Journal of Documentation, Vol. 53, No. 4, September 1997

INFORMETRIC ANALYSES ON THE WORLD WIDE WEB: METHODOLOGICAL APPROACHES TO 'WEBOMETRICS'

TOMAS C. ALMIND* *and* PETER INGWERSEN
tomas@information4u.com & pi@db.dk

Centre for Informetric Studies, The Royal School of Librarianship
Copenhagen, Denmark

This article introduces the application of informetric methods to the World Wide Web (WWW), also called Webometrics. A case study presents a workable method for general informetric analyses of the WWW. In detail, the paper describes a number of specific informetric analysis parameters. As a case study the Danish proportion of the WWW is compared to those of other Nordic countries. The methodological approach is comparable with common bibliometric analyses of the ISI citation databases. Among other results the analyses demonstrate that Denmark would seem to fall seriously behind the other Nordic countries with respect to visibility on the Net and compared to its position in scientific databases.

1. INTRODUCTION

THE AIM OF THIS ARTICLE is to introduce and argue for the interesting idea that it is possible to utilise informetric methods on the World Wide Web (WWW). While informetrics is the research into information in a broad sense and not only limited to scientific communication, the approach taken here will be called Webometrics, which covers research of all network-based communication using informetric or other quantitative measures. It is obvious that informetric methods using word counts and similar techniques can be applied to the WWW. What is new is to regard the WWW as a citation network where the traditional information entities, and citations from them, are replaced by Web pages. These pages are the entities of information on the Web, with hyperlinks from them acting as citations.

The use of informetric methods on the WWW is very exciting and allows for analyses to be carried out almost in the same way as is traditional in the citation databases. Until now these ideas have been used in only a few publications [1, 2].

The future for the use of informetric methods in the field of electronic communication was observed by William Paisley in 1990 [3, p. 286]:

*Tomas Almind lost his life in a road accident while this paper was in proof

Aslib, The Association for Information Management
Information House, 20–24 Old Street, London EC1V 9AP
Tel: +44 (0) 171 253 4488, Fax: +44 (0) 171 430 0514
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 53, No. 4, September 1997

> In the future, a large proportion of all text that now appears in books, journals, magazines, and newspapers will be contained in electronic databases. In fact, the electronic databases will contain more than the published record because there will be unpublished data collected just for the databases. This vast collection of electronic information is the future domain of bibliometric research.

In order to substantiate this framework and show how informetric methods can be used on the WWW we can take the traditional use of informetric methods as our starting point.

Tague-Sutcliffe [4] describes the following uses of informetrics:

- statistical aspects of language, word, and phrase frequencies, in both natural language text and indexes, in both printed and electronic media;
- characteristics of authors – productivity measured by number of papers or other means, degree of collaboration;
- characteristics of publication sources, most notably the distribution of papers in a discipline over journals;
- citation analysis: distribution over authors, papers, institutions, journals, countries; use in evaluation; cocitation-based mapping of disciplines;
- use of recorded information: library circulation and in-house book and journal use, database use;
- obsolescence of the literature, as measured both by use and citation;
- growth of subject literature, databases, libraries; concomitant growth of new concepts;
- definition and measurement of information; and
- types and characteristics of retrieval performance measures.

Many of the proposals above are based on investigations made in citation indexes.

Throughout this article, the main citation indexes considered will be the files from ISI as they are stored on Dialog (Knight-Ridder Information. Inc.). The citation databases are characterised as containing a general bibliographic representation of the indexed documents, together with a searchable list of a document's citations.

It is therefore given that the citation databases are among the most important tools for carrying out informetric research, as detailed above. The full-text dimension of modern database systems somewhat alters the scene.

Some of the above analysis methods make use of the special potential provided by full-text databases with respect to the different forms of word counting and census. It is possible to carry out all these types of analysis on the Web. This is mainly made possible via the tags which are added to each of the information objects in the form of the HyperText Markup Language (HTML) codes. HTML is actually a formatting tool, but it is possible to use the HTML codes to search and retrieve information, and thus also to perform informetric analysis.

Aslib, The Association for Information Management
Information House, 20–24 Old Street, London EC1V 9AP
Tel: +44 (0) 171 253 4488, Fax: +44 (0) 171 430 0514
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 53, No. 4, September 1997

Today HTML is widely used in an information retrieval (IR) context but very little in an analysis context. The HTML syntax, together with the meaning of the codes and their general application will not be discussed in this article, although a comparison between HTML and the field codes of the citation indexes will be discussed later; reference can be made to Graham [5], Powell [6] and Raggett [7] for elaboration of the HTML tags. This means that the possibilities available in citation indexes and full-text databases respectively can be combined on the WWW where it is possible to search a citation network that also contains the full texts. At the same time, the WWW provides the potential for investigating multi-media objects, when they are incorporated in the Web pages being searched.

In the following section some of the opportunities for using informetric methods on the WWW will be elaborated upon, together with a discussion of the problems arising when implementing informetric analysis on the Web. It may be noted that others are investigating the WWW from a quantitative but not a webometric viewpoint [8]. In order to illustrate the options for such informetric analyses, Sections 4 and 5 describe a case study where the core of these options for informetric analysis on the Web is tested. The analyses basically circumscribe the characteristics about which information is required. They generate a background for further analyses which could be citation analyses. A further aim is to illustrate how it is possible to become acquainted with the characteristics of information on the WWW, and how they can be used. The method is described in Section 4 and the results are displayed in Section 5. Section 6 discusses what has been demonstrated in the preceding sections, followed by a summary and conclusion in Section 7.

2. COMPARISON OF THE WWW AND THE TRADITIONAL ISI DATABASES

The World Wide Web is well designed for informetric investigations of the references, i.e. the links, between the information objects, as both the objects and the quoted information are easily accessible. Simultaneously there are, via HTML codes, many different possibilities for differentiating between individual documents, whether for isolating or gathering data, or for data analysis, as shown in Figure 1. What is lacking is any enforced conformity of form and content in the Web pages, and therefore these codes can be used differently from Web page to Web page. Another dimension is that of time. The citation databases are retrospective, whereas the Web is constantly in real time.

To illustrate that it is possible to carry out the same informetric analyses on the WWW as is possible via a citation database, Figure 1 compares the individual field codes from a citation database with a description of how the same information may be found on the WWW. The functional similarities between the field codes found in the ISI files and the HTML tags used on the Web are described below.

In the ISI files the document abstract is available in the AB field. It is obvious that the Web page contains the full text, as it is not a document representation like an ISI record. Descriptors are found in three forms in the ISI indexes:

Aslib, The Association for Information Management
Information House, 20–24 Old Street, London EC1V 9AP
Tel: +44 (0) 171 253 4488, Fax: +44 (0) 171 430 0514
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 53, No. 4, September 1997

	HTML	*ISI files*
Abstract/Full Text	Web page	AB
Author Keywords	<EM>, <STRONG>	DE
KeyWords Plus	Robot frequency	ID
Research Fronts/Best of	Most cited (Lycos 250)	RF
Title	<TITLE>, <H1>	TI
Author	URL, <ADDRESS>	AU
Information on cited entity	<A>	CA, CP, CR, CW, CY
Corporate Source (Affiliation)	URL	CS, ZP
Document Type	File extension (ex. .gif)	DT
Journal Name	<TITLE>, <H1>	JN
Language/Country Name	Manual identification	LA
Number of References	Manual identification/Web index	NR
Publication Year	Manual identification	PY
Subject Category	Manual identification	SC, SF
Size of information entity	Manual identification/Web index	not possible

The < >s indicate that the information is found within a HTML tag.

Figure 1. *Search codes for searches on the WWW and in citation indexes*

Author Keywords, KeyWords Plus and Research Fronts. The main point here is that the descriptors are either given by an author or by frequencies, and so are the subject access points of Web pages, where an author can use tags such as <em> and <strong>. Frequencies of terms or links are measured by some Web indexes.

The titles of the Web pages are found either within the <TITLE> or/and <H1> tag. A Web page can uniquely be identified by the URL of the page. Whether the author is a person or corporate source can only be identified manually, and only if the information is contained in the Web page. The corporate source or affiliation for Web pages is given by the first part of a URL; but the institution hosting a Web page is not necessarily connected with the author of the Web page in any way. The rest of the data elements are of less interest in an informetric study of the Web and will therefore not be discussed further.

There are some disadvantages in using the WWW for informetric analyses. These are mainly identical to those relating to the citation indexes. The problems traditionally found are often created by restrictive file structures together with flaws in the data validity itself. Ingwersen and Hjortgaard [9] have examined the pitfalls of online informetric analysis using the ISI files which illustrate these points. Data validity is a special problem on the WWW, as each individual author marks up and thereby indexes his or her own information entity. On the other hand, the full text is available and retrievable on the Web and can be manipulated as is necessary.

Aslib, The Association for Information Management
Information House, 20–24 Old Street, London EC1V 9AP
Tel: +44 (0) 171 253 4488, Fax: +44 (0) 171 430 0514
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 53, No. 4, September 1997

3. THE WWW SCENARIO AND ITS ACCESS PROBLEMS

The preparatory work necessary to carry out a quantitative analysis of any information on the WWW is very cumbersome as regards time and effort. This will be illustrated in the case study below. Certain problems for informetric analysis exist which are specific to the WWW. It is only possible to carry out an asynchronous analysis due to the dynamic and real-time nature of the Web. Further, the WWW is a distributed hypertext system. This fact has the effect that all data are unstructured, and there are no controls or requirements for the use of mark-up codes.

Another problem which affects data collection is that it is not always easy to sort or identify the Web pages one wishes to investigate. Nevertheless, it is possible to identify all Danish Web pages, because a Danish Web page must have '.dk' in its URL. However, it is not possible to identify all Web pages from New York or Washington DC, as there is no requirement for a Web page from these places to carry a special character combination in their URLs.

The enormous amount of data available on the Web means that it is very difficult to find exactly what one needs. The solution to this problem is to use the databases on the WWW that attempt to index as much as possible. Their indexing and coverage are very mixed and erratic. Some of these databases are now becoming more professional in the sense that they have a wide coverage and offer reasonable search facilities. The largest problem is still, however, their size and the dynamic nature of the WWW. Most Web databases are monolithic indexes that try to contain the entire WWW. This requires inappropriately large resources, especially as the nature of the Web requires very frequent updating. A more suitable strategy is to create a distributed index of the Web. The Nordic Web Index uses this pattern. Each of the Scandinavian countries does the indexing and maintains a database of their respective share of the WWW [10]. The conclusions that can be drawn from an informetric analysis performed on the Web are not illustrated in the paper, but like all informetric methods the results are given as indicators which either correlate with, or diverge from, other indicators which may have more qualitative properties.

4. THE CASE STUDY

The analysis described is an example of what is possible when exploring the Web structures. Further, the examples provide a good picture of a country's (Denmark) share of the World Wide Web, and illustrate the general Web page characteristics.

4.1 The collection of data

The starting point for this investigation is the entire population of Danish Web pages. In this context 'Web pages' refers to both text documents and hypertext documents, as well as to pictures and binary code files such as compressed files and programs. The only requirement is that the URL for the information entity via the HTTP protocol should be Danish. This is checked by ensuring that the

Aslib, The Association for Information Management
Information House, 20–24 Old Street, London EC1V 9AP
Tel: +44 (0) 171 253 4488, Fax: +44 (0) 171 430 0514
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 53, No. 4, September 1997

URL contains '.dk' as the server suffix. To ensure the highest possible level of completeness in identifying pages fulfilling these parameters, several different search engines and sources were used in parallel to collect Danish Web pages. This can be seen as a form of search using polyrepresentation [11], in that the different sources provide different cognitive and functional results as well as representations of the objects for which the search is being made.

The term 'home page' is only used for those Web pages which are classified as being a home page according to the classification in Section 4.2.4. The sources used for data collection can be divided into the following types:

- Server based
  - Data collection based on access to the server's own file hierarchy
- Client based
  - Web indexes
  - Lists, such as Yahoo
  - Databases/Indexes, such as Lycos
- Known Web pages, which can contain links to pages which have yet to be found
- Browsing through the Net.

The optimal method for data collection is if one has direct access to the data located on a server which is relevant for the analysis in question. This will ensure that all the Web pages which should be investigated are present. This method is appropriate where there is only a single institution to be investigated. In such a case there are but a few Web servers to which one needs to have access. If, on the other hand, the investigation has to cover a large number of Web servers, as is necessary if a country or domain is to be examined, it will usually not be possible to obtain access to all the relevant servers' file hierarchies. In these circumstances this procedure can still be used to check on the completeness of the results given by the other methods. This is done by investigating a server's file hierarchy compared with the Web pages found on the server by other methods.

Web indexes have been the method used first and foremost in this analysis of the Danish part of the WWW, as this is the method which gives most 'hits'. The disadvantage with Web indexes is that their output in the form of document representations is very individual and does not contain the same data elements.

The Web indexes that were used, Lycos and OpenText, were judged by many sources to be the best available, measured both qualitatively and quantitatively at the time the searches were carried out [12–14].

Searches were also carried out in a Scandinavian WWW index, the Nordic Web Index [15]. Examining known Web pages to find further Web pages is a very time-consuming and laborious task, and can be equated to manual citation counting. Despite this, the method is to be recommended in that it produces an estimate of the completeness of the other search results.

Browsing is also very time-consuming, and can only be recommended as a check on the completeness of the other methods.

To ensure that the Web pages collected could be shown to contain the total

Aslib, The Association for Information Management
Information House, 20–24 Old Street, London EC1V 9AP
Tel: +44 (0) 171 253 4488, Fax: +44 (0) 171 430 0514
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 53, No. 4, September 1997

population of Danish Web pages, the different sources were compared to examine how much the results overlapped. The hypothesis was that a large overlap meant that a suitably large number of Web pages had been collected so that pages missed would not make any meaningful difference. In this way, approximately 47,000 Danish URLs were found on the Nordic Web Index and these were used as the starting point. All other Danish URLs were searched for among these and added if they were not already present. Approximately 200 new URLs were found by this method. There is therefore a very large overlap between the sources used. Next, a simple random sample of 200 Web pages was extracted by collecting Web pages after the selection of URLs from the population by random numbers. The sample was stored in a database with the following elements: URL, title, type, domain, size and number of links. The individual elements were found via the method described in Figure 1. This sample and database can be seen as a snapshot of the Danish part of WWW at that time (December, 1995).

4.2 Analysis method

In the following sections the methods for the analyses are described. These analyses are: first, an analysis of Denmark's position on the Web; second, an analysis of the distribution of Danish Web pages on large centres of learning in Denmark; third, a method for the analysis of the sample, distributed over scientific domains; fourth, an analysis of the sample Web pages distributed over type of document; and last, an analysis of selected frequency distributions for the sample Web pages.

4.2.1 Analysis of Denmark, Sweden and Norway To give a value for Denmark's position on the WWW, the total number of Danish Web pages was compared to the number of Swedish and Norwegian Web pages. The resulting distribution was then compared to the three countries' presence in the traditional citation indexes, and further compared to a similar Web investigation carried out in April 1995 [16]. In total, the analyses display two unique snapshots of the Web and a proportional distribution of the research published in the three countries. Only the latter part of the analysis is directly reproducible by means of the ISI databases.

More specifically, the three countries were searched for in different Web indexes. It has to be assumed that the incompleteness in coverage is the same for each country on these Web indexes so that the incompleteness or potential bias has no influence, as long as the only measure being considered is the relative proportions between the three countries.

Searching the WWW was done by means of Lycos [17] and the Nordic Web-Index [15], and the results are shown in Figures 4A-4B below. The search parameters used were ((Denmark or Danmark) or .dk). The first parameter finds the country name in the document itself; the second finds the country code in the document's URL. Search parameters were specified in the same way for Sweden ((Sweden or Sverige) or .se) and Norway ((Norway or Norge) or .no).

Aslib, The Association for Information Management
Information House, 20–24 Old Street, London EC1V 9AP
Tel: +44 (0) 171 253 4488, Fax: +44 (0) 171 430 0514
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 53, No. 4, September 1997

The investigation of the citation indexes available on Dialog was restricted in the same way, in that not all Scandinavian research is published in sources that are indexed in the citation indexes [9]. It must therefore again be accepted that this lack of completeness is the same for all three countries. The search arguments used were gl=Denmark/1991:1997, gl=Sweden/1991:1997 and gl=Norway/1991:1997, where gl stands for geographic location. One should note that, using the search statement 'Denmark or Danmark' in a search both on the WWW and in the citation indexes (using Corporate Source (CS)) produces some incorrect sources, in that it will collect documents which are not Danish; for example, 'Denmark Street, London, UK' (CS) or 'I was in Denmark last summer' (WWW). This does not, of course, apply to the Geographic Location (GL) field on Dialog, which therefore is the preferred entry point.

4.2.2 Analysis of the large centres of learning in Denmark As in the previous analysis a comparison was made to a search in the citation indexes. On the WWW the extracted 200-page sample formed the data window investigated. This was compared with a similar analysis from April 1995 [16], which was performed on a population of 388 Web pages. The Dialog search was carried out using the Corporate Source (CS) field (with its inherent possibility of errors) combined with the GL field. On the WWW the centres of learning were identified from the Web server name. This search requires that the connection between the URL, institution and geographic location is known. The search terms are illustrated in Figure 2 and the distribution is shown in Figure 7.

To illustrate the connection on the WWW between institutions and geographic locations, it should be explained that '.ku' is the University of Copenhagen, '.diku' is the University of Copenhagen, Department of Computer Science, and '.hhk' is the Copenhagen Business School.

There are some differences in what is measured in the ISI files and on the Web. The study on the Web is divided so that the number for each town is all academic Web pages; from the sample, all private sector Web pages are gathered in the category 'others'. The search in the ISI files is not divided into private or academic sectors. This makes it hard to compare the two searches directly, as each town or other category can contain private R&D institutions. However the fact that the data have been divided for the Web makes it possible to observe and compare a measure of the proportion of academic and private

City	*Citation indexes (CS)*	*Sample Web pages*
København	copenhagen? or kobenhavn?	.ku .diku .hhk
Lyngby (Lundtofte)	lyngby	.dtu .dth .dtv .dtb
Roskilde/Risø	roskilde or risoe	.ruc .risoe
Odense	odense	.ou
Århus	aarhus?	.aau
Ålborg	aalborg?	.auc

Figure 2. *Search terms for the large centres of learning in Denmark*

Aslib, The Association for Information Management
Information House, 20–24 Old Street, London EC1V 9AP
Tel: +44 (0) 171 253 4488, Fax: +44 (0) 171 430 0514
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 53, No. 4, September 1997

sector Web pages, in the snapshot samples (Figure 5). This division between academic and private sector is much easier to perform where the domain names '.edu' and '.com' are used.

4.2.3 Domain analysis To identify the difference between scientific domains, seven broad domains were isolated. This categorisation was taken from the classification found in the citation indexes' SubFile field (SF). These disciplines were: Science; Medicine/Clinical medicine; Physics/Chemistry/Inorganic chemistry; Agriculture; Engineering; Social science; and Humanities (Figure 6).

The 'Science' category consists of science objects which were *not* sub-classified into one of the other science domains. The 'Engineering' domain includes all computer-related resources. The classification 'Computer' was applied to the Web pages which predominantly consisted of computer or WWW information without references to other topics. These Web pages were counted both in the 'Engineering' domain as well as the 'Computer' category. This classification is not particularly appropriate. However, the only alternative is to use the Subject Category (SC) field which would give a far too specific differentiation for our purpose. Each object can be assigned to more than one domain. This fact implies that the total percentage is higher then 100%. In this analysis, no differentiation has been made between the academic and the private sectors.

It is possible to compare 'Science and Technology', 'Social Sciences' and 'Arts and Humanities', as the 'Science and Technology' domains have been placed together in the 'Total Science and Technology' category (Figure 6). On the Web the sample from the 1996 snapshot was used, and compared to the analysis from April 1995 [16].

4.2.4 Document types This analysis is carried out solely on the WWW sample as we are of the opinion that at present no meaningful comparison can be made between document types contained in the ISI databases and on the WWW. The data made available on the WWW consist of many different types of documents. These are classified according to Almind's method of classification [16, pp. 24–25]. The intention is to classify Web pages according to the function given them by their authors:

- personal home page: a home page whose main purpose is to represent an individual;
- institutional/organisational home page: a home page whose main purpose is to represent an organisation;
- subject defined/*ad hoc* home page: a home page whose main purpose is to represent a subject;
- pointer document/index page: a Web page whose function is primarily to make a number of hyperlinks available;
- resources: Web pages which primarily make data available, for example, in the form of text, sound, pictures or film.

The sample of Web pages was divided up according to this classification and is

Aslib, The Association for Information Management
Information House, 20–24 Old Street, London EC1V 9AP
Tel: +44 (0) 171 253 4488, Fax: +44 (0) 171 430 0514
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 53, No. 4, September 1997

shown in Figure 7. Each Web page could be assigned to more than one group. The analysis is compared to the similar sample analysis from April 1995.

4.2.5 The frequency distribution for Web pages The investigation included a study of how the extracted samples of Web pages were apportioned with respect to size, number of hyperlinks and the size per link ratio for the sample Web pages, called link density. These parameters were also classified by type and discipline in order to observe whether this distribution can show anything useful or characteristic about Web pages. The material for these analyses is extracted from the sample database mentioned in section 3.1 and shown in Figures 8–11. Each of the Figures 8 to 11 contains a measure of the average of the total sample and the average for either the type or the domain. This makes it possible to observe which types or domains differ from the average.

5. RESULTS OF THE CASE STUDY

5.1 Comprehensive analysis by country: Denmark, Sweden and Norway

The comparative analysis of the three Scandinavian countries: Denmark, Sweden and Norway in the citation indexes and on the WWW results in the distribution shown in Figures 3 and 4. The data from the Web are the total population, not a sample.

Figures 4A and 4D show that the Web indexes in average gave about 540 % more pages for the chosen countries than the number of pages found twenty months earlier. This result undoubtedly relates more about the way the *indexes themselves have expanded* than about the expansion of the WWW; note for example, the changes for Sweden and Norway on 10.01.97 for Lycos. A similar growth on the Web is of course possible. The proportion between the three countries is almost the same in the first four searches in Lycos.

Figure 4B from the Nordic Web Index shows a slightly different distribution

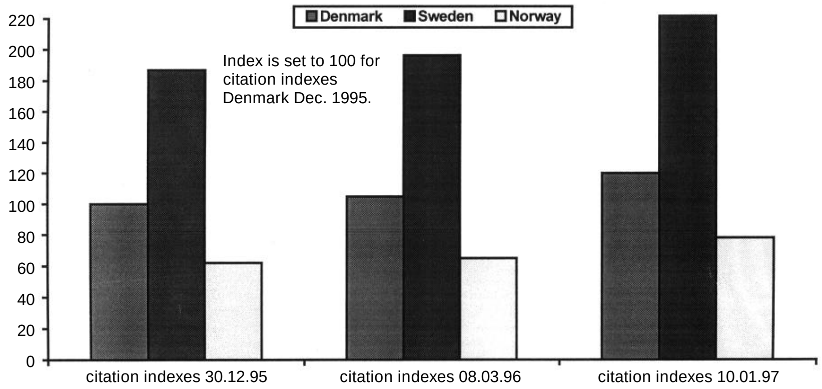

Figure 3A. *The distribution between Denmark, Sweden and Norway for citation indexes*

Aslib, The Association for Information Management
Information House, 20–24 Old Street, London EC1V 9AP
Tel: +44 (0) 171 253 4488, Fax: +44 (0) 171 430 0514
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

	Denmark		*Sweden*		*Norway*	
	No.	*Index*	*No.*	*Index*	*No.*	*Index*
Citation indexes 30.12.95	35,694	100*	66,903	187	22,106	62
Citation indexes 08.03.96	37,399	105	70,076	196	23,127	65
Citation indexes 10.01.97	42,712	120	84,205	236	27,800	78

(* = index base-line = 100)

Figure 3B. *The distribution of publications between Denmark, Sweden and Norway. Numbers and corresponding index figures*

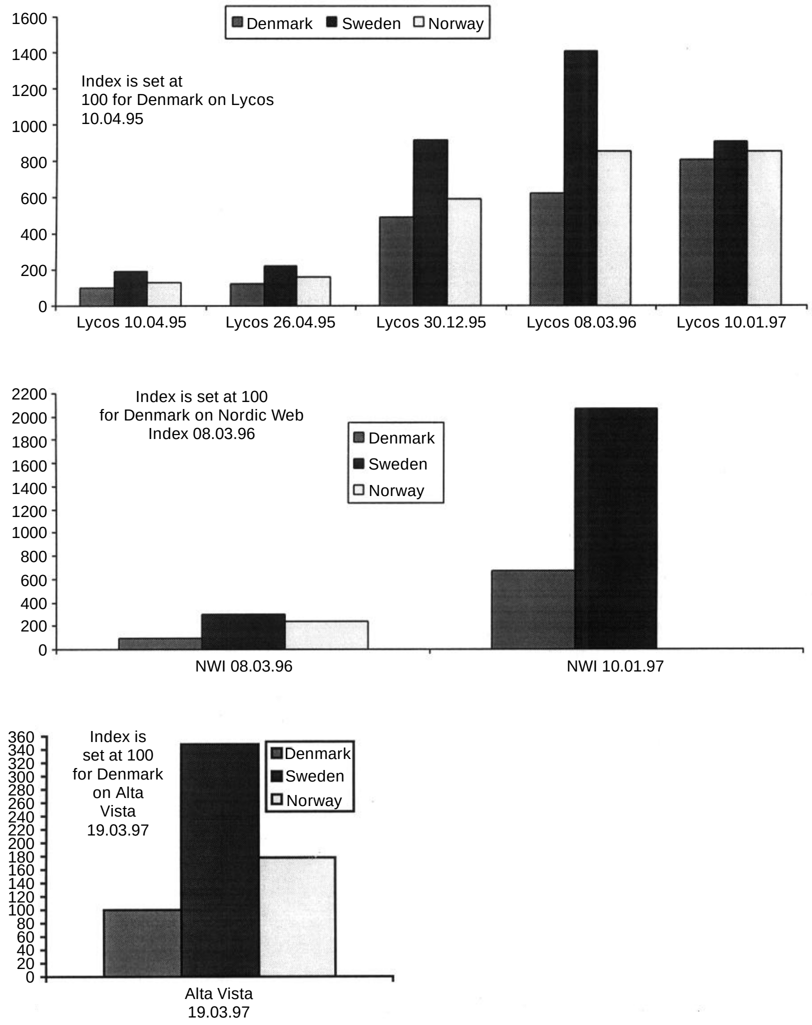

Figure 4A – 4C. *The relationships between Denmark, Sweden and Norway for the WWW plotted against time by means of snapshots* (NWI – Nordic Web Index)

Journal of Documentation, Vol. 53, No. 4, September 1997

	Denmark		*Sweden*		*Norway*	
	No.	*Index*	*No.*	*Index*	*No.*	*Index*
Lycos 10.04.95	1,957	100*	3,692	189	2,597	133
Lycos 26.04.95	2,469	126	4,439	227	3,164	162
Lycos 30.12.95	9,708	496	17,862	913	11,656	596
Lycos 08.03.96	12,156	621	27,573	1409	16,770	857
Lycos 10.01.97	15,838	809	17,752**	907	15,771**	806
Nordic Web Index 08.03.96	50,201	100*	152,238	303	120,029	239
Nordic Web Index 10.01.97	333,979	665	1,036,379	2,064	–***	–
Alta Vista (Host:) 19.03.97	139,109	100	484,134	348	248,224	178

(* = index base-line = 100), (** There is no explanation for the drop for Sweden and Norway) and (*** Norway is no longer a part of the Nordic Web Index service)

Figure 4D. *Relationships between Denmark, Sweden and Norway for the* WWW *plotted against time by means of snapshots. Numbers and corresponding index figures*

which is probably a more accurate picture, as this index concentrates more on the Scandinavian part of the WWW. It demonstrates that Denmark is falling behind the other Nordic countries, perhaps more rapidly than shown on Figure 4A. This can be stated although Norway is not represented in the last search, because of changes in coverage for the Nordic Web Index.

Figure 4C from Alta Vista using the host: search facility shows that the changes for the last Lycos and Nordic Web Index searches do not show a change in the distribution between the countries, but rather changes in the indexing of the Web indexes. The Alta Vista (Figure 4C) search shows a similar distribution to the Lycos and Nordic Web Index searches. The results from Dialog, Figure 3A, are confirmed by an investigation made by Olsen *et al.* [18, p. 12], which shows a proportional distribution of publications between Denmark, Sweden and Norway of 1 : 2 : 0.65.

The most interesting result is the inverted distribution between Denmark and Norway with respect to R&D publications and Web distribution. The proportion 1 : 0.65 for R&D publications between Denmark and Norway can be explained by the fact that Danish researchers are better at getting their results published in international journals and the number of researchers is higher. The inverted proportion for the WWW visibility – with Norway as the much more exploiting country in all snapshots except for the last search in Lycos, (Figures 4A–4C) – must then be explained by the assumption that a wider range of Norwegian public and private sectors and scientific domains make more entries on the WWW than the equivalent sectors do in Denmark. Until now, Norway would also seem to have exploited the Net more rapidly than Denmark.

Sweden's high share of both the WWW and Dialog is certainly due to the fact

Aslib, The Association for Information Management
Information House, 20–24 Old Street, London EC1V 9AP
Tel: +44 (0) 171 253 4488, Fax: +44 (0) 171 430 0514
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 53, No. 4, September 1997

Table 1. *The relationship between Denmark, Sweden and Norway for* GDP, *population, research output and Web pages*

Proportions	*Denmark*	*Sweden*	*Norway*
Population	1	1.7	0.8
Gross domestic product (1995)*	1(1)*	1.6(1.3)*	1(0.85)*
Dialog (Citation indexes 10.01.97)	1	1.9	0.65
WWW (Nordic Web Index 08.03.96)	1	3.0	2.4

The numbers marked with * are the 1995 prices.

that Sweden's population (8.84 million) is around the same as both Norway and Denmark combined (4.37 and 5.25 million), and so has much larger industrial, IT and research sectors than these countries. Another measure is the gross domestic product (GDP), taken from 1995 in billion dollars at 1990 prices and 1990 exchange rates, where the GDP for Sweden is 231.4, Denmark 142.6 and Norway 137.3, which correlate nicely with the population and Dialog proportions. At 1995 prices the GDP is: Denmark 172.9, Sweden 229.1 and Norway 146.6 billion dollars, which gives a different picture from the 1990 prices, largely due to the weak Swedish krone.

The distribution shows that the R&D proportions are almost similar to the population proportions, but that Danish visibility on the Web does not match the expectations arising from both the population and the R&D position.

5.2 Analysis by large centres of learning in Denmark

The results of the analysis by large centres of learning in Denmark on the WWW, Figure 5, show large changes occurring for each individual town in its presence on the Web. These differences cannot be explained by the differences in what is measured in the ISI files and on the Web. This is due to shortcomings in the method used for the first snapshot from the Web (April 1995); secondly, it is believed that changes in the management and organisation of the WWW in some institutions have caused alterations. These changes between towns are not visible in the ISI searches. Here, the distribution is quite stable.

It is worth noticing that the town of Ålborg is over-represented in both WWW snapshots compared to its presence in the ISI files; on the other hand, the presence of the town of Copenhagen on the Web is less than could be expected from the ISI searches. These differences can be explained by differences in institution policies for the use of the Web. They can also be explained by the fact that large universities produce a great number of R&D publications but may not have more Web pages than smaller institutions. The Web survey, which was performed on the chosen samples, shows an increase in the private sector Web pages found in the 'other' category, moving from 19.3 % in April 1995 to 28 % in December 1995. It is believed that the private sector portion of the Web pages will continue to grow at a faster pace than the academic part.

Aslib, The Association for Information Management
Information House, 20–24 Old Street, London EC1V 9AP
Tel: +44 (0) 171 253 4488, Fax: +44 (0) 171 430 0514
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 53, No. 4, September 1997

	ISI April 1995		*ISI March 1996*		*WWW April 1995*		*WWW Dec. 1995*	
	No.	*%*	*No.*	*%*	*No.*	*%*	*No.*	*%*
Ålborg	936	3.0	1,128	3.0	86	22.2	60	30
Århus	5,732	18.4	6,948	18.6	110	28.4	28	14
Odense	2,381	7.6	2,863	7.7	9	2.3	5	2.5
Roskilde	1,736	5.6	2,141	5.7	6	1.5	2	1
København	14,209	45.6	16,792	44.9	46	11.9	29	14.5
Lyngby	2,685	8.6	3,208	8.6	56	14.4	20	10
Other locations in DK	3,468	11.1	4,319	11.5	–	–	–	–
Private Sector	–	–	–	–	75	19.3	56	28
Total	31,147	99.9	37,399	100	388*	100	200	100

(* This number represents the whole population from the April 1995 analysis [16])

Figure 5. *Distribution by chosen Danish towns of learning*

5.3 Analysis by domain

Remarkable changes have occurred for some individual disciplines in the analysis of the Web samples, as shown in Figure 6. The social sciences have

	ISI March 1996		*Web pages April 1995*		*Web pages Dec. 1995*	
	No.	*%*	*No.*	*%*	*No.*	*%*
Science (not subclassified)			19	19.4	12	6.0
Medicine/Clinical medicine	26,443	58.1	3	3.1	5	2.5
Physics/Chemistry/ Inorganic chemistry	8,747	19.2	4	4.1	8	4.0
Agriculture	5,648	12.4	0	0	2	1.0
Engineering (including computer)	2,872	6.3	50	51.0	103	51.5
Computer	–	–	43	43.9	85	42.5
Total Science & Technology (including overlap)	41,400 (43,710)	91.0	61 (76)	62.2	122 (130)	61
Social science	3,251	7.1	7	7.1	25	22.5
Humanities	837	1.8	2	2.0	9	4.5
Other	1,134	2.5	30	30.6	46	23
Total (including overlap)	45,488 (48,932)	102.4	98 (115)	158.1	200 (210)	136

Figure 6. *Distributions of scientific domains with respect to the citation indexes and the WWW samples (Denmark)*

Aslib, The Association for Information Management
Information House, 20–24 Old Street, London EC1V 9AP
Tel: +44 (0) 171 253 4488, Fax: +44 (0) 171 430 0514
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 53, No. 4, September 1997

seen a dramatic growth, humanities has seen a smaller increase. For the science and technology domains it is not surprising to observe that the engineering domain contains the most pages, and that the majority of this domain consists of computer-related Web pages. In comparison with the citation indexes there are two disciplines which fare particularly badly on the WWW, these being Physics/Chemistry/Inorganic chemistry and Agriculture. This leads to the assumption that there is not the same distribution and visibility over domains in the traditional databases and on the Web.

5.4 Document types

The classification of document types, Figure 7, shows a marked difference to the figures given by Almind [16] in April 1995. The reason for this large difference is due to the small population, only 400 Web pages, used in the first investigation, most of these having been found via Web indexes which only carried out a superficial search of Danish Web servers. In December 1995, adding to these indexes, a new Web index was used – the Nordic Web Index – which carried out a far more thorough search of each individual Web server. This means that the population of Danish Web pages is now up to approximately 47,200 pages. It is therefore plausible to assume that the measurement made in December 1995 is more accurate than that taken in April 1995. It also seems more plausible that 9.5 % of the sample consists of home pages instead of the previous figure of 39.2 %. It is also more realistic that 79.5% of Web pages consist of resources, rather than 39.2%. This indicates that for each home page there are about nine other pages; thus it may be said that every home page is associated with about nine other pages.

			April 1995		*December 1995*	
			No.	*%*	*No.*	*%*
Web pages	Home pages	Subject defined	11	9.2	2	1
		For organisations	20	16.7	7	3.5
		Personal	16	13.3	10	5
	Pointer documents		26	21.7	34	17
	Resources (text, graphic etc.)		47	39.2	159	79.5
Total + overlap			100+20	100	200+12	100

Figure 7. *Web pages classified by document type, from the two snapshots of the Web*

Aslib, The Association for Information Management
Information House, 20–24 Old Street, London EC1V 9AP
Tel: +44 (0) 171 253 4488, Fax: +44 (0) 171 430 0514
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 53, No. 4, September 1997

5.5 *Frequency distribution for Web pages*

To display a simple picture of how the Web page size was distributed in the sample investigated, Figures 8–11 have been constructed. These distributions show the mean size for the types and domains.

It may be noted from Figure 8A that the home pages are smaller then the pointer documents and resources. The results of the size by domain shows, not suprisingly, that Humanities shows the biggest average size. The very small sizes for domains such as Inorganic chemistry and Agriculture may have been skewed by the fact that the sample contains very few of those.

Another aspect investigated is the number of links per page and the size per link – or Link Density – shown in Figures 9 and 10.

The number of links per page (Figure 9) can easily be compared to more traditional investigations. In 1973 Donahue [19, p. 24] found that the documents represented in the citation indexes contained an average of nearly fifteen citations per document. The result of our analysis is that the average number of links per page is a little under ten. The difference is most likely due

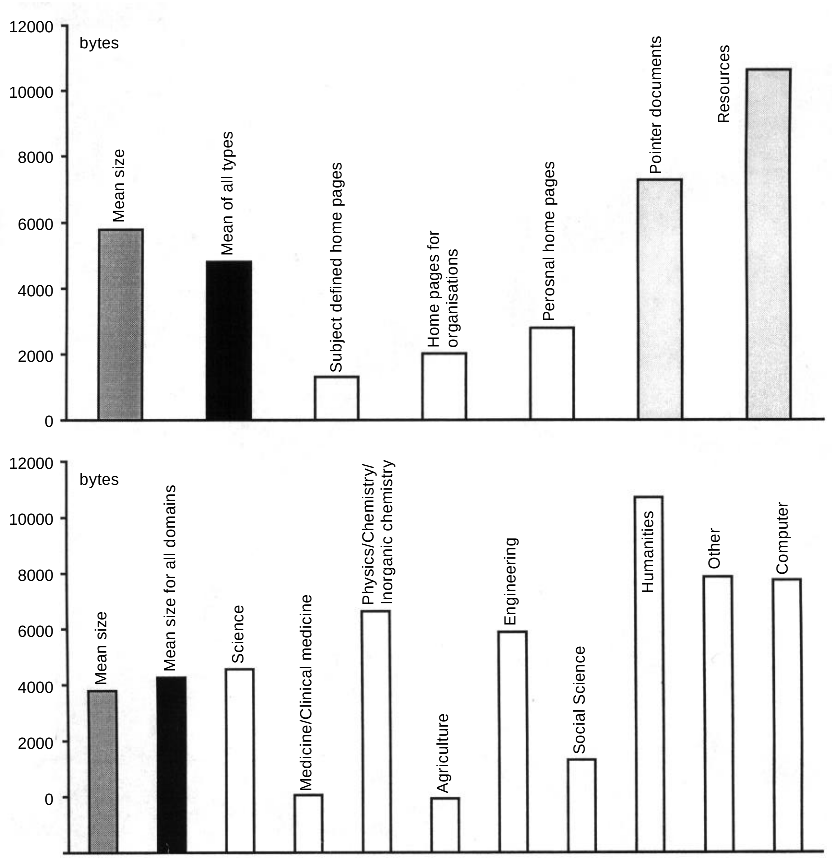

Figure 8A–8B. *Mean size of Web pages by type and domain, December 1995*

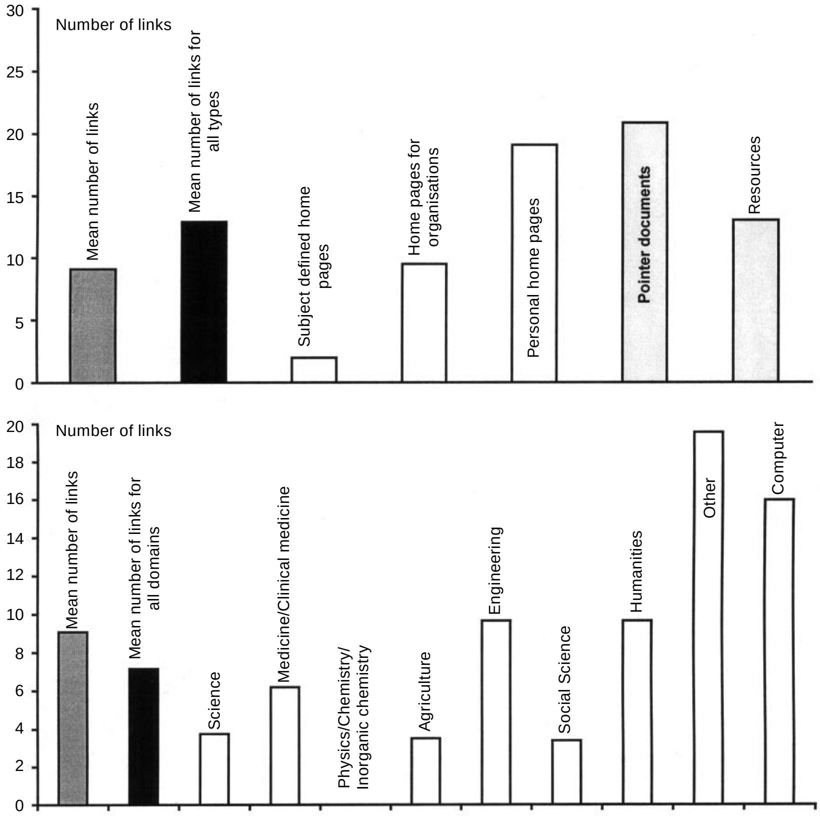

Figure 9A–9B. *Mean number of hyperlinks per Web page, December 1995*

to the fact that the information entities investigated do not contain the same types of information, and do not try to satisfy the same needs.

The Link Density, which measures the size per link ratio (Figure 10) is an interesting new measure as it brings together and normalises the two measures of size and number of links. The smaller the number of bytes the lower the link density for equal numbers of links. For instance, although both organisational and personal home pages are smaller than the average size (Figure 8A) they contain such a low number of links on average that they both display a very low Link Density (approximately 200 bytes per link). Consequently, the visibility or marketing of the ego on those home pages is merely done by referring to other pages on the Net. In contrast, subject home pages are also small in size but do not link up to other pages to a great extent. Their Link Density is hence quite high and above average. This home page type is far more descriptive and self-contained. The pointer documents are interesting with respect to Link Density. By definition, they ought to contain many pointers to other locations on the Net. That characteristic is certainly a fact (twenty links per page, Figure 9A). However, due to their large size (Figure

Journal of Documentation, Vol. 53, No. 4, September 1997

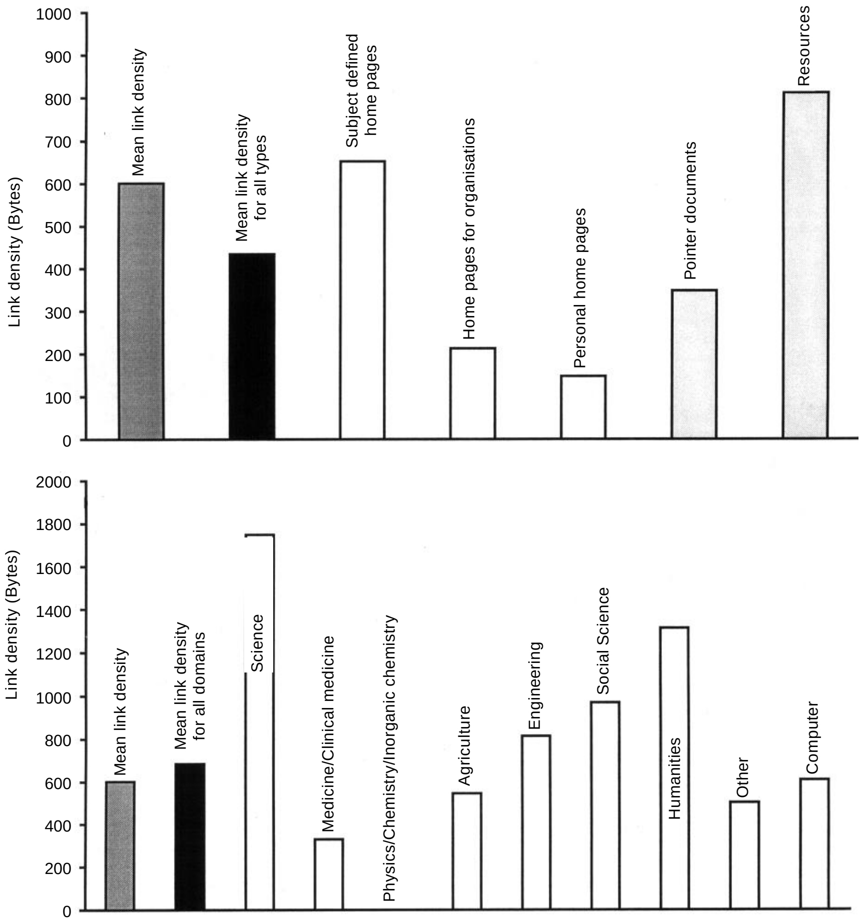

Figure 10A–10B. *Link Density in Web pages, distributed over Type and Domain, December 1995*

8A), the pointer pages on average are surprisingly less dense than the above mentioned home page types (Figure 10A): generally speaking, much more text is used on pointer pages to describe and lead up to a link than in the personal and organisational home pages types. The same phenomenon can be observed for the humanities and social sciences domains (Figure 10B). This characteristic corresponds to the publication patterns common to these fields. Not surprisingly the Link Density for resources shows that they are larger than the average in Figure 8 and have very few links.

To sum up these analyses:

1. the average Web page is 5,779.55 bytes in size;
2. it has 9.09 links per page on average;
3. the Link Density measure tells that the average Web page contains 635 bytes per link in the Web page.

Aslib, The Association for Information Management
Information House, 20–24 Old Street, London EC1V 9AP
Tel: +44 (0) 171 253 4488, Fax: +44 (0) 171 430 0514
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 53, No. 4, September 1997

The results that deviate the most are evidently for those groups that have a very small number of Web pages forming part of the sample.

Others have made similar surveys, but made use of different statistics for the Web and Web pages. One such survey has been done by Bray [20]. He used 1.5 million objects from the search engine Open Text Index as a sample. The other survey was performed by Woodruf *et al*. [21]. They used data collected by the Inktomi search engine which comprised 2.6 million HTML documents at the time of the investigation (1995). Both the Bray and the Woodruff studies measured the average size for Web pages, Bray reporting the average size to be 6,518 bytes, Woodruff *et al*. to be 4,500 bytes. When these numbers are compared to the average size that we found, 5,779 bytes, one may claim that the differences are due to differences in what is measured and a very high standard deviation. Woodruff *et al*. found the average number of links per page to be seventeen. We found the average number of links per page to be nine. The difference is mainly due to the fact that the Woodruff investigation only investigated HTML documents, whereas our study also covered non-HTML objects, such as data programmes, which cannot contain links.

6. DISCUSSION

One of the aspects of informetric analysis of the Web is the importance of, and difficulties connected with, carrying out a comprehensive data collection.

	Size (bytes)	*Number of links*	*Size per link (bytes)*
Total	5,779.55	9.09	635.81
Type	4,804.50	12.91	372.15
Subject defined home pages	1,304.00	2.00	652.00
Home pages for organisations	2,017.38	9.50	212.36
Personal home pages	2,799.50	19.10	146.57
Pointer documents	7,279.13	20.88	348.62
Resources	10,622.50	13.06	813.36
Domain	6,267.07	7.17	874.07
Science	6,554.67	3.75	1,747.91
Medicine/Clinical medicine	2,042.60	6.20	329.45
Physics/Chemistry/Inorganic chemistry	8,630.63	0.00	–
Agriculture	1,910.00	3.50	545.71
Engineering	7,875.06	9.67	814.38
Social Science	3,298.32	3.40	970.09
Humanities	12,701.00	9.67	1,313.44
Other	9,847.88	19.57	503.21
Computer	9,737.50	15.98	609.36

Figure 11. *Data for Figures 8 to 10*

Journal of Documentation, Vol. 53, No. 4, September 1997

The problems with data collection occur primarily for two reasons. Firstly, the current tools available are not good enough. It is only possible to carry out a fully comprehensive data collection manually: every single URL must be compared against those URLs which are already known. This problem will be minimised by the evolution of more sophisticated search engines, with enhanced capabilities of performing very detailed searches as well as combining them. The first improvements have become visible in facilities such as those of Alta Vista [22]. In this engine it is possible to search for specific elements such as ‘link:’ and ‘host:’, so that a count of citations to an institution, say The Royal School of Librarianship, will be found by searching for link:*.db.dk, where db.dk is the host name. The number of Web pages on the Web server of the Royal School of Librarianship will be found by searching for host:*.db.dk. Lastly the number of citations from other hosts only to The Royal School of Librarianship can be found by searching for link:*.db.dk and not host:*.db.dk.

Secondly, the means used for collecting data, both manually and with the use of Web indexes, have the effect that there are Web pages which will not be found. All together this makes it very cumbersome to perform the very vital data collection activity.

Figure 12 illustrates that a document which is not cited or which does not contain citations (page X) can only be found by direct access to the Web server’s file hierarchy. This is because Web indexes collect URLs by following links to and from individual Web pages. They get access to a Web server either via links or by starting at the highest Web page in the file hierarchy. Moreover, Web pages which only contain citations and are not cited in other documents can also only be found via direct access to a Web server’s file hierarchy. It may be said that it is not particularly productive to design a Web page which neither cites nor is cited. More likely is the situation where there is a small collection of Web pages which are not cited outside the group. Figure 12 could illustrate such a cluster that neither cites nor is cited outside itself. Nor will the individual Web pages be found if there are no outside citations to any of them. Figure 12 therefore illustrates that it is worth using several different data collecting

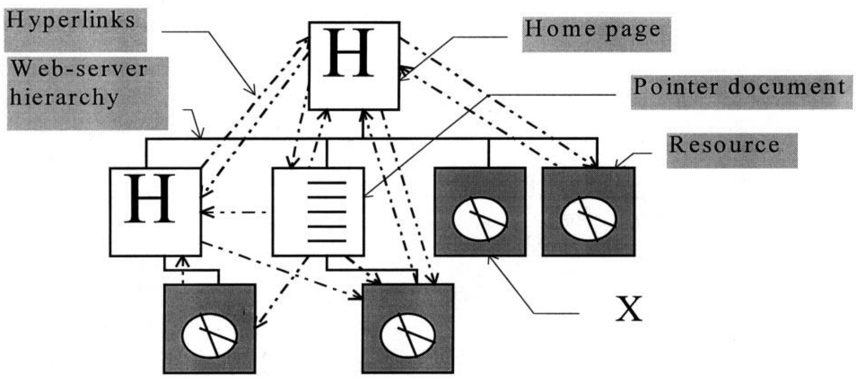

Figure 12. *Simplified model of relationships between Web pages*

Aslib, The Association for Information Management
Information House, 20–24 Old Street, London EC1V 9AP
Tel: +44 (0) 171 253 4488, Fax: +44 (0) 171 430 0514
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

methods and sources. The principle can be easily related to the way polyrepresentation is used within IR [11].

One of the problems with informetric analyses of the WWW is that they can only be carried out asynchronously. This problem can only be solved if databases and statistics are created which illustrate the development of the WWW over time. Finally, it is also difficult to isolate *all* the citations to one particular Web page or a cluster of pages.

7. CONCLUSION

The previous sections have confirmed that it is possible to use informetric methods on the WWW, i.e. to perform Webometric studies.

The case study demonstrates a method which can be used for Webometric analyses. It consists of defining the population which is to be investigated. All the relevant URLs are then collected and from them a sample for further analysis is chosen. The reason it is so important to collect all Web pages falling within the original definition is to avoid bias in the final sample.

The method for each individual analysis mainly consists of extracting the data elements that are to be investigated. Data collection and data extraction are very time consuming. At present, it is not possible to automate these processes.

The case study has drawn a picture of Denmark's use of the WWW compared to both other countries and traditional databases, together with a view of the types of Web page, discipline, size and number of links. The picture shown is very varied. Denmark's position on the WWW compared with its position in traditional databases is much weaker than those of Sweden and Norway. Generally speaking, the analysis by type shows that 75% of the Web pages investigated are resources, and that only 9% are home pages. The investigation of size and number of hyperlinks shows that the average Web page size is about 6,000 bytes and has about 9 hyperlinks per Web page. The number of links per Web page is smaller than the number of citations found in traditional academic texts. About 40 % of the Web pages from the sample had no hyperlinks. All the Web pages with no links are resources, i.e. texts, images, sound, objects. The reason for that is simply that non-HTML pages cannot link to other pages.

This investigation has not been concerned with why citations are given on the WWW, as we consider the amount of data available to be too small to be able to describe anything about the use of citations on the WWW.

Citation analysis on the WWW has not been tested in practice. The problem with citation analysis on the WWW is to find a collection of Web pages that are tightly enough linked, so that there is something to measure.

Further it is important to understand the idea of using the WWW and HTML as analytic tools which can be used for many purposes, in addition to their use in a purely search or mark-up context. It became very obvious during the work that informetric methods on the WWW can be used for such dissimilar tasks as issue management, gathering of business intelligence and research evaluation,

Aslib, The Association for Information Management
Information House, 20–24 Old Street, London EC1V 9AP
Tel: +44 (0) 171 253 4488, Fax: +44 (0) 171 430 0514
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 53, No. 4, September 1997

which have already been investigated by Cronin and McKim [23] and Cronin *et al*. [24].

The fact that data structures on the WWW are unstructured can be considered from an analytic viewpoint to be an advantage, as this means that it is possible to carry out analyses which cannot be made using databases with highly structured fields and data structures.

REFERENCES

1. Camron, R.D. A universal citation database as a catalyst for reform in scholarly communication. http://elib.cs.sfu.ca/projects/ElectronicLibrary/project/papers/citebase/citebase.html, 1995.
2. Larson, R.R. Bibliometrics of the World Wide Web: an exploratory analysis of the intellectual structure of cyberspace. *In: Proceedings of the 59th ASIS Annual Meeting, Baltimore, Maryland, 1996*. http://Sherlock.berkeley.edu/asis96/asis96.html.
3. Paisley, W. The future of bibliometrics. *In*: Borgman, C.L., *ed. Scholarly communication and bibliometrics*. Sage, 1990, 281–299.
4. Tague-Sutcliffe, J. An introduction to informetrics. *Information Processing & Management*. *28*(1), 1992, 1–3.
5. Graham, I.S. *The HTML sourcebook*. John Wiley & Sons, 1995.
6. Powell, J. Spinning the World-Wide Web: an HTML primer. *Database, 18*(1), 1995, 54–59.
7. Raggett, D. HyperText Markup Language Specification Version 3.0. http://www.w3.org/hypertext/www/MarkUP/html3/html3.txt.
8. McMurdo, G. The Net by numbers. *Journal of Information Science, 22*(5), 1996, 381–390.
9. Ingwersen, P. *and* Hjortgaard Christensen, F. Data set isolation for bibliometric online analysis of research publications. Fundamental methodological issues. *Journal of the American Society for Information Science*, *48*(3), 1997, 205–217.
10. Ardö, A. Nordic WebIndex – projektbeskrivning. http://www.ub2.lu.se/NNC/projects/NWI/ [11.12.95], 1995.
11. Ingwersen, P. Cognitive perspectives of information retrieval interaction: elements of a cognitive IR theory. *Journal of Documentation*, *52*(1), 1996, 3–50.
12. Coutois, M.P., Baer, W.M. *and* Stark, M. Cool tools for searching the web. *Online*, *19*(6), 1995, 14–32.
13. Scales, B.J. *and* Felt, E.C. Diversity on the World Wide Web: using robots to search the Web. *Library Software Review*, *14*(3), 1995, 132–136.
14. Winship, I.R. World-Wide Web searching tools: an evaluation. *VINE, 99* (June), 1995, 49–54.
15. Ardö, A. Nordic WebIndex (experimental service). http://www.ub2.lu.se/NNC/projects/NWI/NNCdemo.html [11.12.95], 1995.
16. Almind, T.C. *Informetrisk undersøgelse af den danske del af World Wide Web*. Masters Thesis, Danmarks Biblioteksskole, Copenhagen, 1995.

Aslib, The Association for Information Management
Information House, 20–24 Old Street, London EC1V 9AP
Tel: +44 (0) 171 253 4488, Fax: +44 (0) 171 430 0514
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 53, No. 4, September 1997

17. Lycos. http://www.lycos.com.

18. Olsen, T.B., Foss Hansen, H., Luukkonen, T., Persson, O. *and* Sivertsen, G. *Nordisk forskning i internasjonal sammenheng: – en bibliometrisk beskrivelse av publicering og siteringer i naturvitenskapelig og medisinsk forskning*. Nordisk Ministerråd (TemaNord 1994:618), 1994.

19. Donahue, J.C. *Understanding scientific literatures: a bibliometric approach.* MIT Press, 1973.

20. Bray, T. Measuring the Web. *Fifth International World Wide Web Conference, Paris, France, May 6–10, 1996*. http://www5conf.inria.fr/fich_html/ papers/p9/overview.html.

21. Woodruff, A., Aoki, P.M., Brewer, E., Gauthier, P. *and* Rowel, A. An investigation of documents from the World Wide Web. *Fifth International World Wide Web Conference, Paris, France, May 6–10, 1996*. http://www5conf.inria.fr/fich_html/papers/p7/overview.html.

22. Altavista. http://altavista.digital.com.

23. Cronin, B. *and* McKim, G. Markets, competition, and intelligence on the World Wide Web. *Competitive Intelligence Review*, *7*(1), 1996, 45–51.

24. Cronin, B., Overfelt, K., Fouchereaux, K., Manzvanzvike, T., Cha, M. *and* Sona, E. The Internet and competitive intelligence: a survey of current practice. *International Journal of Information Management*, *14*, 1994, 204–222.

(Revised version received 11 April 1997)

Aslib, The Association for Information Management
Information House, 20–24 Old Street, London EC1V 9AP
Tel: +44 (0) 171 253 4488, Fax: +44 (0) 171 430 0514
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib