic like 'informetrics' developed over time on the Web, that is, a kind of issue tracking investigation. They applied search engines, whereas, for instance, Björneborn (2004) applied a web crawler in order to investigate the nature of the academic UK Web space, with special emphasis on transversal links, short cuts between disparate topical clusters on the Web, and small world phenomena. In none of these and other similar studies has been used a complete data set, only more or less systematic or stratified sampling. Most often, the sampling methods, owing to the distortion possibilities discussed above, have been something one might call 'convenience sampling'. This means that since the total population frequently is unknown, unavailable, very large, and its properties inhomogeneous, the sampling must be done on an unsatisfyingly small scale from which generalizations probably always are statistically difficult or invalid. Applications of dedicated Web crawlers are thus an improvement because the harvested data are reusable and analysable locally. This issue is also evident for link analysis work.

5. LINK ANALYSES AND WEB IMPACT FACTOR STUDIES

In his classic webometric article on site inlinks (named 'sitations'), Rousseau (1997) analysed the patterns of distribution of web sites and incoming links. Although Rousseau, like Ingwersen (1998) later, made use of the old unstable version of AltaVista, his study operated with 343 downloaded sites for further analysis, retrieved from a query on the search keys 'informetrics' + 'bibliometrics' + 'scientometrics'. The analyses are thus more independent of the Web engine characteristics and more robust. The analyses showed that the distribution of sites followed the omnipresent Lotka distribution. Similarly, Rousseau demonstrated that the distribution of inlinks to those 343 sites also followed a Lotka distribution. The proportion of selflinks was estimated to 30%.

Since then many other types of link analyses have been performed. Either the investigations make use of predefined search profiles or sets of URLs by means of commercial search engines, or they apply personal crawlers. First, we briefly discuss Web Impact Factor (WIF) analyses, because there are some methodological problems connected to such analyses. This is followed by other selected link analyses with methodological implications.

5.1 Web Impact Factor Analysis

Ingwersen demonstrated (1998) the difference between counts of inlinks and counts of inlinking pages in his attempt to calculate the Web Impact Factors (WIF) for national domains and individual sites[3]. The underlying idea was that the WIF could say something about the awareness, authority or recognition of national sites (on average) or individual sites – but not necessarily quality. The study found three interesting results relevant from a methodological perspective. 1) Since the AltaVista search engine cannot count the actual number of inlinks to particular sites, but only the number of *pages* that are sources of an inlink at least once, selflinking will not influence the overall WIF. The external node inlinks, for instance, site inlinks, or TLD inlinks, hence becomes the important score to observe. This is because for each new web page within a given site providing one or more links to its own site, both the numerator and the denominator increase with the score 'one', given that the analysis unit is web pages. With aggregation into site or higher levels, this phenomenon does not matter. 2) The WIFs for

[3] Note that prior to Ingwersen, Rodriguez i Gairin (1997) had introduced the concept of information impact on the Internet in a Spanish documentation journal.

individual web sites was more unreliable than that of the top-level domains, such as countries. This was, however, owing to the instability of the 'old' AltaVista at that time found later (Rousseau, 1999). 3) The variance in the WIF calculations, also between engines, could be applied as a Web engine *evaluation measure,* i.e., as an indicator of engine performance. However, the instability and variance was probably fortunate, since it already, in the case of Ingwersen (1998), gave cause to prudence in applying the methodology and the interpretation of results.

In connection to the second result in Ingwersen's study, and with the idea of comparisons to, for instance, citation data or other classical parameters in mind, Smith (1999) as well as Thelwall (2000) further investigated the variance phenomena; however, still applying the unstable AltaVista version. Fortunately, exactly owing to the observed variations they both became suspicious about the coverage and retrieval properties of the engine(s). Had the results continuously been stable, etc. during these reproduced experimental trials, one might not necessarily immediately have questioned the methodology.

Smith (1999) demonstrated some periodic and robust data collection methods and showed how results became distorted owing to retrieval of *noise pages*, e.g., Indonesia (domain code: .id) showed very high WIF because of the retrieval of the URL element 'id' in many sites other than Indonesian. He also showed that the longer the URL string searched for, the more reliable the result. The context of the string should assure its uniqueness. However, later unpublished studies of the actual coverage of the engines, including AltaVista, with respect to the known pages and links on our own local server (ax.db.dk) demonstrated that they do not penetrate to all pages and links. Thelwall confirmed (2000) this negative result by applying AltaVista, Hotbot, and Infoseek in his analyses. The coverage is *not random* in such a way that the WIF denominator and numerator are influenced in identical ways. In short, at the present state of search engine coverage and retrieval strategies, "the exiting concept of WIF appears to be a relatively crude instrument in practice" Thelwall (2000, p. 188). Thus far the outcome when applying Web engines seems highly problematic, and, as stated by both Rousseau (1999), Smith (1999), Thelwall (2000), and Björneborn and Ingwersen (2001, 2004), one would have to apply dedicated web crawlers or direct URL searching to download data for local analyses.

Obviously, WIF calculations can be compared to other academic–like impact measures, like classic journal impact factors for journals that are printed and online; personal or institutional citation impact or number of citations received or publications produced; or other economic parameters like IT expenditure or number of staff, etc. A methodologically interesting study in this respect was the use an alternative WIF compared to the

Research Assessment Exercise (RAE) and research productivity and staff size in UK computer science departments by Li et al. (2003). Both AltaVista and a special crawler were used to collect link and page data (not link page numbers but the actual number of links was retrieved). Two kinds of WIFs were calculated: one with staff size per department as denominator and one with department web pages as denominator. The former WIFs correlated significantly with their *RAE ratings* whereas the latter did less well. The numerator values alone, i.e., the number of inlinks to computer departments, correlated significantly with the *research productivity* of the departments. The RAE rating correlation was interesting since Thomas and Willet (2000) did not find significant correlations between inlinks and RAE for UK LIS departments. The staff number per department seems a better indicator of departmental size than web pages numbers. Probably, there are too many pages per department of less research significance.

The major problems with the WIF are its reliability and its interpretation – as for other kinds of scientometric impact factors. The operational variable is well detectable (the links), although less robust than citations, while the theoretical variable, its meaning, is obscure or only understood to a certain degree.

5.2 Other Link Analysis Issues and Link Motivation Studies

Many comparative analyses have been done at a large scale by the Thelwall project team mainly by means of specialised crawlers (see Thelwall, Vaughan and Björneborn (2005) for methodological details). The general trend was found to be that *links* are better sources of information (Oppenheim, 2000) or indicators than web pages at directory and domain levels (Thelwall and Tang, 2003, p. 156).

Wilkinson et al. (2003) used a random sample of 414 links between UK universities. The links were classified according to scholarly content. Less than 1% dealt with contents equivalent to a scholarly article, whereas 90% was created for a variety of academic or scholarly reasons, including teaching. Wilkinson et al. (2003) showed that links to academic sites are not made solely for formal scholarly motivations. Link counts thus measure a host of *informal* scholar communication. This divide of links can be compared to the Jepsen et al. study (2004) of web page content types of academic nature discussed above. Thelwall and Tang (2003, p. 157) outlined a range of link research connected to academic web entities. For instance, interlinking between UK universities decreased with geographic distance. They found that the correlation between link count and research productivity also exist for Taiwan, outside the UK and Australia. Further, "… the most

highly targeted pages, at least in the UK, typically have little direct scholar content, e.g., university home pages, and demonstrate clear disciplinary and role biases. For example, a page may be highly targeted because it has an information dissemination purpose, or is related to computing or general university education issues." Evidently, data collection, sampling and analyses must take into account both link and page roles and types or genres which introduces issues and problems of classification and typology, i.e., subjective interpretation.

Link creation motivation studies are central for developing an understanding of how counts of links should be interpreted. Nevertheless, they have tended to lag behind statistical correlation studies. According to Thelwall, Vaughan and Björneborn (2005), most motivation studies have actually investigated the context of links, rather than attempting to directly ascertain author motivations. Motivation studies should be viewed in the context of what is known about *web use* in general. It is important to understand that web use is not determined by technology, it is context–specific (Hine, 2000). In particular, academics use the Web in many different (in)formal ways, and this is likely to continue to be true (Kling and McKim, 2000).

In relation to science and technology studies there are several types of formal academic web communication to investigate. For example, traditional journals in paper format, also available in electronic form; real peer reviewed e-journals; university online series, peer or non-peer reviewed; pre-prints in circulation (forever?) prior to final submission; etc. In all these cases, traditional references and citations *and* outlinks and inlinks are central properties to study. Kim (2000) made, for instance, a detailed investigation into authors' motivations for creating outlinks in e-journal articles. These were found to extend paper citation motivations. New ones were *functional*, that is, relating to accessibility and richness of electronic resources.

One may hypothesize that when moving more into technical/commercial web spaces the more functional and rhetorical are the outlinks. Rhetorical links, as rhetorical references, link to *authoritative* web pages or pages that are *profitable* to link to. For instance, this kind of linking can be done with the purpose of self–presentation and emphasis, showing off professional relationships and collaboration. In addition, such pages may be 'hubs' (Kleinberg, 1999), that is, having many outlinks themselves, also functioning as *web junctions*. Then we move into more functional and navigational linking motives, such as, drawing attention to relevant pages, to share knowledge, experiences, etc.

Within the academic web space one may expect also to find *normative* outlinks, aside from rhetorical and functional ones. For each of these generic types of outlinks there exist a large number of specific reasons for outlinking

that are dependent on scientific environments and domains and communication media. Analogously with references, normative linking motives could be acknowledging support, sponsorships, assistance, and providing information of a variety of commercial, academic or entertainment purposes. We have seen above that there exist significant correlations between (normative–like) formal inlinks and research productivity but that functional and rhetoric (informal) linking also display a, albeit weaker, correlation to productivity. Future investigations may reveal associations and degrees of correlations to other interesting parameters significant from the point of view of SandT evaluations. Hence, there seems to be more to distinguishing between inlink genres than commonly done in relation to citation analyses. This seems to be caused by the rather unconventional and fuzzy way linking is done, compared to giving scientific reference on lists, later to be broken up into citations to be counted. Clearly, in traditional citation analysis the motivations for making references may not really matter on large-scale citation analysis because opposite motives become neutralised.

One interesting property of the web linking behavior is that *negative outlinks* are rare or non-existent – in contrast to traditional scientific references. One should also bear in mind the dynamic nature of the Web, i.e., that *time* plays a predominant role. Ageing, i.e., generation, maturity, obsolescence, decline, and death happen faster and are probably less predictive on the Web than in traditional scientific literature (Glänzel, 2003). Methodologically speaking, this dynamic characteristic of the Web makes data collection and analysis highly cumbersome, compared to using the traditional citation databases

6. WEB ENGINE LOG STUDIES OF INTERACTION AND USE

The majority of studies of Web interaction focuses on single sites and is based on server log analyses. This kind of webometrics is a natural bridge to the other major research discipline within information science: integrated information seeking and retrieval studies. For a deeply detailed overview of Web searching research the Jansen and Pooch review (2001) is recommended. The web server log captures ordinary persons' web searching processes and provides data that is useful, not only for web interface and presentation design, but also for the interpretation of the social and psychological impact of the Web on people.

Notwithstanding, there are surprisingly few studies that have focused on *user–centered surveys*, i.e., on the searcher side of Web transactions, e.g.,

children's and high school students' use of the Web to solve assigned specific search tasks (Ingwersen and Järvelin, (forthcoming).

6.1 Large-Scale Web Engine Studies

Large-scaled Web engine studies are often based on log analysis. The Excite studies reported by Jansen, Spink and Saracevic (2000) and Spink et al. (2001) were preceded by the AltaVista study (Silverstein et al., 1999). Later, Wang, Berry and Yang (2003) reported the longitudinal study of an academic Web server over 4 years, 1997 to 2001.

The major limitations of these studies include that they only catch a narrow facet of searchers' Web interaction. The searcher, his/her intentionality, strategies, and motivations are hardly known. On the other hand, log analysis is an easy way of taking hold of data, which can be treated with quantitative methods. We can use the studies to obtain statistically significant data about user selection of search keys and use of syntax in queries.

Silverstein and colleagues (1999) performed an analysis of approximately 1,000 million requests, or about 575 million non-empty queries, from Alta-Vista. Their findings support the notion that Web users behave differently from users of traditional IR systems, they use few query terms, not the advanced IR features, investigate only a small portion of the result list, and seldom modify queries. It is, however, impossible to tell what the situation would have been like if the search engines had similar response times and the same features that professional IR systems have. Aside from searching for known items by means of URLs it is difficult directly to assess the kind of information needs that underlie the queries posed to the systems. The method used for distinguishing between searchers was a combination of the use of cookies by the searchers and IP addresses. That method is not perfect since cookies can be disabled, different searchers can apply the same browser, and floating IP addresses can be assigned to computers. A method to separate sessions is to define a session, as done by Silverstein et al. (1999) as "a series of queries by a single user made within a small range of time". After 5 minutes of searcher inactivity a session is timed out.

Jansen, Spink, and Saracevic (2000) analysed more than 50,000 queries in the query log provided by the Excite search engine, and probably made from cookies. However, the paper says nothing about whether users search for different topics during a session, i.e., one does not know if they tried to solve more than one task in one session. A follow up study based on analysis of one million queries in Excite (Spink et al., 2001) showed that searchers moved towards even shorter queries and that they viewed fewer pages of results per query.

The third large *longitudinal* investigation by Wang, Berry, and Yang (2003) analysed more than 540,000 user queries submitted to an academic Web server from 1997 to 2001. Their log file and queries did not include IP addresses of individual searchers due to privacy concerns. Hence, the sessions of the individual searchers could not be identified from the log data (p. 744). Nevertheless, the study demonstrates valuable results on query level statistics, which reveal users' search activities, as well as the actual queries that uncover both topics and linguistic structures. The observations reported are thus on the user population as a whole.

6.2 User–Centered Surveys

A different and interesting kind of study, still applying server logs but viewing the processes distinctively from a user oriented point of view, can be found in Catledge and Pitkow (1995). They carried out a longitudinal survey at the Georgia Institute of Technology on 107 persons belonging to the Institute who agreed to have their *client logs* captured over a period of three weeks. The client logs contained the URL of the users' current and target page, as well as information on the technique they used to access the target. The data was more controlled than in the previous studies above and analyzed to compute path lengths and frequency of paths and to distinguish between kinds of web users. The survey also gave some insight into which techniques and tools are being used to browse the Web, e.g., following links and using the back button as means of accessing web pages.

User–centered evaluations and direct observations of human interaction with Web search engines have started to evolve by assessing effectiveness as well as usability factors, such as screen layout, and searcher behavior during interaction. Commonly log protocols are created from monitoring actual searchers' seeking processes, e.g., by means of video or talking aloud recordings, screen and keyboard logging and forms of pre and post interviews. Session length is thus known and often applied as an IR performance parameter. Peoples' own information problems as well as assigned topics or search task situations are used as instigators of the process. See for instance the thematic issue of JASIST (Spink, 2002) for a variety of approaches to this kind of surveys.

7. CONCLUDING REMARKS

The contribution has attempted to demonstrate the relationships between the variety of 'metrical' research areas associated with library and information science, within the framework of its established sub-field

informetrics. Fundamentally, *webometrics* is referred to as belonging to cybermetrics and covered by an expanded concept of bibliometrics. Further, the contribution has made an effort to establish a consensus in connection with the outlined link terminology and web node levels of analysis.

The terminology was then applied to discussing methodological facets of webometric research, in particular concerned with academic web spaces, although the methods may very well be applied to other kinds of the Web as well. In relation to the academic part of the publicly available Web an essential aspect was to stress that the analogy between outlinks and inlinks on the one hand, and academic references and citations on the other, should *not* be taken too far. Linking motivations display a greater variety, also on the academic web, than in traditional formal scholar communication. Increasingly, the Web demonstrates tracks of informal interaction and communicative behavior. The most important problematic issues to be aware of when making data collection and analyses for studies of the Web were seen to be:

- Methods of initiating web data collection and their comprehensiveness, e.g., by URLs or appropriate search keys dealing with web page content;
- Search engine and/or specialised crawler retrieval strategies and update frequency of engines;
- Accessibility variations (for download) between search engines;
- Page ranking algorithm applied by the engines used;
- National, genre, and other biases;
- Crawl depth limitations at web sites;
- Page number limitation per site visited;
- Omissions of web pages. Pages are isolated owing to lacking inlinks. The crawlers may not comply with page format, they are protected by measures that do not allow mining or crawling, or by passwords; or servers are momentarily shot down;
- Non-integration of personal home pages in institutional link structures within a site;
- Different domain names used for the same entities under study, e.g., (inter)national corporations under .com, .net, .dk;
- Quality variation as to page genres and other kinds of data on the Web;
- Sampling methods and significance.

Aside from such problematic issues of data isolation, it is important to be aware of *what is measured*. As stated by Björneborn and Ingwersen (2004), there is, for instance, a rather large difference between counting the real number of inlinks to a web site or page and counting the number of in-neighbours in the shape of web pages (or sites) inlinking at least once to

some web node. This difference is often overlooked, in both calculus and applying the terminology.

Lastly, the distinction between *web node levels,* its terminological impact, and the application of a consistent diagram notation is necessary if the topology of the Web is to be understood and investigated. There exists a constant possibility of loosing the point of perspective in such analysis, in particular if terminological rigor is lacking.

REFERENCES

Almind, T.C., Ingwersen, P. (1997). Informetric analyses on the World Wide Web: methodological approaches to "webometrics'. *Journal of Documentation*, 53 (4), 404–426.

Allen, E.S., Burke, J.M., Welch, M.E., Rieseberg, L.H. (1999). How reliable is science information on the Web? *Science*, 402, 722.

Bar-Ilan, J. (1997). The 'mad cow disease': Usenet newsgroups and bibliometric laws. *Scientometrics*, 39 (1), 29–55.

Bar-Ilan, J. (1999). Search engine results over time: A case study on search engine stability. Cybermetrics, 2/3 (1999), paper 1. Visited 08.11.2003: http://www.cindoc.csic.es/cybermetrics/articles/v2i1p1.html.

Bar-Ilan, J. (2000). The Web as an information resource on informetrics? A content analysis. *Journal of the American Society for Information Scienc.* 51, 432–443.

Bar-Ilan, J. (2001). Data collection methods on the Web for informetric purposes: A review and analysis. *Scientometrics*, 50 (1), 7–32.

Bar-Ilan, J. (2002). Methods for measuring search engine performance over time. *Journal of the American Society for Information Science and Technology,* 53 (4), 308–319.

Bar-Ilan, J., Peritz, B.C. (2000). The life span of a specific topic on the Web. The case of ' 'informetrics': A quantitative analysis, *Scientometrics.* 46, 371–382.

Björneborn, L. (2001). *Small–world linkage and co-linkage*. Proceedings of the 12th ACM Conference on Hypertext and Hypermedia (pp. 133–134). New York: ACM Press.

Björneborn, L. (2004). *Small–world link structures across an academic Web space: a library and information science approach*. PhD Thesis. Royal School of Library and Information Science, Denmark. http://www.db.dk/dbi/samling/phd/lennartbjoerneborn-phd.pdf.

Björneborn, L., Ingwersen, P. (2001). Perspectives of webometrics. *Scientometrics*, 50, 65–82.

Björneborn, L., Ingwersen, P. (2004). Towards a basic framework for webometrics. *Journal of American Society for Information Science and Technology* (in press).

Brin, S., Page, L. (1998). The anatomy of a large scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30 (1–7), 107–117.

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., Wiener, J. (2000). Graph structure in the Web. *Computer Networks*, 33 (1–6), 309–320.

Brookes, B.C. (1990). *Biblio-, sciento-, infor-metrics??? What are we talking about?* In: L. Egghe, R. Rousseau (Eds.), Informetrics 89/90: Second International Conference on Bibliometrics, Scientometrics and Informetrics (pp. 31–43). Amsterdam: Elsevier.

Catledge, L. D., Pitkow, J. E. (1995). Characterizing browsing strategies in the World-Wide Web. *Computer Networks and ISDN Systems*, 27 (6), 1065–1073.

Chen, C., Newman, J., Newman, R., Rada, R. (1998). How did university departments interweave the Web: a study of connectivity and underlying factors. *Interacting with Computers,* 10, 353–373.

Clarke, S.J., Willett, P. (1997). Estimating the recall performance of Web search engines. *Aslib Proceedings*, 49, 184–189.

Courtois, M.P., Berry, M.W. (1999). Results ranking in Web search engines, *Online*, May/June, 39–46.

Cronin, B. (2001). Bibliometrics and beyond: some thoughts on web–based citation analysis. *Journal of Information Science*, 27 (1), 1–7.

Cronin, B., McKim, G. (1996). Science and scholarship on the World Wide Web: A North American perspective. *Journal of Documentation*, 52, 163–172.

Cui, L. (1999). Rating health Web sites using the principles of citation analysis: A bibliometric approach. *Journal of Medical Internet Research*, 1 (1), e4 (ISSN: 1438–8871). Visited 08.11.2003: http://www.jmir.org/1999/1/e4/index.htm.

Egghe, L., Rousseau, R. (1990). *Introduction to informetrics: quantitative methods in library, documentation and information science.* Amsterdam: Elsevier.

Glänzel, W. (2003). *Personal communication*. Available – visited 08.11.2003. http://www.oud.niwi.knaw.nl/nerdi/lectures/glanzel.pdf

Herring, S.C. (2002). Computer–mediated communication on the Internet. *Annual Review of Information Science and Technology*, 36, 109–168.

Hine, C. (2000). *Virtual Ethnography*. London: Sage.

Henzinger, M.R., Heydon, A., Mitzenmacher, M., Najork, M. (2000). *On near–uniform URL sampling*. Proceedings of the 9th International World Wide Web Conference, May 2000. Computer Networks, 33 (1–6), 295–308.

Hou, J.Y. & Zhang, Y. (2003). Effectively finding relevant Web pages from linkage information. *IEEE Transactions on Knowledge and Data Engineering*, 15 (4), 940–951.

Ingwersen, P. (1998). The calculation of Web Impact Factors. *Journal of Documentation*, 54, 236–243.

Ingwersen, P., Järvelin, K. *The Turn: integration of information seeking and retrieval in context*. Kluver (forthcoming).

Jansen, B.J., Pooch, U. (2001). A review of Web searching studies and a framework for future research. *Journal of the American Society for Information Science*, 52 (3), 235–246.

Jansen, B. J., Spink, A., Saracevic, T. (2000). Real life, real users, and real needs: a study and analysis of user queries on the web. *Information Processing & Management*, 36, 207–227.

Jepsen, E.T., Seiden, P., Ingwersen, P., Björneborn, L., Borlund, P. (2004). Characteristics of scientific Web publications: Preliminary data gathering and analysis. *Journal of American Society for Information Science and Technology* (in press).

Kleinberg, J.M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46 (5), 604–632.

Kleinberg, J., Kumar, R., Raghavan, P., Rajagopalan, S., Tomkins, A. (1999). The Web as a graph: measurements, models and methods. *Lecture Notes in Computer Sc.*, 1627, 1–18.

Kling, R., McKim, G. (2000). Not just a matter of time: field differences in the shaping of electronic media in supporting scientific communication. *Journal of the American Society for Information Science*, 51 (14), 1306–1320.

Larson, R. (1996). *Bibliometrics of the World Wide Web: an exploratory analysis of the intellectual structure of cyberspace*. Proceedings of the 59th Annual Meeting of the American Society for Information Science, 33, 71–78.

Lawrence, S., Giles, C.L. (1998). Searching the World Wide Web. *Science*, 280, 98–100.

Lawrence, S., Giles, C. L. (1999). Accessibility and distribution of information on the Web. *Nature*, 400, 107–110.

Li, X.M., Thelwall, M., Musgrove, P., Wilkinson, D. (2003). The relationship between the WIFs or inlinks of Computer Science Departments in UK and their RAE ratings or research productivities in 2001. *Scientometrics*, 57 (2), 239–255.

Molyneux, R.E., Williams, R.V. (1999). Measuring the Internet. *Annual Review of Information Science and Technology*, 34, 287–339.

Oppenheim, C., Morris, A., McKnight, C. (2000). The evaluation of WWW search engines. *Journal of Documentation*, 56, 190–211.

Otte, E., Rousseau, R. (2002). Social network analysis: a powerful strategy, also for the information sciences. *Journal of Information Science*, 28 (6), 441–454.

Park, H.W., Thelwall, M. (2003). Hyperlink analyses of the World Wide Web: A review. Journal of Computer–Mediated Communication, 8 (4). Visited 08.11.2003: http://www.ascusc.org/jcmc/vol8/issue4/park.html.

Pinski, G., Narin, F. (1976). Citation influences for journal aggregates of scientific publications: theory, with applications to the literature of physics. Information Processing and Management, 12, 297–312.

Pirolli, P., Pitkow, J., Rao, R. (1996). Silk from a sow's ear: extracting usable structures from the Web. CHI 96 Electronic Proceedings. Visited 08.11.2003: http://www.acm.org/sigchi/chi96/proceedings/papers/Pirolli_2/pp2.html.

Rodriguez I Gairin, J.M. (1997). Volorando el impacto de la informacion en Internet: Altavista, el "Citation Index' de la Red. Revista Espanola de Documentacion Scientifica, 20 (2), 175–181.

Rousseau, R. (1997). Sitations: an exploratory study. Cybermetrics, 1 (1). Visited 08.11.2003: http://www.cindoc.csic.es/cybermetrics/articles/v1i1p1.html.

Rousseau, R. (1999). Daily time series of common single word searches in AltaVista and NorthernLight. Cybermetrics, 2/3, paper 2. Visited 08.11.2003: http://www.cindoc.csic.es/cybermetrics/articles/v2i1p2.html.

Rousseau, R. (2001). *Evolution in time of the number of hits in keyword searches on the Internet during one year, with special attention to the use of the word euro.* In M. Davis, C. Wilson (Eds.), Proc. of the 8th Int. Conf. on Scientometrics & Informetrics. Sydney, 619–627.

Rusmevichientong, P., Pennock, D.M., Lawrence, S., Giles, S.L. (2001). *Methods for sampling pages uniformly from the Web*. In Proceedings of the AAAI Fall Symposium on Using Uncertainty within Computation, 121–128.

Silverstein, C., Henzinger, M., Marais, H., Moricz, M. (1999). Analysis of a very large Web search engine query log. *SIGIR Forum*, 33 (1): 6–12.

Smith, A.G. (1999). A tale of two web spaces: comparing sites using web impact factors. *Journal of Documentation*, 55, 577–592.

Spink, A. (2002). Introduction to the special issue on Web research. *Journal of the American Society for Information Science & Technology*, 53 (2), 65–66.

Spink, A., Wolfram, D, Jansen, B. J., Saracevic, T. (2001). Searching the Web: the public and their queries. *Journal of the American Society for Information Science*, 52 (3), 226–234.

Snyder, H., Rosenbaum, H. (1999). Can search engines be used as tools for web-link analysis? A critical view. *Journal of Documentation*, 55, 375–384.

Tague-Sutcliffe, J. (1992). An introduction to informetrics. *Information Processing & Management*, 28 (1), 1–3.

Thelwall, M. (2000). Web impact factors and search engine coverage. *Journal of Documentation*, 56, 185–189.

Thelwall, M. (2001a). Extracting macroscopic information from web links. *Journal of the American Society for Information Science and Technology,* 52 (13), 1157–1168.

Thelwall, M. (2001b). The responsiveness of search engine indexes, *Cybermetrics*, 5 (1). Visited 08.11.2003: http://www.cindoc.csic.es/cybermetrics/articles/v5i1p1.html.

Thelwall, M. (2001c) A Web crawler design for data mining. *Journal of Information Science,* 27 (5), 319–325.

Thelwall, M., Tang, R. (2003). Disciplinary and linguistic considerations for academic Web linking: an exploratory hyperlink mediated study with Mainland China and Taiwan. *Scientometrics*, 58 (1), 155–181.

Thelwall, M., Vaughan, L., Björneborn, L. (2005). Webometrics. *Annual Review of Information Science and Technology*, 39 (in press).

Thomas, O., Willett, P. (2000). Webometric analysis of departments of librarianship and information science. *Journal of Information Science*, 26 (6), 421–428.

Vaughan, L., Thelwall, M. (2003). Scholarly use of the Web: What are the key inducers of links to journal web sites? *Journal of the American Society for Information Science and Technology*, 54 (1), 29–38.

Vaughan, L., Thelwall, M. (2004). Search engine coverage bias: evidence and possible causes. *Information Processing & Management* (to appear).

Wilkinson, D., Harries, G., Thelwall, M., Price, E. (2003). Motivations for academic Web site interlinking: evidence for the Web as a novel source of information on information scholarly communication. *Journal of Information Science*, 29 (1), 59–66.