International Journal of Scientometrics, Informetrics and Bibliometrics
ISSN 1137-5019

VOLUME 10 (2006): ISSUE 1. PAPER 6

Web indicators – a new generation of S&T indicators?

Andrea Scharnhorst*
Paul Wouters**
The Virtual Knowledge Studio
Cruquiusweg 31
1019 AT Amsterdam
The Netherlands
*E-mail: andrea.scharnhorst@vks.knaw.nl
Web: www.virtualknowledgestudio.nl/en/vks_members/homepage_andrea_scharnhorst/
**E-mail: paul.wouters@vks.knaw.nl

Abstract

In this paper we discuss the need and possibilities for the development of web indicators for the characterization of emergent features of ICT-mediated scholarly practices based on the European WISER project [1]. We summarize these developments as a turn to e-science. We place the research of web indicators in a virtual landscape of measuring and exploring the Internet and the Web. and provide an overview of webometrics research field from this point of view. We raise the issue of how the self-organization and meta-stability of the worldwide web will influence the development of web indicators. Theoretically, the paper is a plea for the combining of complex network theory and semiotics. In our view, this enables the combination of quantitative and qualitative analysis in the interpretation of web phenomena in innovation systems.

Keywords

Web indicators, cybermetrics, complex networks, e-science, self-organization, STI indicators, semiotics

Introduction

Information and communication technologies have become increasingly important in scientific and scholarly research (cf. Thomas, King et al. 2000; Marsh 2001). The emergence of the Internet and the Web have modified the way that scientists and scholars search for and find information. High-end computing has enhanced the research process in a number of fields (e.g. through networked scientific instruments, distributed databases and software tools). The use of Web and Internet-based communication tools has also affected the development of international networks of cooperation and collaboration. Contrary to many expectations (e.g. Nentwich 2003, Science, Technology and Industry Outlook 1998 (chapter 7)), however, ICT has not changed the central role of the traditional scientific article in the system of scholarly publication (Heimeriks 2005). Nonetheless, based on the aforementioned changes, important dimensions of research are being represented on the internet. Three aspects of the application of information and communication technologies are particularly important for the analysis of science and technology:

- The internet and the WWW have made a large number of new sources of data available, including meta data. They also provide new interfaces to existing data sources.

- The internet and the WWW are becoming the dominant platform for new communication media between researchers (including new forms of virtual collaboration) and between research and society at large.

- The internet and the WWW enable the development and use of new web-based research tools in all areas of scientific and academic research and thus create new challenges for science and technology policy.

These factors have direct implications for the development of new indicators. The new data sources provide new data for both existing indicators and for the development of new indicators which may supplement traditional science, technology and innovation indicators. The emergence of the internet and the worldwide web has prompted some modest changes in the academic publication system. Most established journals have a web presence and academic institutions are increasingly creating institutional repositories, partly in response to the Open Access movement **http://www.soros.org/openaccess/**, **http://www.openarchives.org/** (Harnad 2001). Web-based publishing may also blur the boundaries between formal and informal scientific communication. A recent report (Wouters and De Vries 2004) showed that the citation profiles of web-based documents (in computer science literature) differ to Science Citation Index (SCI) documents in this field. In web-based documents, 13% of the references referred to conference proceedings whereas in SCI-covered documents, only 3% of the references cited conference proceedings (Goodrum, McCain et al. 2001). The increase in on-line publishing and referencing may necessitate the development of new indicators. Initial attempts to explore the use of web visibility indicators have shown that measures derived from link counts on university websites vary significantly in accordance with the research ratings of the institutions (Thelwall 2004). Other examples include web citations defined as Google hits on the titles of articles (Vaughan and Shaw 2005).

The emergence of new web-based research tools may also increase the need for new tailor-made indicators for science and technology policy, in particular if major investments in advanced technologies are at stake. Important scientific activities may become partly invisible in measurements undertaken using traditional indicators. For example, if scientific research is carried out in virtual laboratories [2], measurements of national R&D expenditure and personnel will not cover all of the research infrastructure. There may also be an increased need for electronic infrastructural investments in scientific and academic research, in terms of both hardware and software. Traditional S&T indicators may very well underestimate the demand for resources. For example, it is clear that databases play an important role in the process of knowledge creation in a wide variety of fields (Arzberger, Schroeder et al. 2004). Yet the need for investments in database infrastructures and standards are usually not captured by traditional S&T indicators.

Science and technology indicators

Indicator research generally needs a clear and methodologically sound understanding of the basic social processes involved in science and technology. The word “indicator” has roots in the Latin “indicare” which means “to direct or to point out”. It is usually defined as a measurable variable, sometimes hypothetically linked to another (latent) variable which cannot be observed directly (Bollen 2001). The variable describes a characteristic of a social phenomenon which is used, in turn, to define, explain or describe a social science concept in theoretical terms.

Science, technology and innovation (STI) indicators characterize particular characteristics of innovation processes (Kleinknecht, Montford et al. 2002; Godin 2004). They cover all aspects of research, technology and innovation. STI indicators may be classified as input, output or process indicators (Nalimov and Mulchenko 1969; Dobrov 1970; Weingart and Winterhager 1984) [3]. Expenditure on research and development and the number of employees in R&D sectors are examples of input indicators while the number of publications, citations or patents are examples of output indictors. While input and output indicators are among the conventional indicators used in

science and technology documents (European Report on Science & Technology Indicators 2003; Key Figures : Towards a European Research Area ; Science, Technology and Innovation 2004; OECD Factbook 2005), process indicators are less clearly defined. To observe a process, static indicators need to be measured at different points in time. Alternatively, it is possible to construct time-related indicators, such as the relaxation time of processes, (for an approach to making temporal patterns visible, see Liang 2005). For example, some authors propose the use of static co-citation maps as the basis for process indicators indicating the formation of research fronts (Weingart and Winterhager 1984).

In general, the development of STI indicators has two purposes: the scholarly study of structure and process in knowledge creation and the production of information for innovation policy (Moed, Glänzel et al. 2004). Hitherto, most STI indicators have been based on R&D statistics, patent data or bibliometric data and the “literature model” has dominated the quantitative study of scientific communication in scientometrics for many decades. The digitization of scientific publication and communication may undermine this. For example, the increased use of information and communication technologies in science (e.g., on-line publications, digital data production and simulations) is challenging the model of the “journal article” as the prevailing form of scientific communication. The shift of attention to science-based technology and innovation in the 1980s has led to the construction of innovation indicators for researching patent databases. The “embeddedness” of science in technology, and vice versa, can be traced by analysing the differences and asymmetries in citation patterns between the domains of scientific literature and patenting (Schmoch 1997; Grupp and Schmoch 1999; Meyer 2000a; Meyer 2000b). The introduction of web indicators is a further step towards dynamic process monitoring as opposed to static input/output monitoring (Ingwersen and Björneborn 2004). In a recent paper, Day discussed strategies for implementing usage statistics or webometric link data analysis based on digital repositories as e-print archives for research evaluation (Day 2005). This development was triggered by the emergence of funding policies aimed at creating new digital informational infrastructures for research: the turn to e-science.

Indicators and the turn to e-science

The notion of a “turn to e-science” was used in a report about a Strategic Planning Workshop on Future and Emerging Technologies in Brussels April 2001: “Future Science will be conducted largely in the Computer (e-Science) and Computing underpins every other discipline” (FP6 Strategic Planning Workshop, 2001). However, the precise nature of the interaction between the new information and communication technologies (ICTs) and scientific and scholarly research is often obscure. The application of new information and communication technologies, both in research and in other areas, has often been accompanied by overoptimistic expectations. As a result, policy decisions with regard to the application of, and investment in, new information tools and technologies run the risk of being guided more by wishful thinking than sound judgement. It must also be taken into account that statistics are also socially constructed, as the choice as to what to measure and how to measure it depends on relevance and feasibility and are not arbitrary (Godin 2004).

e-Science is generally defined as the combination of three different developments: the sharing of computational resources, distributed access to massive data sets and the use of digital platforms for collaboration and communication (Hey and Trefethen 2002). The precise definition varies between the UK, the US and the Netherlands. The UK definition focuses on the analysis of massive data sets. In the US, the notion of “cyberinfrastructure” is central. In the e-science Dutch initiatives, the e does not stand primarily for “electronic”, but for “enhancement” [4]. Whatever the precise definition, the core idea of the e-science movement (most of which remains in the realm of promise rather than reality) is that knowledge production will be enhanced by the combination of pooled human expertise, data and sources, and computational and visualization tools. This is accomplished by transferring the entire research process “into” the digital environment. To be more precise, all stages of the research process are being inscribed into a digital work environment. This includes digital sensors and measurement devices, digital analytical instruments, digital communication between researchers, both within the research group and with external colleagues, and digital information services, within which the publications produced by the

researchers also find their place. This does not mean that the collection of digital representations is a perfect copy of the inevitably messy process that the creation of knowledge always involves. The way some features of research practice will be moved to the background while others are highlighted is an interesting research problem in itself. It does mean, however, that a much larger area of raw materials is emerging for science and technology indicators.

This provides ample space for a focused analysis of the role of digital traces as the raw materials for science and technology indicators. In an earlier study, we demonstrated how the role and meaning of citation analysis can be analysed from a semiotic perspective (Wouters 1999). This becomes even more relevant in the analysis of e-science because what is involved here is not merely a matter of translating traditional existing science and technology indicators to the internet or the web. The core features of e-science will create a space for potential indicators which is not only much larger than the existing indicator space, but also different in character. The study and development of new indicators in this space involves trying to understand the role of these digital traces in the research process itself. It is our expectation that complex systems theory will help us to understand core features of the dynamics of these indicator spaces, partly by simulating key questions. The role of digital traces and of indicators based on these traces also requires a qualitative understanding. It is here that the combination of semiotics with the sociology of scientific knowledge can make an important contribution. Semiotics is understood here as the systematic study of the role of signs in social life (Eco 1976).

The key issue in indicator research is that signs, such as the citation frequency or the number of hits on a web page, do not have a fixed function or meaning. The signs can be linked to each other in more or less arbitrary ways. The history of citation analysis and the creation of the citation index has provided many illustrations of the active creation of meaning by linking different indicators to different practices (Wouters 2000). This has resulted, for example, in the common notion that the impact factor of a journal is a good indicator of scientific quality, irrespective of the social and cognitive structure of the field.

This will be relevant to the turn to e-science because the digitization of research practices and infrastructures will lead to an abundance of digital data that can be used to construct a huge variety of STI indicators. This may affect the political economy of data sets for indicator construction. For example, in the long run, fully-automated web-based citation indexing may threaten Thomson Scientific's monopoly on citation indexes. This may lead to a proliferation of, for example, "amateur citation indexing", which, in turn, may create new problems with regard to the maintenance and governance of indicator quality in the field of scientometrics. It will also render more acute the question as to which data should be used for indicator construction and which should definitely not be used in such a way. Thus, we expect that the further development of digitized platforms for research practice, collaboration and communication could intensify both the construction of STI indicators and the need for its critique.

In this paper we discuss web indicators mainly from the perspective of process indicators. In particular, we refer to complex network theory as a heuristic tool for the methodological foundation of web indicators. Web indicators can be used to address changes in the process of web-based academic publishing and communication, in the information-seeking behaviour of scholars, in new patterns of international collaboration and in new research tools such as high-end computing and databases. All of these changes leave traces on the web and can be analyzed using webometric methods. Our paper also refers to the approach adopted by and initial findings of an EU-funded project on "Web indicators for scientific technological and innovation research" (WISER). Before turning to our experience in the context of this project, however, we would like to discuss the development of the field of webometrics.

The development of webometrics

A growing body of literature on the measurement of science and technology activities on the worldwide web using informetric, bibliometric and scientometric methods has been emerging since the mid-1990s. The term

“webometrics” itself was coined in 1997 (Almind and Ingwersen 1997). Since then, the informetric community has become involved in the investigation of the new electronic media, including the internet (Larson 1996; Rousseau 1997; Ingwersen 1998; Egghe 2000; Thomas and Willet 2000; Björneborn and Ingwersen 2001; Cronin 2001; Bar-Ilan and Peritz 2002). The reason for this development is rather obvious: if academic and scientific research and communication are increasingly shaped by and shaping the internet, analysis focused on printed media will miss out on an important dimension of research. The development of web indicators is a growing research area within webometrics. It was formally initiated with the foundation of the e-journal Cybermetrics by Isidro Aguillo and others in 1997 (Thelwall, Vaughan et al. 2005). The field is characterized by the search for methodological approaches and the testing of the reliability and validity of possible indicators, on the one hand, and insight into the nature of web indicators based on experiments and exploration, on the other. Although traditional S&T indicators have proven deficient with respect to a number of important dimensions of science in the internet era, the development of appropriate web or net indicators is still in its infancy (Björneborn 2004; Thelwall 2004a). The development of this new specialty of “webometrics” or “cybermetrics” has been documented in several recent reviews (Thelwall 2004a; Thelwall, Vaughan et al. 2005). Hitherto, the main topics of webometrics have highlighted issues concerning the best methods of data collection and the use of search engines (Snyder and Rosenbaum 1999), the problem of transferring terms like “citation” to the world of the web (“sitation”) and the definition of impact factors for electronic journals (Rousseau 1999).

It is our belief that the development of web indicators is not completely covered by webometrics. We are convinced that web indicators should be developed by combining qualitative research (for example web surveys and virtual ethnography) with quantitative research (webometrics, information science and complex networks theory). The growing body of literature on webometrics has not yet found access into the world of established STI indicators. Part of the challenge of webometrics is caused by the high state of flux of the internet. Also, despite being widely studied, the basic mechanisms and structural properties of web users are still not very well understood. This will make the development of web-based indicators a challenging undertaking for years to come. Much work in webometrics has been based on a more or less direct translation of research questions and methods from bibliometrics and scientrometrics to the world of online publishing and communication. In our view, however, the field of webometrics may profit from a more systematic incorporation of work from the fields of statistical web analysis, statistical physics and complex networks theory. In the following sections, we provide examples of this type of work and discuss their implications for webometrics and web indicators.

The growth of the web has been measured by the increase in the number of hosts (**http://www.isc.org/**, Internet Domain Survey). Host surveys have also been produced for Europe (**http://www.ripe.net/**), by national initiatives (e.g. for France, **http://www.nic.fr/statistiques/**) and for commercial activities (e.g. **http://www.netcraft.com/survey/**). Measuring the web incorporates size and growth, the geographic distribution of hosts and structural properties such as the connectivity in, and the traffic on, the net. In an overview, Bray posed the following questions (Bray 1996):

- How big is the web? [5]
- What is the “average page” like?
- How richly connected it is?
- What are the biggest and most visible sites?
- What data formats are being used?
- What does the WWW look like?

Some of these questions are also relevant for STI indicator research. For example, the connectivity between the sites of different institutions can be considered as an additional expression of mutual perception and visibility. The community of web analysts has also produced different visualization tools. These maps and graphical tools have been used to support information navigation. One example of such an initiative is **http://www.cybergeography.org/** which has produced a collection of web maps (Brunn and Dodge 2001), including the map on “data traffic flows between different points on the globe” (**http://mappa.mundi.net/maps/maps_021/**) (Figure 1).

Figure 1: Map of backbone flows for Europe, from 13 May 1993 (Courtesy of Brian Reid, reproduced from **http://mappa.mundi.net/maps/maps_021/**)

The knowledge and expertise developed in this type of internet analysis can be used to visualize STI indicators.

A second important approach in the area of web analysis focuses on the statistical properties of the network and its growth dynamics. Research on the non-linear character of the internet has been strongly influenced by the information sciences and statistical physics (see, for example, Bentley and Maschner 2000; Callaway, Newman et al. 2000; Tadic 2001; Bornholdt and Schuster 2002; Pastor-Satorras and Vespignani 2004). In fact, the web has become one of the favoured objects for statistical analysis from the point of view of complex network theory. The availability of web data has triggered the emergence of this new specialty in complexity science (Scharnhorst 2003). There is ample evidence that information topology (Faloutsos, Faloutsos et al. 1999) and the flow of information (Leland, Taqqu et al. 1994) on the internet are scale independent [6] (Adamic 2000). According to Katz 1999, a scale-independent system is “a system that exhibits statistically similar characteristics when examined at the level of individual entities, collections of entities or the system as a whole”. The web has the characteristics of both a small world network and a scale-free network (Huberman 2001). The features of the web as a complex network with certain non-linearities have important consequences for the building of indicators that are robust enough to monitor a dynamic and non-linear changing system (see section on self-organization and complexity). The validity of an average in a skewed distribution poses an interesting problem. If particular characteristics of a system are not distributed according to a Gaussian equation, the mean loses its representational function of the properties of the system. Fluctuations can then no longer be considered as irrelevant. These small perturbations may have a constitutive role for the dynamics of the system. Thus, it has been proposed that the entire power law distribution be taken as a reference and that the deviations from this model be studied as the basis for indicator construction (Katz 2004; Thelwall 2004b).

Power laws are not only relevant at the level of size and link distributions of web pages (websites), they also affect enhanced indicators such as co-link patterns (Katz 2004). The existence of skewed distributions in the web graph influences effective search and surfing strategies (Adamic 2001; Maurer and Huberman 2001). This is also important for monitoring tools, on the basis of which dynamic web indicators will be built. Finally, the structure of connections between different points (link topology) in the internet affects the building of “web communities” (Adamic 2000). This has consequences for the representation of e-science communities on the web and the use of the web in e-research.

Given the easy availability of data, the study of science and technology is

increasingly seen as a particularly suitable area of application for these complex theoretical approaches (Callaway, Newman et al. 2000). Examples of the use of traditional bibliometric data in complex network theory include the study of co-authorship networks and citation networks, resulting in network characteristics such as clustering coefficients and exponents of power laws (Barabasi, Jeong et al. 2002; Goldstein, Morris et al. 2004; Sen 2004; Li, Wu et al. 2005). The extent to which this application of complex network theory to bibliometric data about science systems will actually influence traditional STI indicators and push them in the direction of scale-independent indicators remains an open, however (Katz 1999; Katz 2000; Katz 2004; Katz 2005). Webometrics approaches also increasingly incorporate complex network theory (Björneborn 2004).

In summary, the research carried out hitherto in the field of cybermetrics has mainly dealt with basic questions. Broader analyses that would provide a suitable basis for science policy reasoning are still rare. In our view, research of the development of web-based indicators would appear to be particularly promising in two respects. Firstly, the analogy between printed and electronic media has been used to transfer terms and concepts from traditional scientometrics to webometrics. However, only a few studies have attempted to investigate the link between printed and non-printed scientific products or the influence of electronic media on non-electronic media (Harter 1997; Kling 2000; Cronin 2001). This topic is important in estimating the actual influence of the use of the web on scientific communication. Secondly, the development of scientometrics was strongly influenced by the need for effective evaluation methods in the shift from “little science to big science” and the creation of new accountability regimes in science (Cozzens, Healey et al. 1990, Wouters 1999). Ultimately, the aim of webometrics will be to create methods, tools and numbers for the analysis and evaluation of science in the turn to e-science. This will have to include a critical approach to the use of web-based indicators in evaluative contexts. After all, the availability of previously inconceivable masses of data may lead to inappropriate or misplaced use of these data in evaluation exercises. For this reason, the critique of the use of S&T indicators in evaluative contexts must be taken into account, e.g. the widespread application of the Impact Factor (Schubert and Glänzel 1983; Moed and Van Leeuwen 1995).

Initial attempts at applying web indicators for evaluative purposes (Thelwall 2004a; Day 2005) show that a reflexive approach is needed (Leydesdorff and Scharnhorst 2002). In principle, all indicators should be used in a reflexive way. This means that the use of the indicator and its interpretation should include an awareness of the circumstances of its creation, the nature of the data and the nature of the processes to be described. Given the ambiguity of the web, both in terms of obtaining reliable data traces and interpreting them, extreme caution must be exercised. Web indicators should reflect the nature of the processes (in terms of the turn to e-science) they are supposed to measure.

European webometrics projects

An initial feasibility study funded by the EC stated in 1999 that “the opportunities for using informetric methods [on the Web] are not yet well elaborated” (Boudourides, Sigrist et al. 1999). This study was part of a EU-funded project “The Self-organization of the European Information Society (SOEIS)” [7]. Since then there have been enormous developments in this field, as is also demonstrated by the different funding initiatives. Since the definition of the creation of a European Research Area was defined as crucial goal in European science policy, the need for timely and relevant web-based S&T indicators has become more urgent. The internet and the web will obviously be crucial tools and media in the construction of this new research area. The availability of appropriate tools for the monitoring of the development of the European Research Area and possible bottlenecks and problems may, therefore, be rather important (Future research domains at the frontiers of science and technology: on the road to the 6th FP 2001). The creation of European collaboration and communication networks in science and scholarly research is one of the key elements of the European Research Area. This means that collaboration and connectivity indicators need to be further developed than they are at present. For example, studies on changing patterns in research networks may help to identify crucial points for concerted actions for investment in RTD infrastructure. Various initiatives have already been launched which are relevant for the creation of a European Research

Area. These include foresight/forecasting methods such as Delphi studies, expert ratings and brainstorming conferences. The development of new quantitative indicators is, however, probably an indispensable additional method for the study of newly emerging web-based phenomena.

The European Union acknowledged the importance of web-based measurements and e-science analysis in recent years through its funding of two major projects: EICSTES (European Indicators, Cyberspace and the Science-Technology-Economy System, **www.eicstes.org**, 2000-2003) and WISER (Web Indicators for Science, Technology & Innovation Research, **www.wiserweb.org**, 2003-2005). The aforementioned SOEIS project is a predecessor of the EICSTES project. Other projects, including SIBIS (Statistical Indicators Benchmarking the Information Society, **www.sibis-eu.org**, 2001-2003) in the European Union and OCLC (Online Computer Library Center, **www.oclc.org**, 1967-) in the USA, have also produced web measurements.

The SIBIS project, which focused on the link between the internet and scientific communication, designed survey-based approaches to cover the internet behaviour of researchers. These approaches concentrated on issues such as the frequency in the use of on-line and off-line information sources, the presence of websites and e-mail usage (Barjak and Harabi 2003). The EICSTES project started with a study of possible internet and web indicators (Heimeriks 2002). It undertook large-scale data collection and developed methods and tools for a variety of data such as link data, search-behaviour data and data about retrieval behaviour from scientific databases. It developed web indicators comparing bibliometric and web data (Heimeriks, Hörlesberger et al. 2003; Heimeriks and Besselaar 2006), including analytical tools (such as tools for creating Self Organized Maps) and visualization tools. It validated different approaches to using web data for indicator development in a variety of qualitative and quantitative case studies (focused on the subfield level or specialties) within the scientific system and at the interface with society (intermediaries).

As compared with EICSTES, the WISER project has placed the focus on the theoretical and methodological foundation of web indicators. The constituent case studies of the WISER project focus on methodology and included, for example, a study on visibility on the web and an examination of the issues pertinent to web-page persistence. The application of different crawlers (commercial and non-commercial tools) in WISER led to a systematic investigation of web crawls (Cothey 2004). These methodological undertakings were combined with the analysis of large data sets of web data at the level of research groups, departments, universities and countries.

In this section of the paper we will focus primarily on agenda-setting in the WISER project and its findings. The WISER project developed a specific research agenda incorporating elements of an evolutionary search strategy for the detection of the targets and tools of measurement. The project started by exploring a wide range area of tailor-made tools (crawlers) and commercial software toolkits for extracting information about websites and linking structures. It also produced an overview of measurable items on the web, such as characteristics of web pages and of the web resources that constitute web pages [8] (e.g. images, pdfs, java applets), the so-called invisible web, traffic data, e-mail lists, forums and chat rooms. The rate of mutation and variance in the search was high in this exploratory phase of the project. In the course of the project we introduced different selection criteria for the purpose of focusing our research. One selection criterion was the availability of measurement tools. Finally, WISER concentrated on crawler development and search engines. Two crawlers were developed (Cothey 2004; Thelwall 2004a) and search engines were explored systematically. The second selection criterion concerned measurable items. WISER excluded traffic data, the analysis of e-mail lists and chat rooms and concentrated on information available on websites. More specifically, WISER concentrated on university websites. This is unique subset of the STI Web domain. In the final phase of the project, science policy requirements are being converted into an additional selection criterion so as to make it possible to concentrate on a specific set of web indicators and visualization tools suitable for science policy issues.

WISER aims to cover features of the science system in the turn to e-science. The overall objective of the WISER project is to start developing a new generation of web-based S&T indicators by exploring the possibilities and

problems of web-based S&T indicators. The project also has several specific key objectives:

1. To lay the foundations of web-indicator research through the investigation of data collection techniques and the production of a survey of practices manual.
2. To develop a set of indicators covering key aspects of e-science, i.e. primarily those represented on the web.
3. To apply these indicators in a series of experimental case studies which, it is hoped, will blaze a trail for the future use of such information in a science policy context.

The methodological approach of WISER

The WISER project is concentrating on the study and development of the foundations of web-based STI indicators. Therefore, it is attempting to take a systematic approach to measurement problems. It combines experience in the measurement of the web in general (i.e. not only the scientific world) with the specific needs for the monitoring of innovation activities. Thus, a set of case studies was set up for the purpose of identifying the development of e-science qualitatively and quantitatively. Links to other areas of internet studies (e.g. ethnographic and sociological studies) were made in the case studies. The "webometric approach" can be seen as a key element in the project. However, the project covers a broader area while at the same time being more focused than webometrics. On the one hand, it concerns measurable units and quantifiable phenomena in knowledge production and in this differs from "pure" output-oriented communication studies: traditional science and technology indicators concerning innovation infrastructure have been taken into account here. On the other hand, it is more restricted than "webometrics" because the focus is on information, maps and indicators relevant to science policy. Figure 2 shows how the various research areas of web measurement and internet studies serve as sources for the WISER approach. WISER uses techniques from the field of internet studies such as the manual inspection of web pages and interviews to investigate how social norms, behaviour and community formation are shaped by the use of the internet and the web (see **www.aoir.org**). WISER also relies on webometric studies of hyperlink networks and web performance and compares the search for indicators with traditional bibliometric and scientometric approaches. WISER adopted concepts of the structure, dynamics and stability of networks from complex networks for the discussion of static versus dynamic indicators.

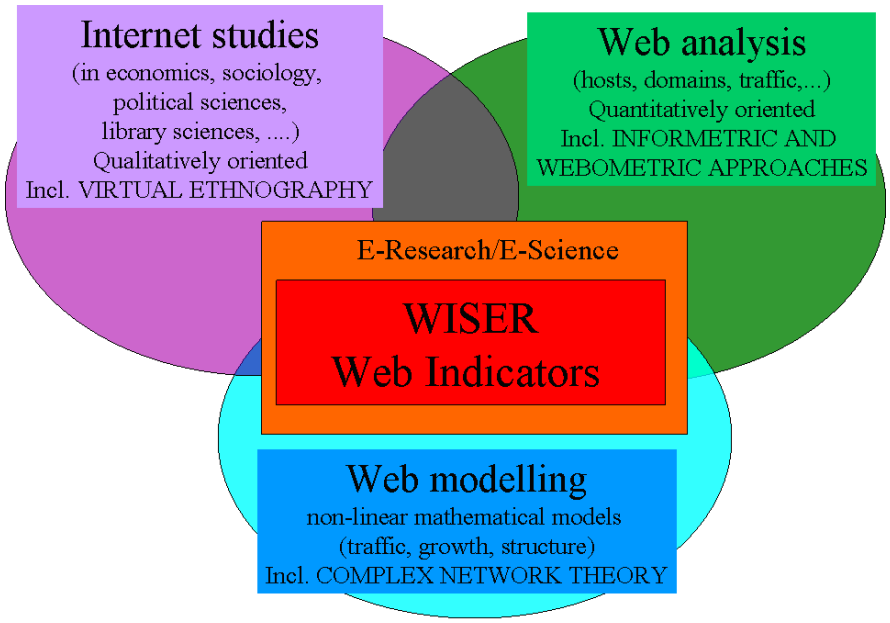

Figure 2: The location of WISER with respect to other internet studies

One stream of the WISER project involved the production of data sets of web data that are open to the public which enable comparative analysis. An example of these are the co-link data files produced by Sylvan Katz (**http://www.wiserweb.org/Reports/co-link/co-link.html**). The SCIT

group at the University of Wolverhampton started an “Academic Web Link Database Project” in which raw data and software were presented for the purpose of analysing link structures (**http://cybermetrics.wlv.ac.uk/database/index.html**). This group also published a crawler called SocSciBot (**http://socscibot.wlv.ac.uk/**). One target of the WISER project is the archiving of raw data for public purposes and the development of standards and best practice for crawling (Cothey 2004). This includes the comparison of different crawlers (i.e. commercial and tailor-made).

Self-organization and complexity – remaining stable while moving

The worldwide web can be analysed both as a tool for research and other activities and as a space in which different activities are mediated (Wouters et al. 2007). For web indicators, the web is used to trace academic activities. Therefore, it is important to keep in mind that the web emerged due to a vast number of individual acts of web page creation. Although these individual activities may be shaped by institutions and regulations, the web as such is a phenomenon that emerges spontaneously and on a self-organized basis. Despite this apparently uncoordinated and chaotic behaviour at the micro-level of web creation, the web displays remarkably regular structures on a macro level. These structures entail power laws in the degree distribution of hyperlinks (both concerning in-links and out-links), in the correlation between the number of nodes and the number of links, in the size of web pages and the emergence of community-like structures in hyperlink networks (Pastor-Satorras and Vespignani 2004). As already mentioned, such phenomena have attracted the interest of branches of complexity theory and, in particular, complex network theory which searches for “laws of the web” (Huberman 2001). Indeed, the availability of web data for automatic data mining and analysis has prompted the emergence of an entire branch of complex network theory (Scharnhorst 2003). Before we discuss the implications of seeing the web as a self-organized system for the construction of web indicators, we will address another feature that is related to the self-organized nature of the web, i.e. the non-stationary character of the web. Processes in which the values of certain variables or their probability distribution do not change over time are stationary. Thus, by non-stationary, we mean that all characteristics of the web (its size, composition and content) are in a constant state of change. The question arises as to whether this also leads to structures (i.e. order out of chaos). The dynamic nature of the web entails not only growth processes, but also changes related to the structure and content of the network itself. This has often been stated as a specific characteristic of new media. Search engines, for example, are continuously overwriting their own history (Wouters et al. 2004, Hellsten et al. 2006). They are “tied to updating cycles of the Web and the Internet, rather than to the historical development of their structure. The construction of tailor–made archiving and crawling tools seems, therefore, urgent if one wishes to retain either structural or historical information on the Internet” (Wouters et al. 2004, Hellsten et al. 2006). Initiatives like the internet archive and the wayback machine (**http://www.archive.org/**), webarchivist (**http://www.webarchivist.org/about.htm**), as well as websites referring to “old” browser technologies (**http://browsers.evolt.org/**) indicate the dynamic character of the web and the attempt to create more stable web archives. Other fields in which issues of reliability of web sources have been extensively discussed can be found in library sciences and archive sciences (Fetterly, Manasse et al. 2004). The most prominent example is Koehler’s analysis of web page persistence (Koehler 1999; Koehler 2002; Koehler 2004). Koehler addresses different kinds of stability and persistence. These features can be embedded in a broader picture of the changeable web (see Figure 3).

The stability of the web can be problematized from different perspectives. The nature of the nodes and the links in the web may differ, depending on the units of analysis. The definitions used by the W3 consortium refer to the web resource as an elementary unit. A web resource is identified by a URI (Uniform Resource Identifiers). A web page, which can contain elements such as pictures, games and Word documents, is, therefore, itself a network. If we conceptualize the web as a network or graph, temporal stability or instability can be linked to all components of the graph (the links as well as the nodes).

Web indicators must deal with both the “technical and content identity” of web

pages. Koehler proposed that a differentiation be made between constancy and permanence: “Constancy measures the rate at which web documents are changed in any way over time” and “permanence measures the probability that web documents will carry the same URL over time” (Koehler 2002). “Permanence” and “constancy” address issues of URL and content stability of nodes, whereas “intermittency” addresses the stability of links. These features are part of an interplay of the stability and instability of all elements of a web graph (Figure 3). By differentiating between the content of a node and its URL address, we create two different images of the web graph: one based on an “identity” of a web page and the other based on its “address”. (Figure 3) The life cycle of a web page could be correlated to both its content and its location in the lexically defined web graph. Furthermore, we can differentiate between documents that bear the same identity but a different URL, or the other way around. An example of the first is an on-line document in a digital repository (a journal publication) which may have different URLs over time because of changes in the technical structure of the repository. If the content of a web page is included in its “life time”, almost 80% of so-called dead or broken links can be re-found using a strategy of combining the URL structure with searching for keywords in search engines (**http://www.earlham.edu/~peters/wirewise/no2.htm**). In these cases, content is simply replaced in the graph (technically). However, the opposite case also arises if a URL is stable, but the content changes frequently. Changes in both the content and the URL may affect the link structure of the web graph. In visualizations of a web graph, both as a URL structure and as a content based graph (Figure 3), nodes and links can be the subject of re-production (growth) and decline.

In our diagram, it is possible to differentiate between the different processes of change. These processes influence web indicators differently according to the constituent elements of the graph that make up the web indicators. Web indicators, such as web impact factor, are influenced by changes concerning both the URLs of the nodes and the in-link structure to them.

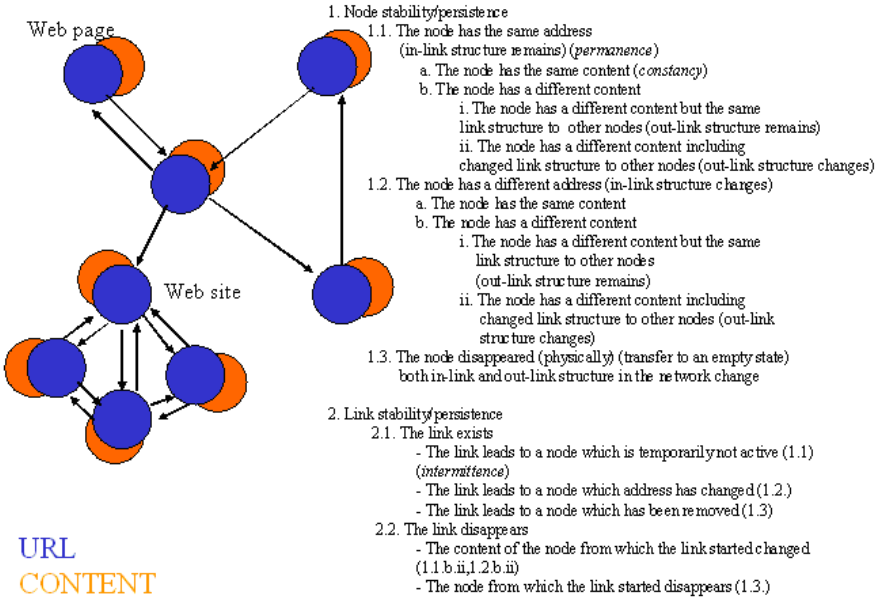

Figure 3: Scheme for the classification of temporal changes in a web graph

Temporal changes are only one aspect of stability. Another aspect is related to the measurement possibilities. We start again from the web as a graph. The web graph is specific because its links are directed. The web graph retrieved will look different, depending on the starting points for a crawl and the parameter of a crawl (Cothey 2004) (Figure 4). Indicators are usually assumed to be based on reliable, reproducible and standardized data. In the case of the internet or, more specifically, the web, indicator research is confronted with changing representations of the communications of different communities (Leydesdorff, Scharnhorst 2002).

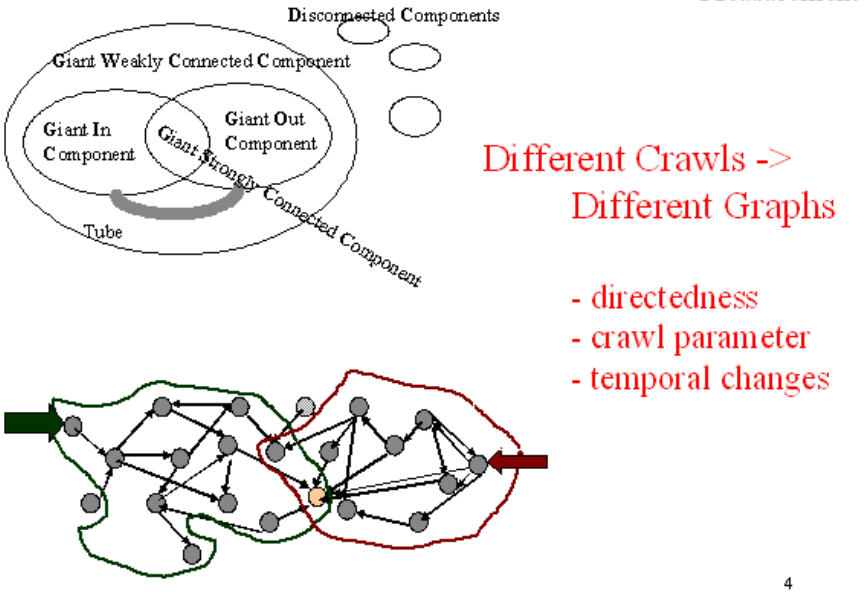

Figure 4: The directed web graph (The image in the top left corner represents Broder et al.'s "bow-tie" model (Broder, Kumar et al. 2000) (see also Björneborn 2004)

There are several possible answers to this problem. Firstly, web indicators need to be reflexive in a way whereby their interpretation pays specific attention to the circumstances of the data collection, on which they are based (Leydesdorff and Scharnhorst 2002). In other words, their presentation and interpretation must be sensitive to the method of data collection and data analysis (Thelwall 2004a). The publication of both raw data and data analysis is one possibility here (see, for example, **http://www.wiserweb.org/Reports/co-link/co-link.html**). Due to the instable nature of web data, however, this publication of web data can not be intended as a contribution to a reference database. Instead of attempting to create web data repositories as standards of reference, libraries of tools (as crawlers) with a test data set and criteria for evaluation of the tools could/should be created (WISER 2003, internal communication). However, it is important to note that the study of the dynamic web also has disadvantages. For example, the analysis of data cannot be repeated. The crawling of large amounts of data has another disadvantage. The graph will change during the crawl time. Thus, instead of having a snapshot of a web graph at one point in time, one obtains a time series of snapshots of the graph. Having said that, it would also be possible to aim to publish crawl results to make the raw data available for further analysis. (Katz, 2005, personal communication)

Another way of taking the dynamic nature of the web into account is to consider how web indicators are constructed. Web indicators are "snapshots" of dynamic representations of formal and informal communication around scholarly activity. Thus, we must differentiate between the dynamics of the representations and the dynamics of the (multi-)media in which the representations manifest themselves. It is also possible to examine similar measurement problems in dynamic systems in natural sciences. One way to deal with noise in complex systems is to analyse large ensembles, in which random fluctuations are small compared with trend variables. Thus, there is some hope that when large web graphs are crawled, the statistical features of these graphs remain stable, even if some parts of the graph change (Cothey 2004, personal communication). This means, for example, that we can expect that the exponent of the power law of an indegree-distribution for a fixed set of web pages or websites is stable over time. It was, at least, the claim of complex network theory that the values of such exponents have a certain universality and point to unique features in the creation of the graph. Thus, if two crawls produce different exponents, it is important to ask whether this is not due to the existence of different systems and different mechanisms and not to the fact that it is just another representation. Cothey et al. provide another example on this subject in their paper which compares the size and the impact of web sites defined using different algorithms (Cothey, Aguillo et al. 2006).

The statistical analysis of web graphs relies on the characteristics of distributions (Scharnhorst 2003). Examples include the in-degree distribution, the out-degree distribution, clustering coefficients and shortest path lengths. Most of the characteristics in web graphs, such as the number of pages in a site, the number of in-links to pages or the size of clusters, are seen to be skewed distributions. Therefore, the comparison of averages relating to quantities or absolute quantities does not appear to provide any relevant information. However, one can compare the distributions and in this way create benchmark indicators (Thelwall 2004b; Katz 2006). For example, the number of in-links to different web pages has a skewed distribution: many web pages have few in-links and only a few web pages have large numbers of in-links. This skewed distribution could vary for different samples of web sites. The question remains as to whether an optimal distribution exists or whether different distributions express different functions of the underlying networks. Another way to obtain benchmark indicators is to compare the observed number of in-links with an expected number of in-links. In this case the expectation comes from a skewed distribution, from which the measured points may deviate.

Conclusions

Web-based STI indicators can be understood as specific representations of innovation activities that are dynamic in nature. Web-graph measurements merely provide snapshots. The challenge is to create robust measurement methodologies that allow for the construction of stationary stable re-presentations of the web graph. For the development of web indicators, one would either rely on relatively stable structural features of the web graph (as degree distributions) or use time series of network characteristics (as development of clustering coefficients over time). It is then necessary to be aware that changes in these time series are a product of both stochastic elements and deterministic trends.

Studies carried out as part of the WISER project have shown that web indicators cannot be retrieved through purely automated crawling; a combination of quantitative and qualitative studies is required to:

- determine the unit of analysis (which technical representation may change);
- determine the boundaries of the measurable due to a certain technical approach;
- determine the meaning of the measured.

There is a tension between qualitative and quantitative approaches because they usually cover different concepts of samples and different sample sizes. Qualitative investigations involving manual elements of annotating web pages or sites can provide complements to large automated web-data gathering. If a reasonable sampling strategy is developed, it may be argued that more general conclusions can be drawn.

Bibliometric indicators rely on the “literature model” of science. Analogously, web indicators require a “model” of the purpose and function of the web activity of scholars. There is no consensus about such a model. Similar to citation indicators, web indicators can be seen as experiments with additional data sources. It is our belief that the combination of semiotics and the sociology of scientific knowledge can make an important contribution to this analysis and reflection.

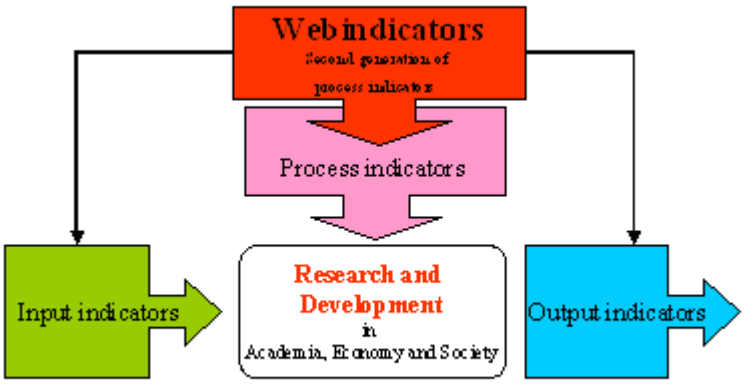

Figure 5: Web indicators located in the landscape of science, technology and innovation indicators.

Our experience and the research questions to be answered in the future may be summarized as follows:

1. The ambiguity of representation influences the measurement methodology and the interpretation of the results. The temporal aspect is important.

2. What kind of structures emerge in the web graphs we sample? How do they differ from those expected? Can we transform mathematical variables characterizing topological features of networks into indicators? Will they be different across different samples (clustering coefficient, power law exponents)? To which differentiation would they point?

3. If the structures we see are the result of a process of self-organization, what kind of self-organization takes place? What are the collective moments? To give one example: each act of hyperlinking may be unique and carry its own meaning, but can we see a pattern in the frequency of actions/motivations/decisions, i.e. criteria on the basis of which linking takes place, can we group link activities into different categories? Is it possible to distinguish systematically between sites one links to? What is the frequency with which large hubs occur and develop?

In terms of finding answers to these questions, the analysis of the dynamics of the system may well prove to be an important research element.

Acknowledgement

We are grateful to the members of the WISER consortium (**www.wiserweb.org**) for their stimulating contributions and suggestions. We would like to thank Sylvan Katz in particular for his thorough and detailed comments. This study was supported by a grant from the “Common Basis for Science, Technology and Innovation Indicators” part of the Improving Human Research Potential programme of the European Commission’s Fifth Framework for Research and Technological Development. It is also part of the WISER project (Web indicators for scientific, technological and innovation research) (Contract HPV2-CT-2002-00015).

Notes

1. WISER Web indicators for scientific, technological and innovation research. Members of the consortium: Isidro Aguillo, Viv Cothey, Sylvan Katz, Hildrun Kretschmer, Colin Reddy, Andrea Scharnhorst, Mike Thelwall, Paul Wouters. For futher information see: **www.wiserweb.org**

2. Virtual scientific communities or virtual scientific laboratories are also called “collaboratories”. William Wulf … coined the term “collaboratory” to describe the concept of using information technologies to make geographically separate research units function into a single laboratory in 1989. Wulf defined a “collaboratory” as a “…‘center without walls’ in which the nation’s

researchers can perform their research without regard to geographical location—interacting with colleagues, accessing instrumentation, sharing data and computational resources, and accessing information in digital libraries" *Science and Engineering Indicators 2000*, chapter 9)

3. Descriptive indicators are sometimes also used (**http://www.nsf.gov/sbe/srs/nsf01336/p2s2.htm#rd**)

4. See **http://www.wtcw.nl/nl/projecten/eScience.pdf**.

5. See also Lawrence, S. and C. L. Giles (1999). "Accessibility of information on the web." **Nature** 400: 107-109.

6. Scale-independence or scale-invariant structures have also been discussed as features of self-organizing systems. They have been analysed in the context of fractal structures (Mandelbrot, B. B. (1983). The fractal geometry of nature. New York, W. H. Freeman.) or self-organized criticality (Bak, P. (1996) How Nature Works: The Science of Self-Organised Criticality, New York, NY: Copernicus Press)

7. EU TSER Project PL97-1296 1997-99

8. From a technical point of view, the worldwide web consists of "all the resources and users on the Internet that are using the Hypertext Transfer Protocol" (**http://searchcrm.techtarget.com/sDefinition/0,,sid11_gci213391,00.html**). "HTTP (Hypertext Transfer Protocol) is the set of rules for transferring files (text, graphic images, sound, video, and other multimedia files) on the World Wide Web" (http://whatis.techtarget.com). The files can be labelled as web resources. Examples of web resources include electronic documents and imagesas well as services, such as information provided from a databasea, e-mail messages and Java classes.

References

(2003) **European Report on Science & Technology Indicators** : Towards a Knowledge-Based Economy 2003 **<http://cordis.europa.eu/indicators/third_report.htm>** (accessed 31 July 2006).

(2001) **Future research domains at the frontiers of science and technology: on the road to the 6th FP**, FP6 Strategic Planning Workshop, Brussels, 26-27 April 2001, REPORT on discussions held in PANEL 4: COMPUTING, **<http://cordis.europa.eu/ist/fet/6fp-7.htm>** (last modified 21 December 2004, accessed 29 July 2006).

(2004) **Key Figures : Towards a European Research Area ; Science, Technology and Innovation** 2003-2004 **<ftp://ftp.cordis.europa.eu/pub/indicators/docs/ind_kf0304.pdf>** (accessed 31 July 2006).

(2005) **OECD Factbook** : economic, environmental and social statistics / Organisation for Economic Co-operation and Development 2005.

(2000) **Science and Engineering Indicators** 2000, **<http://www.nsf.gov/statistics/seind00/>**.

(2004) **Science and Engineering Indicators** 2004, **<http://www.nsf.gov/statistics/seind04/>**.

(1998) **Science, Technology and Industry Outlook** / Organisation for Economic Co-operation and Development 1998, Chapter 7: The Global Research Village: How Information and Communication Technologies affect the Science System, p. 189 – 238. **<http://www.oecd.org/dataoecd/9/30/2754574.pdf>** (accessed 31 July 2006).

Adamic, L. A. (1999), The Small World Web, in: S. Abiteboul, A.-M.

Vercoustre (eds.), **Research and Advanced Technology for Digital Libraries: Third European Conference, ECDL'99, Paris, France, September 1999. Proceedings** LNCS 1696, Springer, Berlin/Heidelberg, 443-452.

Adamic, L. (2000), Zipf, Power-laws, and Pareto - a ranking tutorial. < **http://www.hpl.hp.com/research/idl/papers/ranking/** > (last modified 10 April 2000, accessed 29 July 2006). “

Almind, T. and P. Ingwersen (1997), Informetric Analyses on the World Wide Web: Methodological Approaches to "Webometrics", **Journal of Documentation** 53: 404-426.

Arzberger, P., P. Schröder, A. Beaulieu, G. Bowker, K. Casey, L. Laaksonen, D. Moorman, P. Uhlir, P. Wouters (2004), Science and Government: An International Framework to Promote Access to Data. **Science** 303(5665): 1777-1778.

Bak, P. (1996), **How Nature Works: The Science of Self-Organised Criticality**, Copernicus Press, New York.

Barabasi, A. L., H. Jeong, Z. Neda, E. Ravasz, A. Schubert, T. Vicsek (2002), Evolution of the social network of scientific collaborations., **Physica A** 311(3-4): 590-614.

Bar-Ilan, J. and B. C. Peritz (2002), Informetric Theories and Methods for Exploring the Internet: An Analytical Survey of Recent Research Literature, **Library Trends** 50(3): 371-392.

Barjak, F. and N. Harabi (2003), Internet for R&D. Final Report of the SIBIS Project. < **http://www.fhso.ch/pdf/unternehmer/sibis_final.pdf** > (access 31 July 2006).

Bentley, R. A. and H. D. G. Maschner (2000), A Growing Network of Ideas, **Fractals-Complex Geometry Patterns and Scaling in Nature and Society** 8(3): 227-237.

Björneborn, L. (2004), **Small-World Link Structures Across an Academic Web Space: a Library and Information Science Approach**, PhD Thesis, Department of Information Studies, Copenhagen, Royal School of Library and Information Science, **<http://www.db.dk/lb/#phd>** (access 31 July 2006).

Björneborn, L. and P. Ingwersen (2001), Perspectives of Webometrics, **Scientometrics** 50(1): 65-82.

Bollen, K. A. (2001), Indicator: Methodology, in: N.J. Smelser and P.B. Baltes (eds.), **International Encyclopedia of the Social and Behavioral Sciences**, Elsevier Academic Press, Amsterdam, 7282-7287.

Bornholdt, S. and H.-G. Schuster (2002), **Handbook of Graphs and Networks: From the Genome to the Internet**, Wiley-VCH, Weinheim.

Boudourides, M. A., B. Sigrist, et al. (1999). Webometrics and the Self-Organization of the European Information Society < **http://hyperion.math.upatras.gr/webometrics/** > (last accessed 29 July 2006)

Bray, T. (1996), Measuring the Web, **Computer Networks and ISDN Systems**, 28(7-11): 993 – 1005.

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

Brunn, S. D. and M. Dodge (2001). "Mapping the ‘Worlds’ of the World Wide Web: (Re)Structuring Global Commerce Through Hyperlinks." **American Behavioral Scientist** 44(10): 1717-1739.

Callaway, D. S., M. E. J. Newman, S.H. Strogatz, D.J. Watts (2000), Network Robustness and Fragility: Percolation on Random Graphs, **Physical Review Letters** 85(25): 5468-5471.

Cothey, V. (2004), Web-Crawling Reliability, **Journal of the American Society for Information Science and Technology** 55(14): 1228-1238.

Cothey, V., I. Aguillo, N. Arroyo (2006), Operationalising "Websites": lexically, semantically or topologically?, **Cybermetrics** 10(1) **<http://www.cindoc.csic.es/cybermetrics/articles/v10i1p4.pdf>**

Cozzens, S. E., P. Healey, A. Rip, J. Ziman (1990), **The Research System in Transition**, Kluwer, Dordrecht.

Cronin, B. (2001), Bibliometrics and Beyond: Some Thoughts on Web-Based Citation Analysis, **Journal of Information Science** 27(1): 1-7.

Day, M. (2005), Institutional Repositories and Research Assessment, **<http://www.rdn.ac.uk/projects/eprints-uk/docs/studies/rae/rae-study.pdf**> (access 31 July 2006).

Dobrov, G. M. (1970), **Aktuelle Probleme der Wissenschaftswissenschaft** (in German). Akademie Verlag, Berlin.

Eco, U. (1976) **A Theory of Semiotics**, Indiana University Press, Bloomington.

Egghe, L. (2000), New Informetric Aspects of the Internet: Some Reflections - Many Problems, **Journal of Information Science** 26(5): 329-335.

Faloutsos, M., P. Faloutsos, C. Faloutsos (1999), On Power-Law Relationships of the Internet Topology, **Comput. Commun. Rev.** 29: 251.

Fetterly, D., M. Manasse, M. Najork, J. Wiener (2004), A large-scale study of the evolution of Web pages, **Software Practise and Experience** 34(2): 213-237.

Godin, B. (2004), The Who, What, Why and How of S&T Measurement, Project on the History and Sociology of S&T Statistics, Working Paper no. 26, 16 page, **<http://www.csiic.ca/PDF/Godin_26_a.pdf>** (Published also under the title: Pour une sociologie de la statistique sur la science et l'innovation, **Le Banquet: revue politique** 19(1): 159-170).

Goldstein, M. L., S. A. Morris, G.G. Yen (2004), A Group-Based Yule Model for Bipartite Author-Paper Networks, **<http://arxiv.org/abs/cond-mat/0409205>** (access 31 July 2006)

Goodrum, A. A., K. W. McCain, S. Lawrence, C.L. Giles (2001), Scholarly Publishing in the Internet Age: a Citation Analysis of Computer Science Literature, **Information Processing & Management** 37(5): 661-676.

Grupp, H. and U. Schmoch (1999), Patent Statistics in the Age of Globalisation: New Legal Procedures, New Analytical Methods, New Economic Interpretation, **Research Policy** 28: 377-396.

Harnad, S. (2001), The Self-Archiving Initiative, **Nature** 410(26 April): 1024-1025.

Harter, S. P. (1997), Scholarly Communication and the Digital Library: Problems and Issues, **Journal of Digital Information** 1(1): **<http://jodi.ecs.soton.ac.uk/Articles/v01/i01/Harter/>**.

Heimeriks, G. (2002), Development of Web-Indicators. Report. **<http://www.eicstes.org/EICSTES_PDF/Deliverables/Development%20of%20Web%20Indicators.PDF>**

Heimeriks, G. and P. van den Besselaar (2006), Analyzing Hyperlinks

Networks. The Meaning of Hyperlink Based Indicators of Knowledge Production, **Cybermetrics** 10(1), **<http://www.cindoc.csic.es/cybermetrics/articles/v10i1p1.pdf>**

Heimeriks, G., (2005), **Knowledge Production and Communication in the Information Society: Mapping Communications in Heterogeneous Research Networks**, PhD Thesis, University of Amsterdam.

Heimeriks, G., M. Hörlesberger, P. van den Besselaar (2003), Mapping Communication and Collaboration in Heterogeneous Research Networks, **Scientometrics** 58(2): 391-413.

Hellsten, I., L. Leydesdorff,, and P. Wouters (2006), Multiple Presents: How Search Engines Re-write the Past, **New Media and Society** 8(6) (in print)

Hey, T. and A. E. Trefethen (2002), The UK e-Science Core Programme and the Grid, **Future Generation Computer Systems** 18(8): 1017-1031.

Huberman, B. A. (2001), **The Laws of the Web: Patterns in the Ecology of Information**.The MIT Press, Cambridge Mass...

Ingwersen, P. (1998), The Calculation of Web Impact Factors, **Journal of Documentation** 54(2): 236-243.

Ingwersen, P. and L. Björneborn (2004), Methodological Issues of Webometric Studies, in: H.F. **Moed**, W. **Glänzel**, U. **Schmoch** (eds.), **Handbook of Quantitative Science and Technology Research : The Use of Publication and Patent Statistics in Studies of S&T Systems**. Kluwer Academic Publishers, Dordrecht/Boston/London, 339-370.

Katz, J. S. (1999), The Self-Similar Science System, **Research Policy** 28: 501-517.

Katz, J. S. (2000), Scale Independent Indicators and Research Evaluation, **Science and Public Policy** 27(1): 23-36 .

Katz, J. S. (2004), Co-link indicators of the European Research Area, WISER report, < **http://www.sussex.ac.uk/Users/sylvank/pubs/Co-Link.pdf**> (access 31 July 2006).

Katz, J. S. (2005). "Scale independent bibliometric indicators." **Measurement** 3(1): 24-28.

Katz, J. S. (2006), Indicators for complex innovation systems, **Research Policy (in print)** **<http://www.sussex.ac.uk/Users/sylvank/pubs/ICIS-RP.pdf**>.

Kleinberg, J. M., R. Kumar, P. Raghavan, S. Rajagopalan, A. S. Tomkins (1999), The Web as a Graph: Measurements, Models, and Methods, in: T. Asano, H. Imai, D.T. Lee, S. Nakano, S. Tokuyama (eds.), **Computing and Combinatorics: 5th Annual International Conference, COCOON'99, Tokyo, Japan, July 1999. Proceedings**. Lecture Notes in Computer Science, Vol. 1627, Springer, Berlin, 1-17.

Kleinknecht, A., K. v. Montford, E. Brouwer. (2002), The Non-Trivial Choice Between Innovation Indicators, **Economics of Innovation and New Technology** 11(2): 109-121.

Kling, R. (2000), Learning about Information Technologies and Social Change: The Contribution of Social Informatics, **The Information Society** 16(3): 217-232.

Koehler, W. (1999), Digital Libraries and World Wide Web sites and Page Persistence, Information Research 4(4), **<http://InformationR.net/ir/4-4/paper60.html>** .

Koehler, W. (2002), Web Page Change and Persistence - A Four-Year Longitudinal Study, **Journal of the American Society for Information**

Science and Technology 53(2): 162-171.

Koehler, W. (2004), A Longitudinal Study of Web pages Continued: a Consideration of Document Persistence, **Information Research** 9(2), **<http://informationr.net/ir/9-2/paper174.html>**.

Larson, R. R. (1996), Bibliometrics of the World Wide Web: an Exploratory Analysis of the Intellectual Structure of Cyberspace, in: S. Hardin (ed.), **Global Complexity: Information Chaos and Control**, Proceedings of the ASIS&T 59th annual meeting, Available at <**http://sherlock.berkeley.edu/asis96/asis96.html**>

Lawrence, S. and C. L. Giles (1999), Accessibility of Information on the Web, **Nature** 400(6740): 107-109.

Leland, W. E., M. S. Taqqu, W. Willinger, D.V. Wilson (1994), On the Self-Similar Nature of Ethernet Traffic, **IEEE ACM transactions on networking** 2 (1):1-15.

Leydesdorff, L. and A. Scharnhorst (2002), Measuring the Knowledge Base: A Program of Innovation Studies, Report to the Bundesministerium für Bildung und Forschung. Berlin-Brandenburgische Akademie der Wissenschaften. Amsterdam, NIWI-KNAW, <**http://www.sciencepolicystudies.de/dok/expertise-leydesdorff-scharnhorst.pdf**> (access 31 July 2006)

Li, M., J. Wu, D. Wang, T. Zhou, Z. Di, Y. Fan (2005), Evolving Model of Weighted Networks Inspired by Scientific Collaboration Networks, <**http://arxiv.org/abs/cond-mat/0501655**>.

Liang, L. (2005), R-Sequences: Relative Indicators for the Rhythm of Science, **Journal of the American Society for Information Science and Technology** 56(10):1045-1049.

Mandelbrot, B. B. (1983), **The Fractal Geometry of Nature**, W. H. Freeman, New York.

Marsh, J. B. T. (2001), Cultural Diversity and the Information Society – Policy Options and Technological Issues, Final study, <**http://www.euresearch.ch/media/IST_STOA_Cult_Divers_en_2000.pdf**> (access 31 July 2006)

Maurer, S. M. and B. A. Huberman (2001), Restart Strategies and Internet Congestions, **Journal of Economic Dynamics & Control** 25: 641-654.

Meyer, M. (2000a), Does Science Push Technology? Patents Citing Scientific Literature, **Research Policy** 29: 409-434.

Meyer, M. (2000b), What is Special about Patent Citations? Differences Between Scientific and Patent Citations, **Scientometrics** 49: 93-123.

Moed, H. F., W. Glänzel, U. Schmoch (eds.) (2004), **Handbook of Quantitative Science and Technology Research : The Use of Publication and Patent Statistics in Studies of S&T Systems**. Kluwer Academic Publishers, Dordrecht/Boston/London.

Moed, H. F. and T. Van Leeuwen (1995), Improving the Accuracy of Institute for Scientific Information's Journal Impact Factors, **Journal of the American Society for Information Science** 46(6): 461-467.

Nalimov, V. V. and B. M. Mulchenko (1969), **Naukometria** (in Russian). Nauka, Moscow.

Nentwich, M. (2003), **Cyberscience. Research in the Age of the Internet**, Austrian Academy of Sciences Press, Vienna.

Pastor-Satorras, R. and A. Vespignani (2004), **Evolution and Structure on the Internet. A Statistical Physics Approach**.Cambridge University Press,

Cambridge.

Rousseau, R. (1997), Sitations: an Exploratory Study, **Cybermetrics** 1(1), <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(1), <http://www.cindoc.csic.es/cybermetrics/articles/v2i1p2.html>.

Scharnhorst, A. (2003), Complex Networks and the Web: Insights from Nonlinear Physics, **Journal of Computer-Mediated Communication** 8(4) <http://jcmc.indiana.edu/vol8/issue4/scharnhorst.html>.

Schmoch, U. (1997), Indicators and the Relations between Science and Technology, **Scientometrics** 38(1): 103-116.

Schubert, A. and W. Glänzel (1983), Statistical Reliability of Comparisions Based on the Citation Impact of Scientific Publications, **Scientometrics** 5: 59-74.

Sen, P. (2004), Directed Accelerated Growth: Application in Citation Network, <http://arxiv.org/abs/cond-mat/0409154>.

Snyder, H. and H. Rosenbaum (1999), Can Search Engines be Used for Web-Link Analysis? A Critical Review, **Journal of Documentation** 55(4):375-384.

Tadic, B. (2001), Dynamics of directed graphs, **Physica A** 293(1-2): 273-284.

Thelwall, M. (2004a), **Link Analysis: an Information Science Approach**. Elsevier Academic Press., Amsterdam et al..

Thelwall, M. (2004b), Weak Benchmarking Indicators for Formative and Semi-Evaluative Assessment of Research, **Research Evaluation** 13(1): 63-68.

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

Thomas, N., A. King, E.H.G. Jones (2000), Linguistic Diversity on the Internet: Assessment of the Contribution of Machine Translation, STOA Study PE 289 622/Fin. ST., European Parliament, <http://www.serv-inf.deusto.es/ABAITUA/konzeptu/ta/EuroParlament.html> (access 31 July 2006).

Thomas, O. and P. Willet (2000), Webometric Analysis of Departments of Librarianship and Information Science, **Journal of Information Science** 26 (6): 421-428.

Vaughan, L. and D. Shaw (2005), Web Citation Data for Impact Assessment: A Comparison of Four Science Disciplines, **Journal of the American Society for Information Science and Technology** 56(10): 1075-1087.

Weingart, P. and M. Winterhager (1984). **Die Vermessung der Forschung** (in German), Campus Verlag, Frankfurt am Main.

Wilde, E. (1999), **World Wide Web: Technische Grundlagen** (in German), Springer, Berlin.

Wouters, P. (1999), Beyond the Holy Grail: From Citation Theory to Indicator Theories, **Scientometrics** 44(3): 561-580.

Wouters, P. (2000), Garfield as Alchemist, in: B. Cronin and H. B. Atkins (eds.), **Web of Knowledge**, Information Today, Inc, Medford, New Jersey, 65-71.

Wouters, P. and R. De Vries (2004), Formally Citing the Web, **Journal of the American Society for Information Science and Technology** 55(14): 1250-1260.

Wouters, P., Hellsten I, and Leydesdorff, L (2004), Internet Time and the Reliability of Search Engines, **First Monday** 9(10), <http://www.firstmonday.org/issues/issue9_10/wouters/index.html>

Wouters, P., K. Vann, A. Scharnhorst, M. Ratto, J. Fry, A. Beaulieu, I. Hellsten (The Virtual Knowledge Studio) (2007), Messy Shapes of Knowledge - STS Explores Informatization, New Media and Academic Work, in: E. Hackett, O. Amsterdamska, M. Lynch, J. Wajcman (eds.), **New Handbook of Science, Technololgy and Society**, MIT Press (under review).

Received 15/Dec/2005
Accepted 31/Jan/2006

DISCUSSION

Hybrid systems - some comments on the article by Andrea Scharnhorst and Paul Wouters
Peter van den Besselaar