

# **EICSTES DELIVERABLE D1.4**

## **STATE OF THE ART PART B: WP8**

### **State of the Art in BIBLIOMETRICS and WEBOMETRICS**

**Gaston Heimeriks and Peter van den Besselaar**  
Universiteit van Amsterdam  
January 2002

#### Contents

##### Introduction

1. Historical Overview of BIBLIOMETRIC Analyses
  2. Print Based Indicators
    - 2.1 Collective production of scientific knowledge.
    - 2.3 The landscape of knowledge production: Disciplines and Specialties
    - 2.4 The research front: Research Topics
    - 2.5 Geographical distribution and Transnational research collaboration
  3. Changing knowledge production
  4. From Bibliometrics to Webometrics
    - 4.1 E-journals
    - 4.2 Intermediate communications
    - 4.3 Internet
    - 4.4 Electronic Indicators
      - 4.4.2 Coordination mechanisms in various disciplines
      - 4.4.3 The landscape of knowledge production: Disciplines and Specialties
      - 4.4.4 The research front: Research Topics
      - 4.4.5 Geographical distribution and Transnational research collaboration
  5. Methodological considerations
    - 5.1 E-journals and webometrics
- (Contribution of ARCS: Agelika Zartl, Edgar Schiebel)*

- 5.1.1 E-journals
- 5.1.2 Web publications
- 5.1.3 Webometric Studies
- 5.2 Using Artificial Neural Networks for Clustering and Mapping of Science  
*(contribution of Xavier Polanco, INIST)*
  - Artificial neurons and networks properties
  - Reasons for using ANNs in metrics studies of science
  - Using ANNs in Scientometrics
  - Using ANNs in the Cyberspace
  - Knowledge indicators
- 5.3 Scientometrics and Network analysis  
*(contribution of Xavier Polanco INIST)*
  - Galois lattice or conceptual clustering
  - Semantic networks
- 5.4 Social Network Analysis, Chaos Theory and Complex Networks  
*(contribution of Moses A. Boudourides University of Patras)*
  - Social Network Analysis
  - Chaos Theory and the Internet
  - Complex Networks on the Internet
- 6. Conclusions and Future Research
- 7. References

# INTRODUCTION

This report provides a theoretical framework for studying techno-scientific developments with quantitative tools, as well as an overview of bibliometric and webometric indicators.

Bibliometric indicators of scientific activity are now at the heart of the debate over linkages between advances in science and technology and economic and social progress. There is a growing awareness that the advantages of basing research, and subsequent political choices, on criteria that lend themselves for more quantitative evaluation (Okubo, 1997). Bibliometrics is a tool by which the state of science and technology can be observed through the traces of communication in the science-technology system, most notably the published papers in refereed journals. Originally, it is means for situating a country in relation to the world, and institution in relation to a country. The focus of quantitative analysis of scientific development shifted more and more to the mapping of scientific development in interaction with social, political and economical developments; the so-called second generation indicators. These scientometric indicators are equally suitable for 'macro' analyses and 'micro' studies. They constitute a way to asses the current state of science with respect to the past, which can shed a light on its structure and development. Webometrics –or cybermetrics- can be considered as bibliometric analysis of electronic communications.

In this study, we will focus on indicators for mapping the development of the science system in interaction with economic and political developments. The central idea of this paper is that the enterprise of science is best captured by the concept of an evolving communication system. Scientists can be considered the carrying nodes. Central to this model is the idea that science can be characterized by the distributed communication of 'papers', each proposing a new quantum of knowledge. Of course, many other relevant modes of communication can be identified, but the corpus of science is best defined in terms of papers since these are the formal communications to which reference takes place.

“Whatever scientists think or say individually, their discoveries cannot be regarded as belonging to scientific knowledge until they have been reported to the world and put on permanent record” (Ziman, 1984)

From this perspective, it follows that science can be considered a literary genre, strongly associated with the print medium. Knowledge is produced by recombinations and associations of existing papers. New knowledge claims are related to previous research by means of cited references. It follows that there are some advantages to looking at science in its textual form. Printed papers not only provide a wide accessibility and dissemination of knowledge, it has also more fundamental attributes that legitimate it to be the focus of analysis. It provides a more codified mode of communication than other media. This raises some interesting issues for the quantitative analysis of electronic communications. It is clear that we cannot apply the same bibliometric methods that have been developed for print media to electronic communications without carefully taking into account the differences and similarities between print and electronic media. Not only are we in need of new kinds of indicators, but also new theories and models that improve our understanding of how the information revolution in society at large, and science in particular initiates new information and communication patterns. This revolution finds its roots in the intimate interaction between information and communication technologies (ICT), and advanced scientific and technological research and innovation.

As mentioned, the key concept of this paper is the idea that the enterprise of science is best captured by the concept of an evolving communication system. Evolving communication systems are not given or fixed, they are the result of a continuing process during which more complex forms of organization emerge. This undirected evolution is resulting from ongoing interaction

between system and environment. This concept not only provides a valuable theoretical framework for scientometric analysis, it allows provides an opportunity to elaborate on the consequence of the informational turn in science. The organization of the communication patterns defines a domain of reflexive interactions in which it can act with relevance to the maintenance of itself (Maturana, 1980). Processes of feedback and of self-reference are vital in maintaining a social system. This insight allows us to articulate some of the operations that construct the boundaries of the science system like peer reviews (Wouters, 1999) and editorial boards of scientific journals (Fujigaki, 1998). As Wouters points out, scientific communication can also be communicated reflexively, by measuring scientometric attributes of the communications, which in turn can be communicated. In terms of the information cycle representation, the indicators appear in the form of a second cycle. This cycle processes information about the primary information cycle, in the form of meta-information (Wouters, 1999). Together, these cyclic communication patterns result in a system that can be described as a self-maintaining evolving communication system with well definable boundaries and well definable boundary constructing processes.

The science system is an example of a self-organizing system. Many investigators have observed social activities and structures, particularly in the science system, that are best described by a power law distribution. A power law is one of the common signatures of a non-linear dynamic process, i.e., a chaotic process, which is at a point of self-organized criticality (Bak, 1991) or residing on the boundary between order and disorder. Such a system is often called a self-organized system because it exhibits structure not merely in response to inputs from the outside but also, indeed primarily, in response to its own internal processes (Krugman, 1996).

This perspective has implications for the way the science system can be studied quantitatively by means of indicators. In this paper we will start with a historical overview of scientometric indicators. We will discuss a number of central questions related to the study of scientific communications in print and electronic media and how scientometric indicators can be used in answering those questions. A number of recent studies will be discussed in more detail followed by suggestions for further research.

# 1. Historical Overview of BIBLIOMETRIC Analyses

Okubo (1997) describes that in 1969, Pritchard coined a new term –‘bibliometrics’- for a type of study that had been in existence for half a century. The fact that Pritchard felt the need to redefine the scope of an area that hitherto covered for fifty years by the term ‘statistical bibliography’ (Hulme, 1923) demonstrated that a new field of quantitative analysis had emerged; the application of mathematical and statistical methods to books and other means of printed communication (Pritchard, 1969, pp 348-349). Bibliometrics has become the generic term for a whole range of specific measurements and indicators; its purpose is to measure the output of scientific and technological research through data derived not only from scientific literature but from patents as well. Bibliometric approaches, whereby science can be portrayed through the results obtained, are based on the notion that the essence of scientific research is the communication of new contributions to the body of knowledge in scientific literature. Patents indicate a transfer of knowledge to industrial innovation and a transformation into something of economic and social value; for this reason they constitute an indicator of the tangible benefits of an intellectual and economic investment. The idea that to publish their work is the paramount activity of scientists has long been contented by science analysts. According to Price, a scientist is “..any person who has published a scientific paper” (Price, 1969). His catchphrase “publish or perish” would suggest that publication of research findings is at the forefront of scientists’ activities.

According to Okubo (1997) and Wouters (1998) the 1970s brought a quantum leap in the number of bibliometric studies initiated by the coming into existence of a database of scientific publications, the Science Citation Index (SCI). Founded by Garfield in Philadelphia in 1963. Garfield’s initial idea was to give researchers a quick and effective way to find published articles in their field of research. But he soon extended his work to evaluation of the references he compiled. The first advantage is that the SCI covers all science fields. This is a necessity if one is looking at whole research systems. In addition, SCI coverage is unambiguous because every item from every journal is indexed. The second advantage is that all author addresses listed on the paper are included in the SCI. This is a necessity for studying institutional output, as collaboration is so extensive. The third advantage is that references are included in the SCI and only the SCI. Citation counts can be derived from these references and used as a partial indicator of the impact previous research has had on succeeding work.

A second important step in the development of scientometrics came in the 1980’s with the introduction of co-word analysis (Callon et al, 1983). Co-word analysis was developed by proponents of social constructivism and the technique is as theory-laden as any scientific tool. Co-word analysis implicitly depends upon a number of beliefs in the context of Actor Network Theory (ANT). First, that technology is heterogeneous, that there is no way to separate the scientific from the social and political. The scientist is an interest driven person who uses whatever resources are available, from social influence to raw data, to reach some profit (whether in prestige or funding). Second, that technology is an emergent phenomena without any autonomously set trajectory. Rather, it is shaped by a web of social and technical forces every step of the way. Third, that the primary tool of scientists, the weapon used to exert force on the socio-technical web, is the text. A successful scientific paper is one which links entities in new ways and exerts enough force to convince other scientists to believe in this linkage. We can think of a scientific paper as a network of important words, the network defines the author’s world-view. Co-Word Clustering focuses on analysis of the title or keywords used by authors.

Parallel to the development of co-word analysis, that emerged in France from a sociological tradition, in the Anglo-saxon world the notion of co-citation came into existence (Small, 1973). Co-citation analysis has traditionally been linked with the ISI Science Citation Index and has mainly been concerned with the evaluation of scientific activities. In a review of Callon's work on co-word analysis, Small points out that "if co-word links are viewed as translations between problems, co-citation links have been viewed as statements relating concepts." (Small, 1988) Each offers an interesting perspective on analysis of literature and helps in the identification of research fronts. Co-citation occurs when two publications are cited by a third, later publication. The greater the frequency of co-citation of a given pair, the greater the likelihood that it defines an established or emerging topic or subspecialty. The citation pair can be used in a citation index search to retrieve related publications. One pair can usually identify a small research front, but active research fronts generally involve several interrelated co-citation pairs. The larger the number of pairs included in a cluster, the broader the scope.

Co-citation and co-word methods can be combined in an analysis. This helps to overcome the limitations of co-citation clustering in certain forms of literature, especially where referencing is limited. Tony van Raan and his colleagues (Braam et al. 1991) have used a combination of these techniques in many of their scientometric studies, such as their study of literature on atomic and molecular biology, and have found that the combination of methods allowed them to gain a clearer picture of the cognitive content of publications.

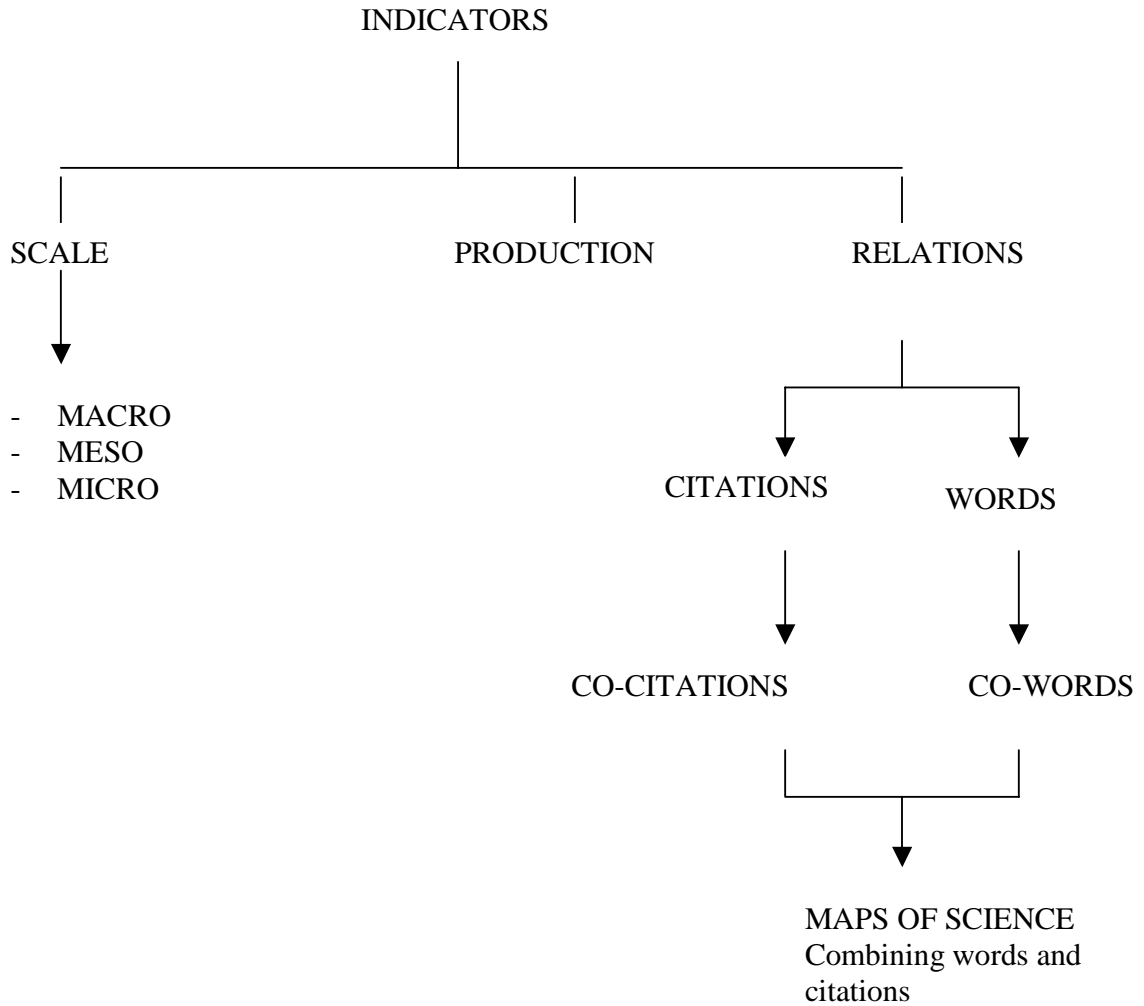
The evolution of European economies and our advancing understanding of technological innovation have led to a call for new types of statistical data and indicators (Katz and Hicks, 1997). Bibliometrics, so successful at portraying research output and impact, can be used to develop new indicators with great potential to address emerging concerns such as institutional level analysis of capabilities and networks. As Katz et al. point out; bibliometric practitioners and their indicators are so firmly associated with these classical uses, that often no further potential is seen. Following Katz et al., we believe that in bibliometric data lie opportunities to develop indicators relevant to central concerns of new theories of innovation, specifically networks within and between national systems, and variety and diversity of capability. As with any type of data, bibliometric indicators will not provide a perfect, all encompassing, ideal picture of the processes we seek to understand. However, they can make a unique contribution to pictures compiled from multiple sources, providing an unrivalled objective, disaggregated and internationally comparable time series signature of networks and capabilities.

According to Katz, classical bibliometrics focuses on the national level and international comparisons. Even with the emerging emphasis on disaggregation, international comparison and analysis of interdependencies will be required, and we illustrate the ease with which national systems can be set in an international context bibliometrically. The sectoral and intra-sectoral level data we have developed are possible due to recent advances in desktop computing. These data can make their most powerful contribution in the context of the new approaches to innovation - although we do not make those connections here (for more detailed efforts in this direction see Hicks and Katz, 1997).

Before we start discussing the various bibliometric indicators, we would like to point out that we do not aim to discuss the so-called performance indicators used to evaluate scientific development for policy purposes. We are interested in quantitatively mapping the structure and development of science in interaction with technological, economic and political developments. Two points are of interest here. First, many well-informed observers (Gibbons, *et al.*, 1994; Price, 1963; Ziman, 1994) of science and technology systems believe that science is an international system. We take it as a fact that science is international (Besselaar and Heimeriks, 2001). Furthermore, we believe that this *global science system* is one foundation on which a *global innovation system* has evolved and it is a product of the dynamic interaction between national systems that partially moulds this meta-system of innovation. And secondly, we can only provide

a *glimpse* of the value of bibliometric indicators for exploring science, technology and innovation systems.

Polanco (personal communication) suggested the following scheme to summarize the various dimensions of scientometric methodologies.



**Figure 1.** Schematic overview of different dimensions of scientometric methodologies.

## 2. Print Based Indicators

This section presents a number of central questions related to the study of scientific communications with bibliometric indicators that are used in studying the development of the science system. They are accompanied by some brief commentary and methodological remarks.

### 2.1 Collective production of scientific knowledge.

For decades there has been a divide in science studies around the question of the appropriate level of analyses for studying the production and communication of scientific knowledge. Qualitative studies of science (sociology of scientific knowledge) have focused mainly on the microscopical level of knowledge production by means of case studies (lab studies) motivated by the hypothesis that science is best described by the exchange of individuals' efforts. However, Gläser (2001) points out that scientometrics answers the question of how knowledge is produced by clearly supporting the idea of collective production on systemic level. The idea underlying the use of bibliometric indicators for measuring attributes of scientific knowledge is that scientists use their colleagues' findings. This basic relation is depicted by the references given or citations received by scientific papers. Consequently, citations show a very skewed distribution as is shown by **Impact or number of citations**. Impact, a classical bibliometric indicator, is a citation-based indicator for the relative importance of a paper or journal. Citation-based indicators point to *one specific, but important quality aspect* referred to as international influence or *impact*. These bibliometric indicators represent the response of the international research community to the published work of a group, expressed in references in scientific literature. By using references in their publications, researchers show how they have built on previous work. Criticisms of sociologists of science on citation analysis are based on the 'reference behavior's of scientists, which would be so unstructured that one cannot base quality assessment on citation data (see for instance Cozzens 1989, and Luukkonen 1997). We disagree with these sociological arguments on the basis of simple statistical considerations. Citation analysis does not concern one publication but a (very) large set of publications.

### 2.2 Coordination mechanisms in various disciplines

We have argued that knowledge production is a systemic process; scientists collectively produce new knowledge by recombinations of knowledge claims in existing papers and adding new claims supported by references to previous findings. The core question is then: how does order emerge and how is it maintained? From the description of collective production given above it follows that order is created not primarily by institutions but by the collective productions' subject matter; the shared body of knowledge (Gläser. 2001).

As mentioned, the key concept of this paper is the idea that the enterprise of science is best captured by the concept of an evolving communication system. Processes of feedback and of

self-reference are vital in maintaining a social system. This insight allows us to articulate some of the operations that construct the boundaries of the science system like peer reviews (Wouters, 1999) and editorial boards of scientific journals (Fujigaki, 1998). As Wouters points out, scientific communications can be represented by two information cycles. The first –primary- cycle is the process of submission to journals and peer review. Scientific communication can also be communicated reflexively, by measuring scientometric attributes of the communications, which in turn can be communicated. In terms of the information cycle representation, the indicators appear in the form of a second cycle. This cycle processes information about the primary information cycle, in the form of meta-information (Wouters, 1999). Together, these cyclic communication patterns result in a system that can be described as a self-maintaining evolving communication system with well definable boundaries and well definable boundary constructing processes.

In order to further refine bibliometric analysis beyond publication counts, over the past ten years, tools to conduct **co-citation and co-word analysis** have been developed (what Daryl Chubin calls the "second generation" of quantitative analysis of science) and applied to cross national comparisons. These methods provide data that allow analysts to 'map' centers of scientific excellence and to show trends and developments in scientific thought. Co-citation analysis-the more developed of the two analytic approaches-involves identifying pairs or groups of articles that are cited together in other articles or publications as cited in the Science Citation Index. This analysis allows for examination of many facets of national scientific activity, including scientific collaboration, the relative strength of centers of research as viewed by peers, and the importance of a scientific contribution over time (Rand Co).

Co-word analysis involves assigning keywords to a paper or article and conducting frequency analysis on the words chosen. AFJ van Raan, one of the pioneers of this method, conducts a two-step word frequency analysis on specific words. In the first step, after the occurrence of words has been established, words that appear a statistically significant number of times in the years being analyzed is determined. Van Raan also uses these methods to identify 'emerging' words-those that indicate recent developments-by using conference papers instead of a journal set or publications. Despite its promise as a more precise indicator than citation counts, the use of co-word analysis is controversial among scholars. Leydesdorff (1989) has shown that words and classifications are an order of magnitude less specific than citations as a static indicator of the output of science. Using co-word citations, analysts can build indices that allow comparisons within and across fields of science. Three examples of indicators that are described in the literature we examined are the Jaccard index that measures the relative degree of overlap between words within a given database, the inclusion index, which measures broad ranges of word and the proximity index which enhances the links between less frequently occurring word, allowing improved representation of new developments

## **2.3 The landscape of knowledge production: Disciplines and Specialties**

The landscape of knowledge production is not a homogeneous area, but consists of dense and empty patterns of communications. The clusters of dense communications can be identified as disciplines and within those disciplines specialties. The concept of membership (and consequently the boundaries) of scientific communication has been a major theme in science studies. In sociology of science, the endeavor has been abandoned altogether, arguing that boundaries take shape in local and temporarily situations. (Gieryn). The delineation of specialties and disciplines depends on the sociometric measure applied. In other words, specialties and disciplines have no inherent boundaries (Woolgar, 1976).

Not only can journals be classified in different fields according to classification schemes provided by the SCI, but also using aggregated journal-journal citations as listed in the Journal Citation Reports (JCR) of the Science Citation Index (SCI), the citational environment of the entrance journal can be defined as all journals that cite or are cited by this journal above a threshold percentage. The method is described in Cozzens and Leydesdorff (1993) and Van den Besselaar & Leydesdorff (1996). Factor analysis reveals clearly identifiable clusters of journals representing scientific disciplines. By repeating this procedure using data from previous years, information is obtained about the recent development of the field and its environment.

A method for empirical analysis of evolving scientific communication systems has been developed that reflects the phenomena of systems boundaries and identity construction. Using aggregated journal-journal citations, the citational environment of the entrance journal can be defined as all journals that cite or are cited by this journal above a threshold percentage. The method is described in Cozzens and Leydesdorff (1993). Factor analysis reveals clearly identifiable clusters of journals representing scientific disciplines. Groups of scientific journals define certain scientific disciplines or paradigms. The use of references and title words within this group of journals is highly codified, which results in skewed distributions of cited references, title-words, publications per journal, publications per author, publications per country, etc. These scientific specialties are expected to behave as self-organizing communication systems that maintain some kind of 'plastic identity' (Maturana, 1980). Of course, despite the maintained identity, the definition of the fields under study does change. If one follows the actors historically (Latour 1987), one obtains a reflexive understanding about how these definitions have changed. From an evolutionary perspective, however, one expects that some of these historical elements have been carried over into the current understanding, while other elements may have faded away.

The characteristics of a cluster of journals, 'a scientific specialty', and its development in relation to its journal environment can provide quantitative indicators for the degree of interdisciplinarity. Traditional scientific specialties generally consist of a stable set of journals in a stable environment while mode 2 oriented specialties tend to have a changing set of journals with a larger amount of citational links to other specialties. (Van den Besselaar and Heimeriks).

## 2.4 The research front: Research Topics

Scientometric indicators like cited references and **(title-) words** can be distinguished by the way they refer to different levels of the textual organization. From an evolutionary perspective on scientific communication, the text –that is words and co-words- provide the variation (Callon, 1986). By referencing, a subset of these texts is selected. References and citations can therefore be considered indicators of these selection processes (Leydesdorff and Wouters, 1999). The way researchers draw on earlier works, and their sharing of a set of exemplars, is considered to be reflected in the referencing practices of the specialty members. On the other hand, the shared interest in a set of research problems and concepts is expected to be reflected in the word-patterns. The congruence in both mapping practices is presupposed in most scientometric studies. However, this assumption was criticized by Braam et al. (1991).

Where the larger research fields can be operationalized as journals systems, research topics are a much lower level of aggregation, of (new) knowledge claims with respect to the topic under study. Research topics can be defined as document systems; that is as related papers, using a relevant criterion. If we are able to identify these document systems, it enables us to investigate the cognitive dimension of the system (what is the research about?) with institutional dimensions of the system (who is conducting this research?). In this way, we may be able to investigate network formation within the STI system, based on the interaction of institutional factors (EU RTD policies, and the process of further European integration), and cognitive factors (knowledge

production on specific topics). Thus, we can distinguish three levels: On the highest level, we have the research fields or scientific specialties, operationalized in terms of sets of journals. The specialty may consist of sub-fields, sometimes in the form of a single – more specialized – journal. STS is the good example, as the various journals represent all different sub-fields within STS. An even lower level is that of specific research topics, represented by related papers. (Besselaar and Heimeriks, 2001)

**Bibliometric maps of science** are landscapes of scientific research fields created by quantitative analysis of bibliographic data. In such maps the 'cities' are, for instance, research topics. Topics with a strong cognitive relation are in each other's vicinity and topics with a weak relation are distant from each other. Data derived from co-citation and co-word analysis enable analysts to perform multidimensional scaling (MDS) and mapping of scientific activity. Van Raan and Peters and others, for example, have used the indices developed through co-word analysis to create a visual representation of scientific activity through cluster graphs and multidimensional scale maps.

Perhaps in part because they are still in the process of being developed, co-citation and co-word analysis and multidimensional mapping created from this data, have a number of weaknesses associated with them: Databases do not carry all published materials, for example. What is carried may not have complete citations. All-author citations can lead to double counting, and obligatory citations to colleagues or professors can distort the counts. Co-word citation analysis depends upon subjective choice of key words and so is subject to bias. Scholars are working on techniques to remove distortions from these indicators. Even so, as Daryl Chubin has noted, bibliometric indicators are, by definition, "statistical proxies for unmeasured parameters in a complex economic political, and social system of knowledge production. They were never designed with policy in mind..." and so would have to be placed in context to be applicable to science policy and research allocation decisions.

## 2.5 Geographical distribution and Transnational research collaboration

The classical measure of research output as measured by paper count, with paper used here to designate various media for scientific texts. Paper counts provide a very simplified and approximate measure of the quantity of scientific output. It has been noted that the number of scientific publications has been growing exponentially since their coming into existence. The output in number of papers per region or country provides a crude measure of the geographical distribution of knowledge production. The number of co-authors on a paper is a measure of co-operation at a national or international level. Glänzel found that international co-authorship, in general, results in publications with higher citation rates than purely domestic papers. International collaboration has, however, not the same influence on publication profiles and citation impact of each analyzed country. In a study of co-authorship patterns in mode 2 fields of science, Van den Besselaar and Heimeriks (2001) found that in biotechnology, a European dimension is emerging while other fields, like information science and artificial intelligence are very much dominated by the USA.

### 3. Changing knowledge production

In an OECD report entitled *The Global Research Village* (1998) is described how information and communication technologies affect the science system. According to this report, ICT related changes underlying the evolving science system have three main sources; technological change in the ICT industry (mostly driven by needs unrelated to science); scientists' efforts to develop their own tools; and government programs specifically designed to foster developments in ICT and apply them to scientific needs. The main technological developments that make this evolution possible are identified; cheaper processing capacity and more user friendly software tools, increased storage capacity and information delivery technologies, and finally electronic networks that constitute the infrastructure which provides scientists with new means of communication.

While there are countless other factors that have contributed to the changes occurring in human consciousness, none is perhaps more important than the shift in communication technologies from print to computer. As Rifkin (2000) points out, great changes in human consciousness have always accompanied changes in forms of communication human beings use to create their social relations. The last great shift in communication technologies, from oral to print culture, came at the dawn of the modern era and changed forever the nature of human consciousness. The print revolution facilitated a way of thinking that was ideally suited to a society of reason and progress. Arguably, science and the print medium made each other's existence possible. To begin with, the new print medium redefined the way human beings organize knowledge. The mnemonic redundancy of oral communication and the subjective eccentricities of medieval script were replaced by a more rational, calculating, analytical approach to knowledge. Print replaced human memory with tables of contents, pagination, footnotes, and indexes. Print organizes phenomena in an orderly, rational, and objective way, and in so doing encourages linear, sequential, and causal way of thinking. By eliminating the redundancy of oral languages and making precise measurements and descriptions possible, print laid the foundation for the modern scientific worldview. Phenomena could be rigorously examined, observed and described, and experiments could be made repeatable with exacting standards and protocols, something that was far more difficult to achieve in an oral culture. In sum, print culture introduces knowledge that can be communicated, reproduced and recombined independent from individuals: the birth of science.

Rifkin argues that electronic communication is organized cybernetically, not linearly. The notion of sequentiality and causality are replaced by a total field of continuous, integrated activity. In an electronic world of communications, subjects and objects give way to nodes and networks, and structure and function are subsumed processes. Electronic communication organizes knowledge differently from the way print technology does. Hypertext replaces the more limited and narrow kind of print referencing. A self-contained book, with a set of number of facts, makes room for an open ended field of information as footnotes and references are expanded indefinitely, creating new subtexts and metatexts. Whereas a printed book is linear, bound and fixed, hypertext is associational and potentially boundaryless. A printed book is exclusive in nature and autonomous in form. Hypertext, however, is inclusive in nature and relational in form. In other words, printed books have a beginning and an end. Hypertext is merely a starting point from which users make connections between related materials. It is never complete. The printed book is a product, while hypertext is a process.

This raises some interesting issues on the development of scientific research and the quantitative methods of mapping this development. As a consequence of the information revolution in society at large, and science in particular, researchers are developing new information and communication patterns. According to NIWI (2000) this revolution finds its roots in the intimate interaction between information and communication technologies (ICT), and advanced scientific and technological research and innovation. Our thesis is that the sciences are in the midst of an informational revolution. Scientific research is going through an informational turn. Of course,

information has always been central to scientific research, but the emergence of digital information and ICT has enabled a radical lowering of costs related to information dissemination, both in pure form and black boxed in technologies. In short, information is a resource, raw material and output in the process of knowledge production. This informational turn in the sciences coincides with the processes of commercialization and commodification of scientific knowledge, the new hybrid roles of academic research institutes and universities, the transformation of economic mechanisms by technology and innovation, a wider variety of types of research output and more attention to norms and values.

Consequently, the processes and operations that maintain the science system are changing. Fujigaki et al (2000) point out that the validation boundaries emerging from the differences between classical and new modes of knowledge production are changing. Traditional scientific research was clearly grounded in paper based scientific communications. The new mode of knowledge production (mode 2) is more heterogeneous in nature; quality control is no longer simply a matter of peer review, consensus is formed outside traditional disciplinary boundaries, research outcomes are influenced by social accountability. This has enormous consequences for bibliometric analysis of knowledge production.

Due to the nature of the print medium, scientific papers are very well codified. It is easy to identify whether or not a paper is communicated within the science system, whether it is scientific or not. Also, the meaning of words within a scientific context is more codified than in other domains (Leydesdorff, 1997) because meaning is reflexively attributed to communicated information within the context of a well-defined system. Although the reasons that cited-references are used can be very diverse on the level of individual scientists (e.g. the negative citations), on the aggregate level of the science system, it provides a very clear, unambiguous representation of the shared knowledge base of a scientific discipline.

Scientific knowledge that is communicated in electronic media is less codified, more heterogeneous than its print equivalent because the operations that define the boundaries of the scientific system do not (fully) apply to electronic communications. This poses an immediate problem for scholars studying the development of science using traces of electronically communicated information. The question arising is, where does science leave off, and society begin? What is science? Or better, where is science located on the web?

However, before going into these questions we would like to point out that the problems of the boundaries of science also provide new opportunities for studying the development of science. It has often been argued that science increasingly develops in interaction with its political and applicational environment. The linking patterns of universities and research institutes might provide immediate information about the context of knowledge production. The knowledge networks that can be mapped on the web do not have any equivalent in print media! As Leydesdorff (2001) points out, the presumed heterogeneity of the nature of electronic communications provides also new opportunities. Le Pair (1988), for example, found that citations do not provide us with an accurate representation of technological achievements, because knowledge can be built into technological artifacts without necessarily leaving the formal trace of a citation in the scientific literature.

It becomes clear, that new methodologies, concepts and theories are required to deal with changing nature of scientific knowledge production and communication. On one hand scientific research provided new tools, database, equipment and media that in turn influence the production of knowledge. The concept of science as an evolving communication system that has been introduced above can be useful in understanding these developments.

## 4. From Bibliometrics to Webometrics

Bibliometrics are still widely used as a generic term for the correlated fields of sciento-, info-, techno- metrics where publications are considered the elementary units of scientific information and the main source of indicators. The diversity of new patterns of communication on the electronic network blurs sometimes the frontiers between formal and informal circulation, between activities taking place inside and outside 'science'.

When we wish to study the development of science using web-based indicators, there are a number of options; electronic journals, intermediate communications and the heterogeneous internet at large.

### 4.1 E-journals

Electronic journals provide an excellent opportunity for quantitative analysis similar to the bibliometric analysis of print articles. Most electronic journals are internet versions of existing print journals. The print based scientometric indicators could be applied to these journals. However, as Aguillo points out, there are some communicational aspects of e-journals that could modify the emerging patterns; as electronic publications permits more flexibility and speed in the dissemination of scientific information. It is expected that electronic publication lead to a wider distribution of drafts with more and faster peer comments. E-journals provide new kind of analysis. Is there a relation between number of times a paper is downloaded and number of citations? How often are papers changed and updated after initial submission? How extensive are the changes? What proportion of preprints are replaced by peer reviewed reprints?

### 4.2 Intermediate communications

Publishing the results of scientific research was, for many years, a symbiotic interaction between researchers and publishers, because the most effective way scientists could disseminate their results was through journals, produced by professional societies and independent publishers. Electronic communication has created new ways to distribute such results and is forcing researchers and publishers to reassess the old procedures and consider new possibilities as we learn to use the Internet. Now, not only can authors easily disseminate their results, but networked readers can have cheap, fast access to more scientific literature and have it in a form that facilitates its use in their own research. A larger number of automated archives for electronic communication of research information have been operational in many fields of physics, and some related and unrelated disciplines, starting from 1991. These archives now serve over 35,000 users worldwide from over 70 countries, and process more than 70,000 electronic transactions per day. In some fields of physics, they have already supplanted traditional research journals as conveyers of both topical and archival research information. Many of the lessons learned from these systems should carry over to other fields of scholarly publication, i.e. those wherein authors are writing not for direct financial remuneration in the form of royalties, but rather primarily to communicate information (for the advancement of knowledge, with attendant benefits

to their careers and professional reputations). These archives have in addition proven equally indispensable to researchers in less developed countries. (Ginsparg, 1996)

## 4.3 Internet

The third option for bibliometric analysis of the science, technology and innovation system on the internet is using web sites in general, not necessarily e-journals. This immediately poses the problem of the unit of analysis. In print communications, papers are communicated in journals, what would be the equivalent of these information carriers on internet? And what is communicated?

Aguillo suggests a new concept of a site; the presence of research groups, universities and other R&D related institutions on the Internet. 'A web site comprising hierarchically grouped pages, and represented by the URL address of the highest level, is considered as a unit if it has been defined according to aspects: it is identified as formally different from other Web sites (documentation unit) and is recognizable as representing a research group with the aim of being present at the web (institutional unit). The institutional character of the proposed new unit eases comparative analysis with external data, as usually there are scientometric analysis available following that institutional grouping criterium. Moreover, the global viability of the database increases since a better identification of Web sites is possible if a physical real counterpart is available. (Aguillo, 1998).

However, one of the main problems of webometrics in this context is related to the 'codification of knowledge'. As Leydesdorff (1997) points out, common languages allow for one layer of reflexivity without confusion. Codification in specialist, scientific languages allows for the (provisional) stabilization of meaning in nearly decomposable reflexive layers of communication (e.g., 'paradigms'). In other words, scientific communications are more codified than non-scientific communications. The meaning of 'words' and 'cited references' are less ambiguous and heterogeneous in scientific communications than in other contexts. But even in a codified, scientific context words and references remain ambiguous in their meaning.

A similar problem can be found in relation to hyperlinks. We do not know for sure why people on the web link up to other pages. Björneborn and Ingwersen write:

Obviously, the breakthrough for everybody to express themselves, practically without control from authorities, to become visible world wide, also by linking to what pages one wants to link, to assume credibility by being 'there', and to obtain access to data, information, values and knowledge in many shapes and degrees of truth, has generated an a reality of freedom of information, also in regions and countries otherwise poor of infrastructure.

Finally, as Björneborn and Ingwersen point out, the web is an information space quite different from common scientific and professional databases, the similarities between electronic and print medium are often superficial. Most notably, time plays a different role on the web.

However, many observers of science believe that the science system and the communication functions of the print medium are inseparable. According to these scholars a two layered communication system will evolve:

Poultney (1996) states that most probably a "two-tier system" will evolve. The first tier is a "free space" for preprints and other "preliminary" communications. It is a representation of the scientific enterprise in "real time". The second tier is the world of more formal publications. [...] The first tier is the place where informal and formal communications will become closely connected. (Van Raan, 2001)

The 'second tier' of scientific communication will consist largely of already existing journals, of which most soon will have a print as well as an electronic form. The electronic form of this journal should be considered semi-electronic since it communicates papers according to print-medium dynamics.

## 4.4 Electronic Indicators

This section presents a number of central questions related to the study of scientific communications with bibliometric indicators that are used in studying the development of the science system. They are accompanied by some brief commentary and methodological remarks.

### 4.4.1 Collective production of scientific knowledge.

In an excellent overview of webometric analyses, Björneborn and Ingwersen (2001) point to some recent investigations in webometrics. They look into citation analysis, i.e., link page analysis in terms of Rousseau (1997) and Web impact factor studies. This study shows that the distribution of top level domains for the sites follows the ubiquitous Lotka distribution. Similarly, Rousseau demonstrates that the distribution of citations to those sites also follow a Lotka distribution. The proportion of self-citations was estimated to 30%. Concluding, at the present state of search engine coverage and retrieval methods, "the exciting concept of Web-IF appears to be a relatively crude instrument in practice" (Thelwall, 2000).

The world-wide-web may revolutionize the way scientists communicate with each other. More and more researchers publish original work on the WWW rather than (or preceding) publication in paper journals. One problem of this is, that for the reader it is difficult to evaluate the quality of the published paper - as opposed to traditional publishing, where for example the scientific "rank" of the journal (as measured by its impact factor) is known. In traditional (paper) publishing, scientometric indicators such as the impact factor are based on bibliometric analysis (citation analysis) and serve as quality indicators for published scientific work. The SCI does not take into account electronic works published on the WWW. Attempts have been made to develop a WCI (WebCitation-Index). In analogy to the SCI, the WCI database indexes and counts citations ("hyperlinks") of one scholarly work published on the WWW citing another scholarly work published on the WWW. It is important to understand that the WebCitation Index does strictly focus on scholarly works published on the web (such as medical articles) and does not include "generic" webpages or whole websites.

The development of a global academic information/communication system also suggests new ways of measuring the impact of scientific contribution that take into account the cooperative aspect of science. The American Physical Society's Task Force's Report on Electronic Information Systems, cited by Harnad, notes that: "Unlike inert publication counts or even citation counts, sensitive measures of "air-time" and "flight-route" for new ideas and findings (how often they are accessed, by whom, and where they lead in subsequent electronic and paper literature) would be helpful not only to those who are trying to evaluate the importance of a given scholar's contribution but also to historian of ideas trying to make sense of the evolution of knowledge."

Candidates for indicators could be remote files retrieval counts (since archiving published or pre-print work on ftp sites is now common practice) or clients hypertext links counts. These can point to "classical" links like citations or author's addresses but they can also assume new forms like the "annotation" or the "is-interested-in" links available on WWW. Similar indicators can be devised in order to assess the impact of information servers or services offered.

## **4.4.2 Coordination mechanisms in various disciplines**

In general sites are positioned in relation to each other. This provides a starting point for analysis. The levels, direction, number, relations, use and the life span of links between sites are all indicators for the nature of clustering of web sites. The meaning and significance of citation is expected to be quite different in the two environments of scholarly publishing and the WWW. The question is whether the types of intellectual mapping of disciplines made possible with citation indexes like Science Citation Index (SCI) and co-citation techniques might be applied to charting the contents of cyberspace. If we assume that the WWW is a prototype of the distributed digital libraries of the future, it would be helpful to know if the tools and techniques developed for the analysis of intellect structure in paper-based libraries will be able to make the transition to this network-based environment (Wilensky, 1995). Counts of ingoing and outgoing links can be seen as citation and reference analysis respectively. However, due to its dynamic and distributed nature, the Web often demonstrates web pages simultaneously linking to each other- a case not possible in the traditional paper based citation world (Björneborn and Ingwersen, 2001).

Ingwersen (1997) introduces the application of informetric methods to the World Wide Web (WWW) also called internetometrics. The paper describes a number of specific informetric analysis parameters. The methodological approach is comparable with common bibliometric analyses of the ISI citation databases. The following online processing tools: Rank, Map, and Target, provided by Dialog, are incorporated in order to perform analyses of citation to and from isolated sets of documents as well as to carry out diachrone journal analyses. These analyses imply further to determine journal impact factors of ISI journals. Measures of the scope of internationalisation of journals are proposed and demonstrated. By the combined application of the Rank and Target commands we demonstrate a hitherto overlooked possibility of working with bibliographic coupling online and mapping of scientific fields

## **4.4.3 The landscape of knowledge production: Disciplines and Specialties**

Björneborn and Ingwersen (2001) identify three main directions to perform knowledge discovery on the web, a set of methodologies to find related clusters of communications. They are concerned with exploiting 1) webpage content, 2) link structures and 3) users' information behaviour (searching and browsing). Their focus is mainly on the exploitation of link structures; an approach with strong kinship to bibliometric citation analysis, but not only by means of strong ties. They point out that links weave together web documents in a complex structured hypertext corpus. Link structures represent implicit human 'annotations' that can be exploited for knowledge discovery, for example inferring web communities (Gibson et al., 1998), identifying authoritative web pages (Kleinberg, 1998; Cui, 1999), topic distillation (Bharat and Henzinger, 1998), or improving search engine ranking algorithms (Brin and Page, 1998).

## **4.4.4 The research front: Research Topics**

Of course, words and co-words can serve as a basis for bibliometric analysis on the web just as in the print world. Again, the (lack of) codification of electronic communication poses a challenge to quantitative analysis. However, besides the words, web pages contain other information in the source code, like key words, that might provide additional and more codified sources of information for analysis.

Also the web equivalents of 'science maps' have been developed. For example, Rogers et al (2000) introduce a preliminary assessment of the potential value of new web mapping techniques that inform and ultimately facilitate meaningful participation in science and technology debates. A number of different attributes of web-based communication can be combined; the freshness of the link, the origin of the information (gov, com, org or edu), and the location of the actors in the communication network.

#### **4.4.5 Geographical distribution and Transnational research collaboration**

The geographical location of a site, publication or person on the web is hard to determine unambiguously since we will not be counting corporate sources but "electronic addresses". There is no assurance that they represent only one person or indeed always the same person. Some information services offer only "hosts access count" and, as we have seen, a host may serve up to hundreds of users. As Gilbert points out, at least 4 types of electronic addresses can be identified. However, the URLs of websites often do provide information about the location and or the type of institution. In a study, Leydesdorff and Wouters (1999) suggested that Triple Helix configurations on the Internet could be searched by using hyperlinks between industrial (www\*.com), academic (www\*.edu), and governmental (www\*.gov) texts (cf. Aguillo, 1999).

Boudourides et al. (1999) used the advanced search technology of AltaVista for the measurement, because it allows for searching on domain names and using Boolean operators. Following this lead, Leydesdorff and Curran were able to analyze the differences and similarities between national systems of innovation as indicated by national domain names (e.g., ".br" or ".nl") as against the so-called "generic Top Level Domains" (like ".edu" and ".com"). Subsequent questions to be addressed concern the use of national languages versus English, the baseline for international comparisons, the differences in using the various search terms, etc.

# 5. Methodological considerations

In this chapter some methodologies will be discussed with some examples of recent research.

## 5.1 E-journals and webometrics

(Contribution of ARCS: Agelika Zartl, Edgar Schiebel)

Due to the increasing amount of information available free or nearly free on the web, scientists are concerned with the application of bibliometrics on the web. In the last years some efforts have been made to apply bibliometric indicators developed for analysis of printed articles and journals also for web pages or hyperlinks. Webometrics displays several similarities to informetric and scientometric studies and the application of common bibliometric methods. The content analysis of web pages corresponds to traditional publication analysis. Counts and analyses of outgoing links from web pages, (outlinks) and of links pointing to web pages (inlinks) can be seen as reference and citation analysis. Outlinks and inlinks are then similar to references and citations, respectively, in scientific articles. The main difference between traditional printed and electronic publications is the fact that due to the dynamic nature of the web web pages are often simultaneously linking to each other. Since the web consists of contributions from anyone who wishes to contribute and usually any kind of peer reviewing is missing, the quality of information or the knowledge value often cannot be determined or is problematically. But citation-like link analysis may reveal clusters of sites to be reviewed (Björneborn and Ingwersen, 2000).

Webometrics can be performed for identification of relevant articles or electronic journals, for mapping a field of interest as well as for determination of the impact of scientific work on research. One part of webometrics focuses on electronic journals as the counterpart of printed journals. Aim of these investigations is to determine the importance of e-journals for scientific research compared with print media based on citation analysis.

The other part of webometrics does not focus on a specific type of publication or web site. Those studies intend to apply bibliometric methods on different kinds on web pages or hyperlinks. Therefore the citation analysis uses links, domains or text strings instead of references – the so called *sitations*. One of the crucial points for citation analysis is the selection of the basic sample of links or domains. Furthermore several aspects have to be considered for interpretation of the results which are of no concern for citation analysis of printed articles.

### 5.1.1 E-journals

In order to evaluate the applicability of existing bibliometric concepts for web analysis and to determine the impact of electronic journals on scholarly communication and research, several studies have been performed in the last years. Besides demographic parameters like the number of articles published, the number of subscriptions, the discipline, the number of articles downloaded or the charging policy, the main indicator used is the impact factor based on citation analysis. In contradiction to most of the indicators the impact factor analyses to which extent researchers are influenced by and build their own work upon research published in e-journals or print media (Harter, 1996).

For determination of the impact factor of publications in print media the Institute of Scientific Information ISI publishes three citation indexes: Science Citation Index SCI, Social Science Citation Index SSCI and Arts and Humanities Citation Index AHCI. The ISI publishes the Journal Citation Reports JCR that includes several citation-based measures of journal impact for the journals indexed. Only few electronic journals are covered by ISI. Furthermore using ISI for evaluation of the impact factor of electronic publications one has to take into account that only references in (important) journals are counted as citations. Hyperlinks to web pages are not included (Harter, 1996).

Harter and Kim (1996) and Harter (1996) compared the impact of citations of scientific articles published in e-journals and print media. Fosmire and Yu (2000) based their study on the results on Harter (1996) and analysed the impact of free available scholarly e-journals. In their terminology 'scholarly' is defined to be peer-reviewed and with a scholarly treatment, that means it contains references (Fosmire and Yu, 2000). Both used the citation analysis and calculated the impact factor, the immediacy factor and the total number of articles published in the journal in the year of measurement.

The impact factor measures the current impact of the typical recently published article in a particular journal. For example the impact factor of a journal in 1995 can be calculated using the following equation:

$$\text{Impact factor for 1995} = \frac{\text{(Number of citations received in 1995 to 1993 and 1994 articles)}}{\text{(Number of articles published in 1993 and 1994)}}$$

The immediacy factor gives the number of citations received in the year of measurement to articles of that year in relation to the number of articles published in that year.

$$\text{Immediacy index for 1995} = \frac{\text{(Number of citations received in 1995 to 1995 articles)}}{\text{(Number of 1995 articles)}}$$

Due to that definition the immediacy index is an indication of how quickly readers incorporate articles from a journal into their research and therefore allows the identification of current journals. In the mentioned studies from Harter (1996) and Fosmire and Yu (2000) the determination of the different indicators is based on the following steps:

- Selection of several e-journals
- Identification of relevant titles in science, technology, and medicine (STM) fields
- Search for the number of citations to the selected titles in the JCR and ISI's Web of Science database
- Determination of the different indicators

Harter (1996) determined only eight out of 39 e-journals which were cited ten or more times over their lifetimes (Table 1).

**Table 1: the eight most highly cited e-journals (Harter, 1996)**

Name of e-journal	Discipline	Print version?	Number of citations to the e-journal
<i>Bulletin of the American Mathematical Society</i>	Mathematics	Yes	> 1,500
<i>Online Journal of Current Clinical Trials</i>	Medicine	No	190
<i>The Public-Access Computer Systems Review</i>	Library and information science	(Yes)	111
<i>Digital Technical Journal</i>	Computer science	(No)	38

<i>Psycoloquy</i>	Psychology	No	35
<i>Interpersonal Computing and Technology</i>	Effects of technology on society	No	14
<i>Electronic Journal of Communication</i>	Communication	No	11
<i>Postmodern Culture</i>	Modern culture	No	10

While Harter (1996) determines a very low impact of e-journals on scholarly communication compared with printed articles, Fosmire and Yu (2000) identified some free scholarly e-journals with significant impact on their specific fields of research. Table 2 shows the results for the impact factor, the immediacy factor and the number of current articles of medical journals in the ISI's Citation Reports 1998 determined by Fosmire and Yu (2000).

**Table 2: Comparisons of the 1999 Impact Factor, Immediacy Index, and Number of Current Articles of Medical Free Scholarly Electronic Journals with the 99th, 90th, 75th, 50th, and 25th percentile Medical journals in ISI's Journal Citation Reports 1998. For the British Medical Journal, 1998 JCR data is reported for that title instead of 1999 data. (Fosmire and Yu, 2000)**

	Impact Factor		Immediacy Index		Current Articles
<b>99<sup>th</sup></b>	<b>42.93</b>	<b>99<sup>th</sup></b>	<b>7.28</b>	<b>99<sup>th</sup></b>	<b>1681</b>
Emerging Infectious Diseases	7.17	British Medical Journal	2.03	British Medical Journal	916
British Medical Journal	5.33	Emerging Infectious Diseases	0.95	Southern Medical Journal	312
Blood, Cells, Molecules Diseases	4.77	Medicine and Global Survival	0.60	<b>90<sup>th</sup></b>	<b>261</b>
Alzheimer's Disease Review	4.38	<b>90<sup>th</sup></b>	<b>0.49</b>	Radiographics	196
<b>90<sup>th</sup></b>	<b>3.03</b>	<b>75<sup>th</sup></b>	<b>0.27</b>	<b>75<sup>th</sup></b>	<b>149</b>
Radiographics	1.88	Radiographics	0.16	Emerging Infectious Diseases	123
<b>75<sup>th</sup></b>	<b>1.81</b>	Blood, Cells, Molecules Diseases	0.14	Annals of Saudi Medicine	122
Eurosurveillance	1.71	Eurosurveillance	0.14	<b>50<sup>th</sup></b>	<b>73</b>
Southern Medical Journal	1.01	<b>50<sup>th</sup></b>	<b>0.12</b>	Blood, Cells, Molecules Diseases	42
<b>50<sup>th</sup></b>	<b>0.94</b>	<b>25<sup>th</sup></b>	<b>0.04</b>	<b>25<sup>th</sup></b>	<b>38</b>
<b>25<sup>th</sup></b>	<b>0.44</b>	Southern Medical Journal	0.02	Eurosurveillance	37
Annals of Saudi Medicine	0.24	Annals of Saudi Medicine	0.01	Online Journal of Issues in Nursing	29
Dermatology Online	0.06	Alzheimer's Disease Review	0	Dermatology Online	8
Online Journal of Issues in Nursing	0.03	Dermatology Online	0	Alzheimer's Disease Review	6

Medicine and Global Survival	0	Online Journal of Issues in Nursing	0	Medicine and Global Survival	5
------------------------------	---	-------------------------------------	---	------------------------------	---

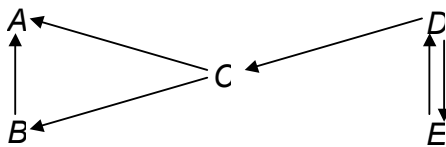
The different results can be explained by the time of the evaluation. Harter (1996) analysed articles published 1993 to 1995 while Fosmire and Yu (2000) focused on publications in 1997 to 1999. Due to these results it can be stated that the behaviour of scientists using the web as communication medium has changed and that the importance of electronic journals compared with printed ones has – at least in some research fields – increased significantly.

### 5.1.2 Web publications

There are three main directions to perform knowledge discovery on the Web:  
 web page contents  
 link structures and  
 users' information behaviour (i.e., searching and browsing).

The exploitation of link structure is closely tied to bibliometric citation analysis (Björneborn and Ingwersen, 2000).

A description of the linking pattern can be found by Björneborn (1999) based on the web topology.



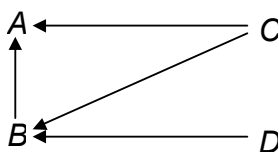
Based on the graph above the link pattern is

- C has a outlink to A
- C has an inlink from D
- D and E are reciprocally linked
- D is transitively linked with A and B through C
- A and B are co-linked from C (co-citation)
- B and C are co-linking to A (bibliographic coupling)

Bibliographic coupling measures the number of references two papers have in common to test for similarity. A clustering based on this measure yields meaningful groupings of papers for research (and information retrieval) by stating "a number of papers bear a meaningful relation to each other when they have one or more references in common" (Kessler, 1963).

The cocitation analysis is based on the assumption that if two references are cited together, in a latter literature, the two references are themselves related. The major refinement between bibliometric coupling and cocitation is that while coupling measures the relationship between source documents, cocitation measures the relations between cited documents.

Using those definitions the co-link searches can be performed by



- co citation: given co-linked A and B - find all linking Cs

example [AltaVista]: link:www.pliant.org/personal/Tom\_Erickson/SocialHypertext.html  
and link:www.ntu.ac.uk/soc/psych/miller/goffman.htm

- bibliographical coupling: given linked B - find all co-linking Cs and Ds

example [AltaVista]: link: www.ntu.ac.uk/soc/psych/miller/goffman.htm

For investigating the link structures on the web graph theoretic methods can be used. The topology of link structures affects possibilities for human and digital agents to traverse and explore the Web. Björneborn (1999) defines different criteria for a link topology, as

- levels: different Web analysis entities: web pages, hosts, domains
- direction: inlinks/outlinks - unidirectional/reciprocal
- number: one, two, many inlinks or outlinks
- relations: analogy with citations ; reflect social attitudes and interests ; linkage motives
- use: strong and weak links: more or less used: "the learning Web"
- life span: links are dynamically created, adapted, and removed - compared to paper media citations accumulated over time

In graph theory a graph is a mathematical representation of a network consisting of vertices (or nodes) connected by edges. The nodes can be Internet servers, documents (in a citation network), concepts (in a thesaurus or semantic network), etc. In a directed graph the edges represent directional relations between the nodes. The Web is an example of a directed graph with web pages corresponding to nodes and hyperlinks to edges. Graph theoretic methods can be used to analyse structural aspects of the Web (Björneborn and Ingwersen, 2000). Based on that theory Björneborn and Ingwersen (2000) discuss the concept of „small-world“-networks and transversal links and their significance and possibilities for web analysis. Transversal links – links between heterogenous web pages – can be used as a tool for information retrieval, improving the exploitation of the web.

One of the first scientists dealing with webometrics and graph theory was Abraham (1997). Abraham's strategy for visualizing and measuring the Web is based on the mathematics of morphogenesis, complex dynamical systems theory. The strategy is based on the fact that the web can be described as a tree consisting of domains, servers, and pages. There are tens of thousands of domains, several servers in each domain, and many pages in each server. Each domain has a unique name (for example, vismath.org), each server has a unique name (eg, www.vis- math.org) and IP address (eg, 162.227.70.1), and each page has a unique URL (eg, http:// www.vismath.org/index.html). These are the main choices for nodes of the web (Abraham, 1997).

The interconnections of the web, as a hypertext and hypermedia system, are links. Links connect pages, but pages are secondary to domains. Thus, given two domains, that is, nodes, all links from any page of the first domain, to any page of the second domain have to be determined. Then this simple count should be normalized. The normalization can be performed by regarding the number of all pages of all servers of the first domain as a width, and all pages of all servers of the second domain as a height, the area of this rectangle (the product of the two page counts) which corresponds to the probability of a link can be calculated. Thus, the connection strength Abraham proposes, is the ratio of the number of links to the product of the width and the height. A more precise measure might take into account the byte size of pages, or equivalently, the total storage served by each domain (Abraham, 1997). In any case, the data to construct the massive connection matrix for the entire web is to be collected by a web crawler or robot, not just once, but repeatedly.

The problems of the time dependence of the search results is also discussed by Björneborn and Ingwersen (2000). Several studies have investigated the behaviour and the results of different search engines (AltaVista, Lycos, HotBot, Infoseek, Excite, NorthernLight) over time. All of them determined inconsistencies and large variations in the determined amount of web pages.

Furthermore in many cases the best results were obtained with AltaVista because it has a large web coverage and provides search features suitable for informetric studies of the web.

Turnbull (2000) at the Georgia Tech Graphics, Visualization, and Usability Web Surveys is concerned with Web demographics. Through adequate programming, data gathering, and extensive statistical analysis the development of an outline of web users and their preferences was possible. The HyperText Transfer Protocol HTTP and the HyperText Markup Language HTML – used to implement the web – were used to perform informetric analysis (Boudourides et al., 1999). According to Turnbull (2000) the first step is to fully enable all possible server logs in order to apply bibliometric methods on the web. Different internal or external server information can be used for bibliometric analysis such as

- email logs to and from the system administrator (the data size and times, not necessarily the content),
- program usage logs
- Server-based - the log file created by the Web server itself.
- Proxy-based - via a firewall (additional layer of server protection) or some server controlled access system.
- Client-based - via client application code that can record and transmit information to the server such as cookies.
- Network-based - the programs that control system-level networking security and access on the server.

In order to analyse those information he discusses the optimal web server configuration, managing log files, downie and web usage and optimal web content setup. By adequate setup, good system architecture and diligent analysis many benefits can be obtained.

The development of a workable method for general informetric analysis of the web is described by Almind and Ingwersen (1997). They determined a number of specific informetric analysis parameters and performed a case study on danish web sites. The methodology applied is due to Almind and Ingwersen comparable with common bibliometric analysis of the ISI citation databases.

While in the ISI's citation databases descriptors are found in three forms: Author Keyword, KeyWords Plus and Research Fronts, on the web descriptors are given either by an author or by frequencies. For the subject access points of the web pages an author can use tags, such as <em> and <strong>. Frequencies of terms are measured by some web indexes. The titles of the web pages are found either within the <TITLE> or/and <H1> tag, and can uniquely be identified by the URL of the page. Whether the author is a person or corporate source can only be identified manually. Corporate source or affiliation for web pages is given by the first part of the URL, but the institution hosting a web page is not necessarily connected with the author of the Web page (Wormell, 1998).

The most common used indicator is the Web Impact Factor, a measure that uses the number of links made to a site to measure of the site's overall influence on the Web (Smith, 1999, 2). Web Impact Factors are related to Journal Impact Factors discussed above. But while for evaluation of the impact factor or other citation-based measures, the JCR is an important tool, no adequate instrument exists up to now for the web. Therefore other methods have to be applied to determine which pages are cited (linked to) by compilers of other web pages (Cui, 1999).

Web Impact Factor studies use hypertext links, which are measured by search engines such as AltaVista, rather than bibliographic citations. Web Impact Factors have been shown to be useful measures of the influence of sites belonging to organisations such as universities and research institutes.

Webometric studies have used large scale Web search engines such as AltaVista. These allow measurements to be made of the total number of pages in a domain (either a top level domain

such as .com or .nz, or a lower level domain such as vuw.ac.nz) or a set of directories, such as the pages in the directories and subdirectories and the links between them. Web search engines provide similar possibilities for the investigation of links between documents to those provided by the citation databases created by the Institute for Scientific Information.

A Web Impact Factor is the number of pages linking to a web space, divided by the number of pages in the web space. The WIF differs in respect to the time period from the journal Impact Factor, which is constrained by the methods used to compile citation indexes. The journal impact factor measures citations made in journals published during one time period, to articles published in another time period. The WIF, in contrast, is a "snapshot" from the search engine database of all links to a web space at the time of measurement (Smith, 1999, 2).

The web pages are analysed corresponding to their links internally and/or externally and the numbers of links is obtained. This method is up to now used from most of the bibliometric researchers. The problems which have arisen using this method are beside others

- inconsistency of different combinations of search operators
- dependence of the amount of identified web pages on the time of the search
- duplicates of web pages and self citation
- dependence of the search results on the search engine applied for the citation analysis

### 5.1.3 Webometric Studies

In several fields like health and medical domains „quality control“ is of great concern due to the fact that the web allows great freedom in information dissemination and linking and lacks any peer-reviewing. Everyone can link to web pages he wants to. This on one hand allows countries and regions with otherwise low infrastructure to contribute in (scientific) communication, on the other hand great uncertainty is created to the users. (Björneborn and Ingwersen, 2000). Therefore the importance of methods and indicators which allow an evaluation of the information found on the web, increases.

#### Analysis of university – industry – government relations

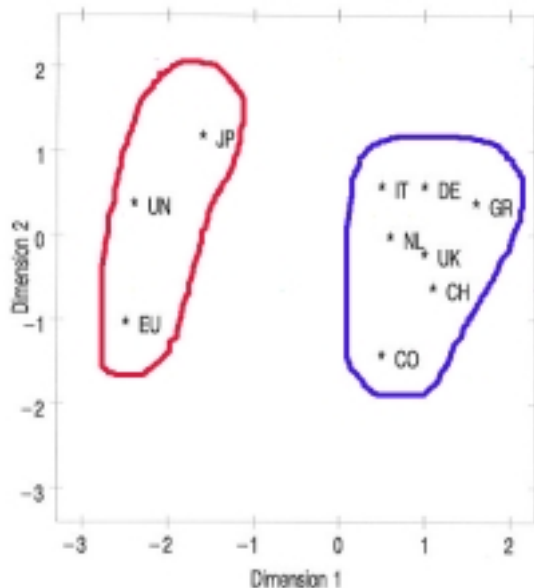
Several studies are concerned with the analysis of the relationship and the knowledge distribution between universities, industry and government, the so called Triple Helix. While Boudourides et al. (1999) focused on the analysis of web sites on the self-organisation of the European Information Society, Leydesdorff and Curan (2000) concentrated on the comparison of national Triple Helix relations of Brazil and the Netherlands.

Boudourides et al. (1999) used for their analysis web server at the level of sub-domains (sdws) and the following search commands which were combined by the AND respectively the OR operator:

**Table 3: search commands used for analysis of sub-domain web servers (Boudourides et al., 1999)**

<i>Keyword</i>	<i>Function</i>
host:name	Finds pages on a specific computer. The search host:altavista.digital.com would find pages on the AltaVista computer, and host:dilbert.unitedmedia.com would find pages on the computer called dilbert at unitedmedia.com.
Link:URLtext	Finds pages with a link to a page with the specified URL text. Use link:altavista.digital.com to find all pages linking to AltaVista.
Text:text	Finds pages that contain the specified text in any part of the page other than an image tag, link, or URL. The search text:cow9 would find all pages with the term cow9 in them.

Sdws contain a big number of web servers. For example the sub domain uva.nl includes beside others the web servers www.uva.nl, www.psy.uva.nl or www.chem.uva.nl (Boudourides et al., 1999). For the test purpose only 10 sub-domains web server were used, six belonging to European universities, one to a Japanese University and three to international sub-domains (like UNESCO or EU). The co-citation matrix was obtained by applying the command link:sdws(i) AND link:sdws(j) for  $i, j = 1, \dots, 10$ . The co-citation matrix was converted in a correlation matrix and multidimensional scaling performed. One map determined is given in figure 1.



**Figure 1: Map of the co-citation behaviour of 10 sub-domain web server on self-organisation of the European Information Society (Boudourides et al., 1999)**

The results show that citation behaviour of the European universities differs from that of more complex sub-domains belonging to international organisations or Japan. Based on these results the conclusion can be drawn that multidimensional scaling can be applied to hyperlinks in the same way as to co-citation in printed media.

Leydesdorff and Curran (2000) determined the following indicators for analysis of the relationship between industry, university and government in The Netherlands and Brazil using the Alta Vista Advanced Search Engine:

- Amount of web-sites
- Consideration of the languages used
- Searching for specific hyperlinks between .com, .edu and .gov sites (the generic Top Level Domains gTLDs)
- Analysis of time dependence (1993-1998)
- Analysis of the links between pages (Alta Vista search link:..)
- Search for words / co-words using the full text respectively the titles

Four search functions of Alta Vista were used: domain, link, text and title. Table 4 summarizes the functions and their search results. Besides those, Alta Vista provides several other search functions like url or host.

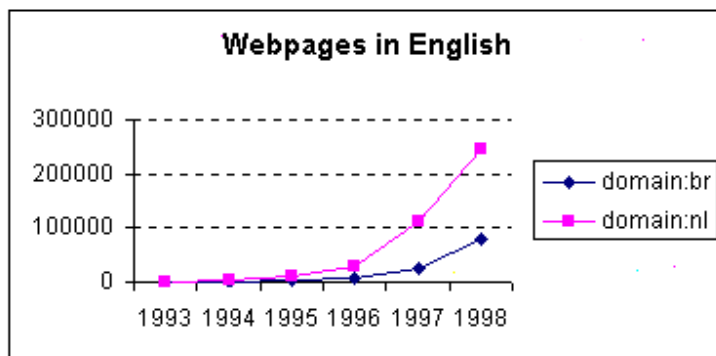
**Tabelle 4: Tags in the Advanced Search Engine of AltaVista (Leydesdorff and Curran, 2000)**

Keyword	Function
---------	----------

domain:domainname	Finds pages within the specified domain. Use domain:uk to find pages from the United Kingdom, or use domain:com to find pages from commercial sites
link:urltext	Finds pages with a link to a page with the specified URL text. Use link:www.zip2.com to find all pages linking to Zip2.com
text:text	Finds pages that contain the specified text in any part of the page other than an image tag, link, or URL. The search text:graduation would find all pages with the term graduation in them
title:text	Finds pages that contain the specified word or phrase in the page title (which appears in the title bar of most browsers). The search title:sunset would find pages with sunset in the title

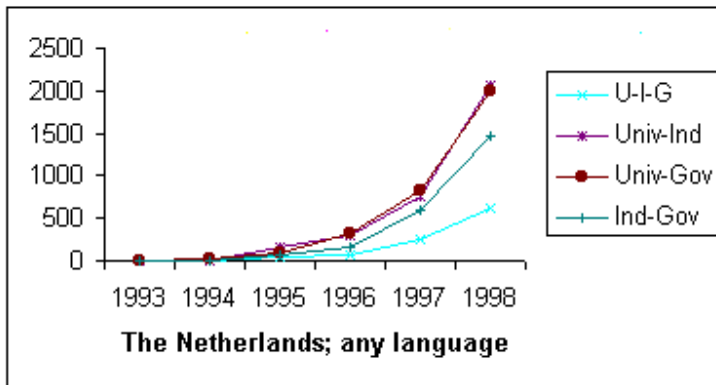
The search was performed for the domains Brazil, The Netherlands and the OR-combination of the generic Top Level Domains .com, .gov, .edu, .org, .mil (military), and .net (internet organisations).

The results show that for example the linguistic behaviour differs between The Netherlands and Brazil. The relative contribution of The Netherlands' web pages to the English-language domain is far more advanced than that of Brazil (figure 2).



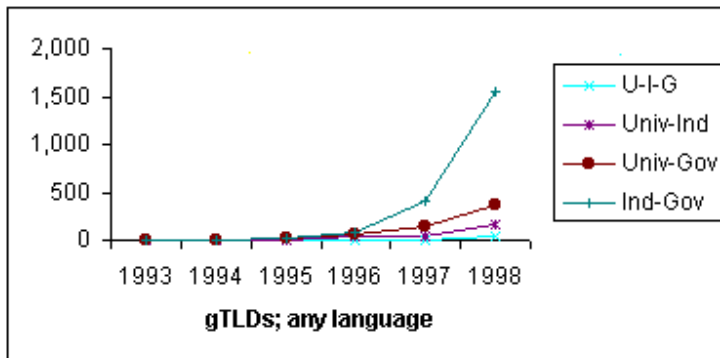
**Figure 2: Contributions of Dutch and Brazilian webpages to the English language domain (Leydesdorff and Curran, 2000)**

In order to determine the relations between university, industry and government web pages searches were performed using the terms „university“, „industry“, „government“ and the AND-combination of them. Figure 3 shows that the relations between universities and government and between universities and industry are the dominant ones, the industry-government relations lag behind.



**Figure 3: Bilateral and Triple Helix relations in the case of The Netherlands (Leydesdorff and Curran, 2000)**

Using the terms „university“, „industry“, „government“ as links Leydesdorff and Curran analysed the linking behaviour. For example the query „link:industry AND link:government“ determines web pages containing both terms in specific URLs. Figure 4 shows that industry and governmental links are dominant. The authors state that the results indicate that industry-government relations are indicators of national economies, while university-industry relations are increasingly knowledge intensive and international orientated. Summarizing the study has shown that the Triple Helix relations can be measured using the web.



**Figure 4: shared links between „university,“ „industry,“ and „government“ at the level of the reference set of combined gTLDs**

A similar study was performed by Smith (1999, 1 and 2) who compared in two studies Australian and New Zealand web spaces, respectively education and research sites in Australasia (Australia and New Zealand) and Latin America (Central and South America).

For this purpose he used the following commands:

- host:vuw.ac.nz/ for determination of the number of pages in the web space D
- link:vuw.ac.nz/ for determination of the total number of pages linking to the web space L
- The number of self-links (links from pages in the same web space) was measured in two ways to overcome the effects of Boolean inconsistency:
  - S1 link: vuw.ac.nz / AND host:vuw.ac.nz/
  - S2 host:vuw.ac.nz/ AND link: vuw.ac.nz /
- The number of external links (links from pages outside the web space) was measured in three ways:
  - E link:vuw.ac.nz/ AND NOT host:vuw.ac.nz/

- E1 link:vuw.ac.nz/ AND NOT (host:vuw.ac.nz/ AND link:vuw.ac.nz/)
- E2 link:vuw.ac.nz/ AND NOT (link:vuw.ac.nz/ AND host:vuw.ac.nz/)

Several measurements were carried out over several days until a measurement was achieved where the Boolean inconsistency was zero or very low.

The selected observations were used to calculate

- the overall WIF: L/D
- the external WIF: (average E,E1,E2)/D
- the self-link WIF: (average S1,S2)/D

A shortcut of the results is given in table 5.

**Table 5: Web Impact factors of university and research web spaces in Australasia and Latin America (Smith, 1999, 2)**

Country	Institution	URL	Web pages	External WIF	Internal WIF	Staff	English pages (%)
Australasian web spaces							
Au	University of Queensland	uq.oz.au	4533	3.58	0.11	1560	95.6
Au	Australian National University	anu.edu.au	44938	2.00	0.56	725	96.5
Nz	Victoria University of Wellington	vuw.ac.nz	9056	1.77	0.51	474	93.8
Au	University of Melbourne	unimelb.edu.au	42944	1.56	0.68	1893	97.2
Nz	University of Auckland	auckland.ac.nz	8657	1.36	0.44	1175	87.2
Au	Latrobe University	latrobe.edu.au	11473	1.20	0.49	2442	94.9
Au	Monash University	monash.edu.au	45981	1.06	0.52	1657	95.4

## Medical websites

Cui (1999) performed a bibliometric analysis of health care related web sites. The analysis focused on the pages linked to in the „other links“ section of the web pages of 19 of the top 25 US medical schools. The identified web pages were ranked according to their cited frequency by examining the links made from these pages. The determined links were analysed by a special software program called Checkweb and cleaned up in order to obtain only the active links to external URLs. The frequency of the URLs was counted taking into account the different URL-levels.

The results were the distribution of the top-level domains (TLDs), the first level domains and the whole domain name web sites. The most highly cited web TLDs are .edu, .com, .gov and .org – all registered in the USA (Table 6). Other countries whose TLDs are frequently cited are UK (.uk), Switzerland (.ch), Canada (.ca), Germany (.de), Australia (.au), Sweden (.se) and Netherlands (.nl). Cui applied the citation analysis and determined the 78 most highly cited web sites out of thousands of cited links. He concludes that the method supports librarians and information scientists in evaluating web sites.

**Table 6: The URLs and the frequency distribution of the "First Level Domains" (Cui, 1999)**

Rank	URLs	Frequency	Percent	Cum Percent
1	<a href="http://www.yahoo.com">http://www.yahoo.com</a>	91	2.47	2.47
2	<a href="http://www.gen.emory.edu">http://www.gen.emory.edu</a>	86	2.33	4.80
3	<a href="http://www.cdc.gov">http://www.cdc.gov</a>	65	1.76	6.56

4	<a href="http://www.ama-assn.org">http://www.ama-assn.org</a>	52	1.41	7.97
5	<a href="http://www.medmatrix.org">http://www.medmatrix.org</a>	36	0.98	8.95
6	<a href="http://text.nlm.nih.gov">http://text.nlm.nih.gov</a>	33	0.89	9.84
7	<a href="http://www.nih.gov">http://www.nih.gov</a>	32	0.87	10.71

A study on pediatric web sites was performed by Hernández-Borges et al. (1999). Aim of this study was the determination of certain website characteristics as possible quality indicators for pediatric websites as well as the evaluation of the agreement of a subset of Internet rating systems editorial boards regarding their evaluations of a sample of pediatric websites. For this purpose a subset of web site ranking systems was compiled. Every web site evaluated by these rating systems that provided information about child health, whether for lay people or health professionals, was included in the study. Some of these rating systems (e.g., Lycos Top 5%) provides a search tool by keyword. In these cases, the websites were selected using the keywords "Pediatrics", "Infancy", "Child health", and "Child Care." For the remaining rating systems, the pediatric websites were compiled manually. Besides different website characteristics that depend on the users' preferences, like the number of daily visits or the updating frequency, the number of websites linked to each of the web sites of the sample was determined using the search engine Infoseek (Table 7).

The authors conclude that some website characteristics as the number of daily visits, their updating frequency and, overall, the number of websites linked to them, correlate with their evaluation by some of the largest rating systems on the Internet, what means that certain indexes obtained from the usage analysis of pediatric websites could be used as quality indicators. On the other hand, the citation analysis on the Web by the quantification of inbound links to medical websites could be an objective and feasible tool in rating great amounts of websites (Hernández-Borges et al., 1999).

Table 7: Top 10 pediatric web sites of the sample (N= 363) by the number of their inbound links. The weeks since the last update, the number of daily visits to the web sites and their editor/author's impact factor are also provided. In parenthesis, the place that each web site would obtain if ranked by the two latter criteria. In *italics*, those web sites indexed at least by two rating systems (Hernández-Borges et al., 1999).

	Uniform Resource Locator	N° of inbound links	Daily visits to web sites	Web site editor/author's impact factor	Weeks since the last update
1	<a href="http://www.merck.com">http://www.merck.com</a>	3574	-	-	13
2	<a href="http://www.ucalgary.ca/~dkbrown/index.html">http://www.ucalgary.ca/~dkbrown/index.html</a>	2366	1620 (3 <sup>o</sup> )	0 (° 60 <sup>o</sup> )	-
3	<a href="http://KidsHealth.org">http://KidsHealth.org</a>	1109	-	-	-
4	<a href="http://www.psych.med.umich.edu/web/aacap">http://www.psych.med.umich.edu/web/aacap</a>	927	-	-	3
5	<a href="http://www.aap.org">http://www.aap.org</a>	896	-	-	1
6	<a href="http://www.chadd.org">http://www.chadd.org</a>	785	-	-	4
7	<a href="http://www.castleweb.com/diabetes">http://www.castleweb.com/diabetes</a>	767	-	-	-
8	<a href="http://www.medconnect.com">http://www.medconnect.com</a>	714	-	-	-
9	<a href="http://www.aaaai.org">http://www.aaaai.org</a>	677	-	-	-
10	<a href="http://www.aacap.org/web/aacap">http://www.aacap.org/web/aacap</a>	612	-	-	4

## Maps

Larson (1996) applied the cocitation analysis for mapping web sites based on their hypertext links. The search was performed for web sites on the topics geographic information systems, earth sciences, and satellite remote sensing. The steps of analysis were:

1. Selection of the core set of items for the study.
2. Retrieval of cocitation frequency information for the core set.
3. Compilation of the raw cocitation frequency matrix.
4. Correlation analysis to convert the raw frequencies into correlation coefficients.
5. Multivariant analysis of the correlation matrix, using principle components analysis, cluster analysis or multidimensional scaling techniques.
6. Interpretation of the resulting "map" and validation.

The search was performed with the Alta Vista search engine. The sample he obtained contained 125 web sites. He calculated the co-citation matrix, determined the co-occurencies and used multidimensional scaling for mapping of the results.

The mappings produced by cocitation analysis of web sites in geographic information systems, earth sciences, and satellite remote sensing seem to produce quite clear, reasonable and interpretable results (figure 5).

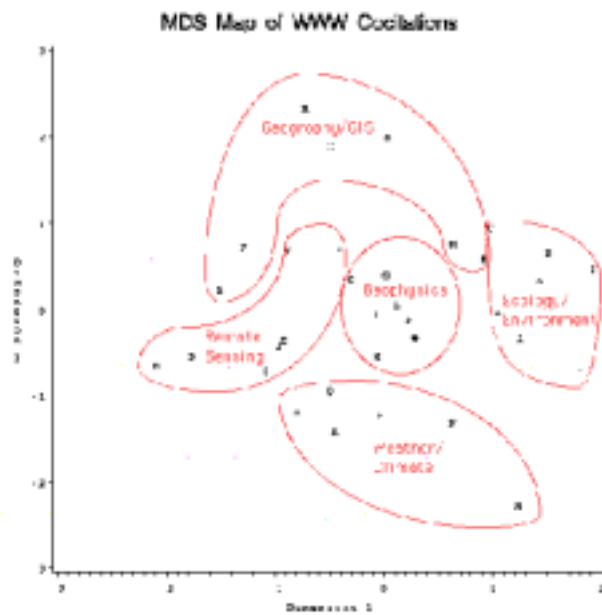


Figure 5: MDS map of web cocitations (Larson, 1996)

The discussions about the application of bibliometric analysis of the web point out that several aspects have to be considered calculating the web impact factor WIF or others:

- While traditionally bibliometric analysis is based on articles published in journals the web does not mainly consist of articles but of web pages with different intention and scientific background.
- Some studies focused on e-journals and scholarly communication in order to obtain the impact of electronic publications on research in comparison to printed ones (Harter and Kim, 1996; Harter, 1996; Fosmire and Yu, 2000). The results have shown that the importance of e-journals for knowledge production has in some research fields increased significantly in the last years.
- Bibliometric analysis is mainly performed using search engines for determination of numbers of hosts, links from one domain or web page to others, domains or text corresponding to a specific topic. The search engine used in most cases is Alta Vista because it allows Boolean searches for host, links, domains and combinations of them (Leydesdorff and Curran, 2000; Boudourides et al., 1999)
- The studies have shown that the search results using search engines have to be handled with great care because the numbers obtained depend on the search string. Using several combinations of the same Boolean operators do not provide the same results. This fact has to be considered for bibliometric web analysis. Inconsistencies have been determined depending on the time of the search and the combination of operators (Smith, 1999, 1 and 2) or due to search engine results duplicates (Ingwersen and Christensen, 1997; Harter and Kim, 1996).
- Indicators determined for webometrics are up to now mainly the impact factor and the immediacy factor. Furthermore the following information can be determined (Leydesdorff and Curran, 2000; Turnbull, 2000):
  - Amount of web-sites
  - Consideration of the languages used
  - Searching for specific hyperlinks between .com, .edu and .gov sites (the generic Top Level Domains gTLDs)
  - Analysis of time dependence

- Analysis of the links between pages
- Search for words / co-words using the full text respectively the titles
- bibliographic coupling: measuring the number of references two papers have in common to test for similarity. He then showed that a clustering based on this measure yields meaningful groupings of papers for research (and information retrieval) by stating "a number of papers bear a meaningful relation to each other when they have one or more references in common" (Kessler, 1963).
- Cocitation analysis: if two references are cited together, in a latter literature, the two references are themselves related. The major refinement between bibliometric coupling and cocitation is that while coupling measures the relationship between source documents, cocitation measures the relations between cited documents.
- An additional fundamental difference between print and web based bibliometric analysis is the possibility of an almost continuous change of contents on the web (Glänzel, 2001). Therefore the citation frequency can increase or decrease at any time. In webometrics special birth-and-death processes can be applied to describe changes of citation frequencies over time. This is on one hand based on the fact that publications can be developed on the web due to input from other scientists or readers. On the other hand this can also lead to different or wrong search results because the web sites may no longer exist or the content could have changed significantly (Smith, 1999, 1).
- For comparison of results of different studies the different cultures and languages have to be taken into account. Different studies have shown that the WIFs or other indicators show a strong dependence on the language and the country.

## 5.2 Using Artificial Neural Networks for Clustering and Mapping of Science

(contribution of Xavier Polanco, INIST)

The artificial neural networks (ANNs) hold that intelligence arises in systems of simple, interacting components (biological or artificial neurons) through a process of learning or adaptation by which the connections among components are adjusted. Processing in these systems is distributed across collections or layers of neurons. Problem solving is parallel in the sense that all the neurons within the network and layers process their inputs simultaneously and independently.

In ANNs systems, processing is parallel and distributed with no manipulation of symbols as symbols. In a domain, patterns are encoded as numerical vectors. The connection among components, or neurons, are also represented by numerical values. The transformation of patterns is the result of numerical operations, usually, matrix applications. The "designer choices" for a connectionist architecture constitute the "inductive bias" of the system.

### Artificial neurons and networks properties

The basis of neural networks is the artificial neuron. An artificial neuron consists of (1) *inputs signals*, (2) *a set of real valued weights*, (3) *an activation level*, and (4) *a threshold function*. Table 1 describes these properties. In addition to these properties of individual neurons, a neural network is also characterized by global properties such as: (1) *the network topology*, (2) *the learning algorithm used*, and (3) *the encoding scheme*. Table 2 describes these global properties characterizing a neural network.

**Table 1:** *Mathematical characteristics of an artificial neuron or node*

Input signals, $x_i$	Typically inputs are discrete, from the set $\{0, 1\}$ or $\{-1, 1\}$ , or real numbers. These data may come from the environment, or the activation of other neurons
A set of real valued weights, $w_i$	The weights are used to describe connection strengths, and the strengths of bias links
An activation level $\sum w_i x_i$	The activation level is determined by the cumulative strength of its input signals where each input signal is scaled by the connection weight $w_i$ along that input line. The activation level is thus computed by taking the sum of the scaled inputs, that is, $\sum w_i x_i$
A threshold function, $f$	This function computes the neuron's final or output state by determining how far the neuron's activation level is below or above some threshold value. The threshold function is intended to produced the on/off state of actual neurons

Source: Luger and Stubblefield (1999), chapter 14 "Machine Learning: Connectionist," p. 661-712.

**Table 2:** *Global properties characterizing neural networks*

	The topology of the network is the pattern of connections
--	-----------------------------------------------------------

The network topology	between the individual neurons. This topology is a primary source of the nets inductive bias
The learning algorithm used	There is a number of algorithms for learning such as: back-propagation learning algorithm, competitive learning algorithm, Hebbian coincidence learning algorithm. And these algorithms can be used in two main modes: supervised and unsupervised learning modes
The encoding scheme	This includes the interpretation placed on the data to the network and the results of its processing

Source: Luger and Stubblefield (1999), chapter 14 "Machine Learning: Connectionist," p. 661-712.

### Reasons for using ANNs in metrics studies of science

- **Non-linear multivariate data analysis**

The scientometric interest in ANNs may be based on the links that exist between multivariate data analysis and the ANNs approaches in the areas of clustering and cartography. The non-linear capabilities and either the supervised or unsupervised learning algorithms that the ANNs represent for clustering and mapping also motivated this interest.

The main techniques of clustering in data analysis are supervised and unsupervised techniques. A clustering technique is called supervised if one compares the unknown pattern  $x$  with all known reference patterns  $y$  on the basis of some criterion. The problem is called unsupervised clustering if one assumes that we do not know the clusters a priori. Nevertheless the data (or the samples) fall in a finite set of categories according to their similarity relations. The term "classification" may be reserved to signify only the supervised technique, which is known also in data analysis as "discriminant analysis" (Lebart et al., 1995; McLachlan, 1992), and the term clustering may be used to signify the unsupervised technique, usually called "cluster analysis."

- **Learning capability**

The power of ANNs is derived from their learning capability defined as a change in the weight matrix ( $W$ ), which represents the strength of the links among nodes. In the self-organizing maps (Kohonen, 1997), the learning is competitive and unsupervised.

Kohonen winner-take-all learning is unsupervised. The winner-take-all learning algorithm (Kohonen, 1984) works with the single node in a layer of nodes that responds most strongly to the input pattern. Winner-take-all may be viewed as a competition among a set of network nodes. The selected node is the node whose pattern of weights is most like the input vector  $X = (x_1, x_2, \dots, x_m)$ , and adjusts it to make it more like the input vector. Learning for winner-take-all is unsupervised in that the winner is determined by a "maximum activation" test. The weight vector of the winner is then rewarded by bringing its components closer to those of the input vector. For the weights,  $W$ , of the winning node and components  $X$  of the input vector, the increment is:  $\Delta W_t = c(X^{t-1} - W^{t-1})$ , where  $c$  is small positive learning constant that usually decreases as the learning proceeds. The winning weight vector is then adjusted by adding  $\Delta W_t$ .

This reward increments or decrements each component of the winner's weight vector by a fraction of the  $x_i - w_i$  difference. The effect is to make the winning node match more closely the input vector. The winner-take-all algorithm does not need to directly compute activation levels to find the node with the strongest response. The activation level of a node is directly related to the closeness of its weight vector to the input vector. For a node  $i$  with a normalized weight vector  $W_i$ ,

the activation level  $W_i X$ , is a function of the Euclidean distance, with normalized  $\|W_i; X - W_i\| = \sqrt{(X - W_i)^2} = \sqrt{X^2 - 2XW_i + 1}$

From this equation, for a set of normalized weight vectors, the weight vector with smallest Euclidean distance,  $\|W_i; X - W_i\|$ , will be the weight vector with maximum activation value,  $W_i X$ . In many cases it is more efficient to determine the winner by calculating Euclidean distances rather than comparing activation levels on normalized weight vectors.

R. Hecht-Nielsen (1990) shows how "winner-take-all" algorithms may be seen as equivalent to the k-means analysis of a set of data. In other words, Kohonen unsupervised clustering of data is basically the same as k-means analysis. On this subject, see also J. M. Zurada (1992).

- **Spatial order and organization**

Another important property of the ANN clustering method discussed in this section is the spatial order and organization in the representation of data with all their interrelationships.

The self-organizing map (SOM) gives central attention to spatial order in the clustering of data following a neighborhood approach. The purpose is to compress information by forming reduced representations of the most relevant features, without loss of information about their interrelationships. Let  $X \in \mathfrak{R}^n$  be a stochastic data vector. We might then say that the SOM is a non-linear projection of the probability density function  $p(X)$  of high-dimensional input data vector  $X$  onto the two-dimensional display,  $\mathfrak{R}^2$ . The smallest of the Euclidean distances  $\|X - W_i\|$  can be used to define the best-matching node or neuron.

During learning, that is, the process in which the non-linear projection is formed, those nodes that are topographically close in the array up to a certain geometric distance will activate each other to learn something from the same input  $X$ . The result is a *local* relaxation or smoothing effect on the weight vectors  $W$  of neurons in this neighborhood, which in continued learning leads to *global* ordering.

## Using ANNs in Scientometrics

The use of ANNs in Scientometrics is recent. One of the first studies of the application of ANNs to matrix keywords  $\times$  documents have resulted in the definition of the Axial-K-Means (AKM) method for cluster analysis. It is inspired from the neural formalism of Kohonen model. AKM applies a modified version of Oja's winner-take-all learning rule (Lelu, 1993). This method was implemented in clustering program called NEURODOC (Lelu and François, 1992a; 1992b), which is used by the URI at the INIST-CNRS and now integrated in STANALYST ®.

Campanario (1995) used Kohonen SOM for mapping scientific journal-to-journal citation data. White, Li, and McCain (1998) compared multidimensional scaling (MDS) versus Kohonen self-organizing map (SOM) for mapping author co-citation data. They showed that, given co-citation data, Kohonen SOM produces results quite similar to those of MDS.

Polanco and co-workers tested some ANNs such as ART1 for clustering, and the Multiplayer Perceptron (MLP) and also the Kohonen self-organizing map (SOM) for cartography (Polanco et al., 1998a). Then they compared and evaluated mathematically the MPL in self-association mode with the standard principal component analysis (PCA) for graphical representation purposes. Such a network as they showed performs a non-linear principal component analysis (Polanco et al., 1998b; 1998c). Finally, they used a non-linear MLP with two hidden layers for mapping the clusters generated by an axial k-means (AKM) algorithm (Polanco and François 2000a).

The final result of this work is a combined two steps computer-based system: firstly the data are clustered by the AKM algorithm, then the MLP maps the clusters and the cluster relationships are designed by means a related component analysis (RCA). The RCA is based on graph theory and defines the related components which represent the relative closeness between clusters on the map surface.

Now, Polanco and his co-workers are turning to the Kohonen self-organizing map (SOM) in order to do in only one step the tasks of clustering and mapping a data set and applying a hypertext multi-map approach (Polanco, François, and Lamirel, 2000b).

## Using ANNs in the Cyberspace

*In Finland, Kohonen and colleagues mapped a set of Usenet news group and placed the results on the Web (WEBSOM). The newest version WEBSOM2 is described in Kohonen et al. (1999, p. 171-182). It is a massive text document collection self-organizing map display. The main problem addressed is the rapid construction of a large document map. The map is used for information discovery. The document map is presented as a series of HTML pages that enable exploration of the grid points. When clicking the grid points, links to the document database enable reading the content of the articles. If the grid is large, subsets of it can be viewed by zooming. An automatic method assigning descriptive signposts to maps regions; in deeper zooming, more signs appear. The signposts are words that appear often in the articles in that map region and rarely elsewhere, and they are used to monitor the search. Available WWW: <http://websom.hut.fi/websom>*

In the EICSTES project, and related to WP 9, we are dealing with an innovation that was firstly introduced for the information retrieval purposes (Lamirel, 1995; Lamirel et al., 2000). It is the multi-map extension of the Kohonen SOM algorithm. This will be from now signified by the name of Multi-SOM. Moreover, the Multi-SOM technique introduces the concept of viewpoints into the domain analysis with its multi-maps displays. The area computation algorithm represents a generalization of the Lin algorithm (Lin et al., 1991) as Lamirel (1995, p. 278-282) presented it. With regard to standard SOM, Multi-SOM introduces a generalization mechanism, and inter-map communication mechanism. An interesting functionality for domain analysis is summarizing the map content into more generic clusters through an on-line generalization process. The communication among self-organizing map that has been first introduced in the context of an information retrieval model (Lamirel, 1995), represents a major amelioration of the basic Kohonen SOM model.

In Polanco, François, and Lamirel (2001) is exposed in detail the Multi-SOM system, and examined the issue of its application. The main SOM improvements introduced by Multi-SOM are recalled (1) the way of the clusters are labelled, (2) the division of the map into logical areas, and (3) the generalization mechanism. Then the authors presented the inter-map communication mechanism as the major innovation. They emphasised the use of viewpoints, each different viewpoint being achieved in the form of a particular map linked to the other viewpoints by the on-line inter-map communication mechanism. A real application is detailed and a method of analysis is also exposed.

## Knowledge indicators

An important reason to use ANNs in quantitative studies of science and technology is their capability to create "higher abstractions from raw data completely automatically. Intelligence in

neural networks ensues from abstractions, not from heuristic rules or manual logic programming" (says Kohonen, 1997, p. 65).

In relation with this remark Polanco, François, and Lamirel (2001) refer to the notion of knowledge indicators. In 1978, there was a "science indicator" versus "knowledge indicator" discussion about the first indicator report edited by the NSF (see Elkana et al., 1978). Polanco, François, and Lamirel (2001) call "knowledge indicators" the following three elements: *keywords*, *clusters*, and *maps*. The keywords or index terms are the indicators of the knowledge content in the indexing documents. The clusters of keywords mean themes in which certain domain knowledge can be featured. The maps of clusters are then considered thematic maps. The maps represent strategic indicators because they provide a comparison way for evaluating the relative position of themes onto an ordered space. This ordered space signifies a space of knowledge defined by the set of clusters and relationships among them.

Polanco, François, and Lamirel (2001) argue that keywords or index terms are by their own nature symbolic objects. The occurrences of that symbols are computed in the statistical analysis. Instead the clusters and maps are both metric and symbolic objects. Some metrics not only computes but also built them and provides a set of symbolic objects having at the same time very useful mathematical properties for their analysis and justification. These are symbolic objects and vehicle meaning or semantic properties, which can be decoded in some signification code. They can be decoded as indicators. Indicators can be defined as a "set of conventions" for describing the knowledge conveyed by documents. In artificial intelligence (Winston, 1977; Lugger and Stubblefield, 1999), this issue is studied under the name of knowledge representation.

## 5.3 Scientometrics and Network analysis

(contribution of Xavier Polanco INIST)

The graph + network paradigm has been evoked at their beginning by both traditions, citation analysis and co-word analysis, but the map as visual and analytical structure became the dominating paradigm.

In "Network of Scientific Papers," Price (1965) represented citations between articles as filled cells in a citing-cited matrix. The citation network is a directed graph, whose vertices can be chronologically ordered, and whose edges connect earlier with later vertices. Callon, Law, and Rip (1986) exposed the co-word analysis approach in terms of "actor-network" sociology. In his "technical issues," Courtial (chap. 11, in Callon, Law, and Rip, 1986) explicitly explain co-word analysis in terms of graphs and networks. In spite of this, the map will remain the central paradigm as structure of analysis and visualization. The idea of using graphs and networks for analysing and understanding the so-called "actor-network" remained undeveloped. In quantitative studies of science (Van Raan, 1987), the mathematical resources of the social network analysis are also absolutely unknown.

The network word might be largely used in social studies of science. One example is Latour (1987, *Science in action*, Part III From short to longer networks, and his *Irreductions*, in *The Pasteurization of France*, 1988). His sociology of science unknowns the social network analysis. The network word delays in his books a qualitative concept.

A very complete academic bibliography about social networks can be found in <http://www.socialnetworks.org> Freeman (2000) gives a good overview about software tools for constructing visual displays of social networks. Degenne and Forsé (1994/2001) have presented the social network in terms of structural analysis in sociology

The Internet and Web diffusion contribute to expand the idea of understanding through graphs and networks. Now the idea becomes topicality. Among many authors, see for example "Graph structure in the Web" by Broder et al. (2001), also Barabasi (2001); Chakrabarti et al. (1999).

Translating the co-word analysis into co-site analysis, Polanco et al. (2001) have proposed the concepts of density and centrality as indicators following the co-word analysis tradition. Polanco (2001) recently says that these concepts should be developed in terms of graphs and network measurements. Clusters of sites as well as clusters of words will be henceforward analyzed and represented according to graph theory and network analysis.

### Galois lattice or conceptual clustering

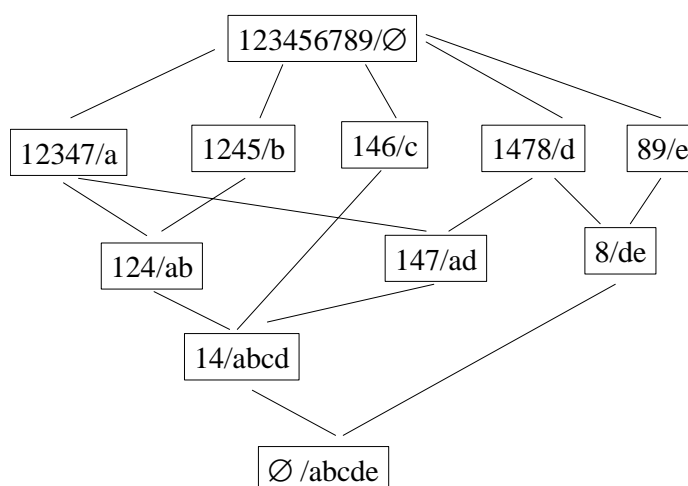
It is interesting to be able to make qualitative studies of resemblance between individuals or objects. This is true also for the analysis of the social networks. Let us suppose for example that one knows that a set of Web sites is indexed by a certain number of metadata. The networks is then described like a family of elements of a set X, i.e. a hypergraph (Berge, 1987). In this hypergraph, the vertex are the elements of the set X and the edges are subsets. The structure is the Galois lattice sometimes called lattice of concepts or conceptual clustering.

Let us take as example a set of 9 sites {1,2,3,4,5,6,7,8,9}, described by 5 metadata {a, b, c, d, e}. The table below presents the membership of associations between the sites at these metadata.

	a	b	c	d	e
1	1	1	1	1	0
2	1	1	0	0	0
3	1	0	0	0	0
4	1	1	1	1	0
5	0	1	0	0	0
6	0	0	1	0	0
7	1	0	0	1	0
8	0	0	0	1	1
9	0	0	0	0	1

Each column carries the symbol of a metadata. We will indicate in "comprehension" the set of two metadata  $a$  and  $d$  by word  $ad$ . The metadata  $a$  can also be defined in "extension", i.e. in the form of a list of its elements:  $\{1,2,3,4,7\}$ ; the same for  $d$   $\{1,4,7,8\}$ . If we are interested in what is common to two metadata  $a$  and  $d$ , we find the intersection  $\{1,2,3,4,7\} \cap \{1,4,7,8\} = \{1,4,7\}$ .

If we reorganize the table placing the columns side to side, and the rows the ones following the others, we would reveal a block filled with 1. Moreover, this block is maximum, i.e. if we add a line or a column to him, it ceases being entirely filled of 1. The graph below represents all the maximum blocks filled with 1 which it is possible to create with the table. Each vertex is a block. It is indicated in comprehension (right part of its name) and in extension (left part). To the bottom, two blocks combine by making the union ( $\cup$ ) of the right part and the intersection ( $\cap$ ) of the left part. Upwards, it is the reverse.



This graph bears the name of "Galois lattice" (Barbut and Monjardet, 1970; Duquenne, 1992). It is another representation of the starting table which can be rebuilt exactly from him. It also highlights the implications ( $\sqsubset$ ). For example any metadata which describes 2 also describes 1 and 4. In the same way  $6 \sqsubset (1,4)$ ,  $7 \sqsubset (1,4)$ ,  $9 \sqsubset 8$ .

*This representation is not limited to the description of individuals strongly dependent between them by common characteristics, it indicates also which are these characteristics. It is thus a tool for qualitative analysis. Freeman (1992) there sees a manner of working on the contents of the social networks.*

GLAD (Duquenne, 1993) is a program for calculating and displaying Galois lattices for analysing and visualizing social networks. Like correspondence analysis, a Galois lattice is designed to simultaneously deal with two sets, the rows and columns of a data table. But the Galois lattice embodies a completely different approach. It displays an order structure, in which the dependencies among the row data, among the column data, and those between the two are simultaneously revealed.

## **Semantic networks**

If we are interested by knowledge indicators, in the sense of empirically representing knowledge that is content in the documents, we need to explore other networks, these are known as *semantic networks* and *concept graphs*. As Sowa (1991, p. ix) says: "Graphic notation for knowledge have been used for centuries in logic, philosophy, psychology, and linguistics. In the 1960s, the early days of artificial intelligence, network notations were among the first knowledge representation schemes to be developed." A semantic network is a structure for representing knowledge as a pattern of interconnected nodes and arcs. Nodes represent concepts of entities, attributes, events, and states. Arcs usually called conceptual relations, represent relationships that hold between the concepts nodes. Labels on the arcs specify the relation types. Conceptual graphs are based on a graph notation for representing natural language semantics. Sowa (1991) argued that a graph logic, such as C. S. Peirce's existential graphs, can represent linguistic structures, which supporting knowledge, more faithfully than the predicate calculus. He combines Peirce's graph with representations from artificial intelligence and linguistics to form his version of conceptual graphs.

Semantic networks and social networks may be related via the Galois lattice or conceptual clustering when the objective is to perform content analysis.

## 5.4 Social Network Analysis, Chaos Theory and Complex Networks

(contribution of Moses A. Boudourides University of Patras)

### Social Network Analysis

A social network is a set of actors and relations occurring among them. Actors can be individual people, objects or events as far as certain relations hold them all together; actors can be also aggregate units such as organizations, institutions, communities, groups, families etc. The very idea of the social network approach is that relations or interactions between actors are the building blocks or the key factors that sustain and define the network (Wellman, 1988; Wasserman & Faust, 1994). Typically interactions between actors result from exchange of resources, either material or informational, such as goods, money, information, services, social or emotional support, trust, influence etc. Each kind of resource exchange is considered a social network relation and actors maintaining the relation are said to maintain a tie. The strength of a tie may range from weak to strong depending on the quantity, quality and frequency of the exchanges between actors (Marsden & Campbell, 1984). Patterns of who is tied to whom reveal the structure of the underlying network: they show how resources flow among actors and how actors are interconnected in the network. A few very well known examples of social network analyses are: Granovetter (1973, 1974) who investigated exchange of job information among acquaintances and found that weak ties are quite operationally strong for the diffusion of such information. Wilson (1997) found that the urban poor in isolated Black ghettos lack connections with sources of work. Burt (1992) studied the dependency of social capital on 'structural holes' (which are particular kinds of network positioning in which a focal actor is connected to other actors which themselves are not connected with one another); thus, according to Burt, social capital is not a direct attribute of actors but rather of their ability to sustain flexible configurations within a network.

Now, computer networks in general and in particular the Internet are clearly social networks (Wellman *et al.*, 1996). In these social networks, actors may be human, such as users, communicants, information producers and consumers, citizens, public or market organizations etc., or non-human, such as computer machines, information databases, (hyper-) documents, multimedia resources etc. Relations among the human Internet actors refer to informative and communicative uses, access, provision, procurement, commerce, work, education etc. Although human actors are always beneath the non-human ones, typical relations among the latter consist of information (data) flows, traffic, exchanges of e-mails and postings in web pages, links, connections, network topologies etc. We should remark that some have been attracted by the idea that in a complex heterogeneous translation network both humans and machines can perform agency. This is the dogma of the 'Actor Network Theory,' in which the development and stabilization of scientific and technological objects (facts and artefacts) results from the construction of heterogeneous networks as concrete alignments between human actors, natural phenomena and social or technical intervening aspects (Callon, 1986; Latour, 1987).

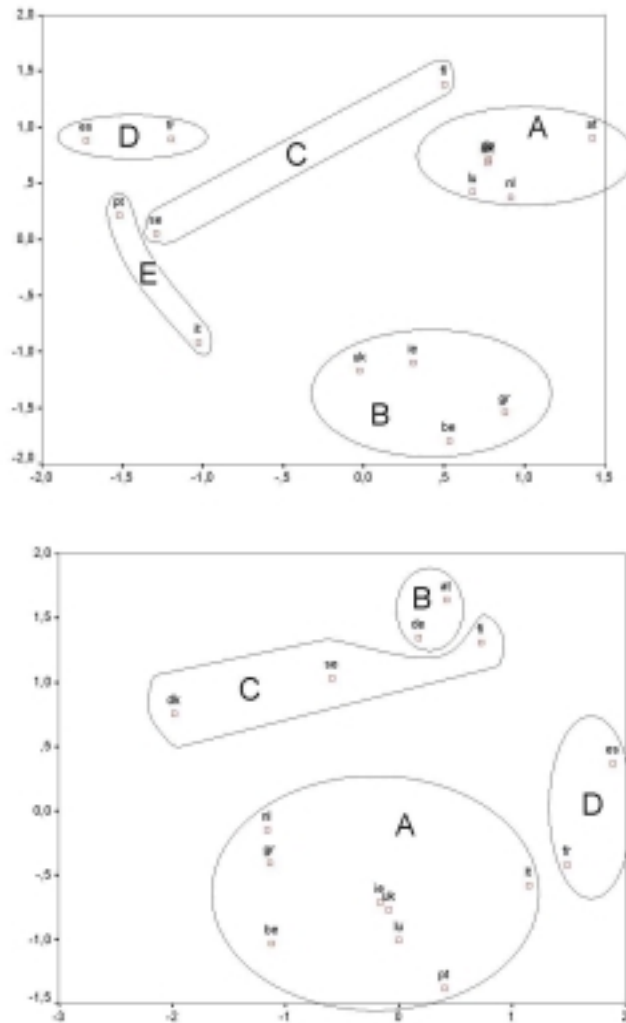
In the methodology of the social network analysis approach, the commonly addressed basic issues are five: cohesion, structural equivalence, prominence, range and brokerage (Haythornthwaite, 1996, pp. 330-336).

- o *Cohesion* (which is a *relational* property) refers to the grouping of actors because of the strength of their relationships with one another (Wasserman & Faust, 1994, pp. 249-290). Cohesive groups of actors form *clusters* or *cliques* depending on whether they are highly or fully interconnected, respectively. A measure of cohesion is the network *density*, which is

calculated as the ratio of the number of actually occurring links (relations) to the number of all possible links (1994, pp. 101-103). A relevant concept is that of *centralization*, measuring the extent to which a set of actors are organized around a central one (1994, pp. 175-177).

- o *Structural equivalence* (which is a *positional* property) identifies actors who have similar patterns of relations with others, even if such actors may not have direct relations with each other (Wasserman & Faust, 1994, pp. 347-424). An actor's pattern of relations constitutes a *role*. Thus, actors playing similar roles occupy similar (or equivalent) structural or status positions. A technique for assessing structural equivalence is known as *block modelling*. In this technique, one first calculates correlations between all pairs of actors and then reorders the actors into sets on the basis of the correlation values in such a way that pairs of actors that are highly correlated (and, therefore, most structurally equivalent) should appear together in the same group (or block).
- o *Prominence* reflects the hierarchical status of an actor (Wasserman & Faust, 1994, pp. 169-174). It can be measured by assessing the *centrality* of an actor in a network, which is derived by measuring the actor's connections in the network, i.e., its *degree*. (This differs from the previously mentioned centralization, which measures the configuration of the network as a whole.) The actor with the highest degree (i.e., the most relationships with other actors) is the most central. Another measure of an actor's prominence is *global centrality* (or *closeness*) and it is derived by measuring the *distance* between this actor and any other actor, which is defined as the number of connections in the shortest path between the actors (Wasserman & Faust, 1994, pp. 184-186). The actor with the lowest sum of distances to all other actors is the most globally central actor.
- o *Range* refers to a combination of network size and heterogeneity that jointly increases the ability of actors to have access to a variety of resources (social support, social capital) (Wellman & Potter, 1999, pp. 65-67). The bigger a network is, the more information an actor will have access to and the more complex the accessed information will be. Moreover, heterogeneous networks may provide a greater variety of social support.
- o *Brokerage* activity puts interested actors in touch with one another so that they might strike a deal (Knoke, 1990, pp. 144-146). It involves at least three actors with the intermediary relegating transactions between the others. According to Gould and Fernandez (1989), there are five ideal-typical roles of brokers: liaison, representative, gatekeeper, itinerant and coordinator. Brokerage can be measured by *betweenness*, the extent to which an actor is located between others in the network (Wasserman & Faust, 1994, pp. 189-191). Where opportunities of brokerage exist but have not been exploited yet, there is a 'structural hole,' in the terminology coined by Burt (1992). However, brokerage indicates not only opportunities to further exploit the network potentialities but also points of possible resistance by those currently playing the gatekeeper's role who have the power to control and filter imported or exported information.

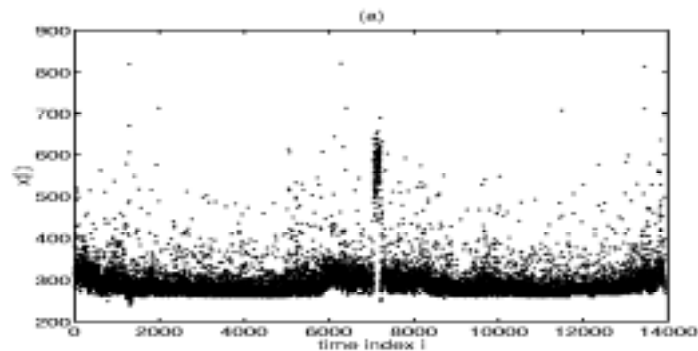
All the above was a general presentation of the theory of social network analysis. As an example of an application of social networks to webometrics, let us consider the links among almost all of the University sites in the 15 EU countries. (These were data collected by our agents and posted at our web-data page <<http://nicomedia.math.upatras.gr/eicstes/webdata.html>>.) Then, the following two sociograms show the relational positioning of the EU countries according to the incoming and out-going links of their University sites, respectively (obtained by MDS while the clustering in groups of countries has resulted by SPSS factor analysis):



## Chaos Theory and the Internet

Some indicators of chaotic and fractal structures have already been applied to webometrics: This is Egghe's (1997) study of the fractal dimension of hypertext systems, which has been calculated in terms of the average number of links per page. But furthermore, chaotic analysis can be applied to time-series of data obtained from the Internet, such as various network traffic data and web cache data. As an example, we will consider the former data and in particular the so-called *ping times data*. As a matter of fact, one way to estimate the traffic on the Internet is by measuring the round-trip times for ICMP *ping* packets sent from one host to a second one. In this way, one obtains a time series (herein called 'ping time series'), which one could process through linear statistical methods and methods of chaotic non-linear analysis. Thus, one may investigate whether certain patterns emerge, i.e., there exists some kind of structure in the data, or the data are completely random, such as, e.g., the outcome of a random number generator. Certainly, one expects that, if any, the underlying mechanism is a complicated one that cannot be identified using the classical statistical tools, such as autocorrelation and Fourier spectrum analyses. (General literature on the field: Kay, 1988, Priestley, 1988, Tsonis, 1992.)

In a previous work of ours (Kugiumtzis & Boudourides, 1998), we had examined the time series composed of ping times in between two hosts, one at the Democritus University of Thrace in Greece and the other at the University of California at Irvine in USA (data collected over a period of a month in 1998).



This ping time series (measured at the scale of 15 sec) was following a repeated pattern, composed of a background, corresponding to a subsided Internet traffic, and succeeded by a sudden peak, corresponding to a short congestion period of the order of a minute or less. In the measured set of round trip times, a data window of larger magnitude was detected corresponding to a large congestion period of more than half an hour, during which the net was overloaded. The presence of this data window introduced nonstationarity and, therefore, it altered the results of some of the applied methods. Particularly, the inclusion of the long congestion period resulted substantially larger estimated linear correlations over many lags, which was apparently a misleading result. Nevertheless, removing the long congestion period, no correlation was detected. Moreover, the False Nearest Neighbors method suggested erroneously that there was more structure in the long congestion period data.

The findings of the applied methods have not given any definite evidence whether the ping data were simply white noise or not. Obviously, the data was following a nonsymmetrical distribution resembling the lognormal distribution and, so, the data was not white noise of a Gaussian or uniform type. Some methods, such as the estimation of the autocorrelation, mutual information and correlation dimension, suggested that there was no correlation and structure in the ping data. On the other hand, the prediction with Autoregressive models and the Local Linear Prediction, as well as the Largest Lyapunov Exponent method, indicated that there were small correlations and some structure in the data, as they have given different results than those expected for the white noise case.

## Complex Networks on the Internet

A wide range of systems in nature and society are described as complex networks: examples are abundant and range from the biological cell to the technological Internet. Furthermore, what is amazing is that many of these complex networks are seen to be governed by certain robust organizing principles of the statistical mechanics of network topology and dynamics (Albert & Barabási, 2001). As a matter of fact, researchers have concentrated particularly on a few properties, which seem to be common in many complex networks, such as clustering (or network transitivity), the 'small-world' property and power-law degree distributions. In the sequel, let us briefly discuss these properties.

**Clustering:** This property (also called *network transitivity*) is a common property of social networks, in which often cliques are formed as fully connected subgraphs. In other words, clustering is manifested by the tendency that any two nodes, which are both linked with (adjacent to) the same third node, tend with an increasing probability to be linked together themselves. In the language of social networks, two friends of a person will have a greater probability of knowing one another than any two persons randomly chosen from the population.

This property is quantified by the clustering coefficient (Watts & Strogatz, 1998; Newman, Strogatz & Watts, 2001) as follows: Fix a node  $u$  and consider the set  $A_u$  of all nodes linked to  $u$ ; let  $C_u$  be the ratio of the number of all the existing links among the nodes of  $A_u$  divided by the number of all possible links among these nodes; then *the clustering coefficient*  $C$  of a graph  $G$  is defined as the average over all nodes, i.e.,  $C = N^{-1} \sum_{u \in V} C_u$ .

**Small Worlds:** The ‘small-world’ property of a complex network occurs when, although such a network might be of large size, there is a relatively short path between any two nodes and also the network is quite highly clustered. Note that the distance between two nodes is defined as the number of links along the shortest path connecting them. More formally: Fix two nodes  $u$  and  $w$  and consider the shortest path length  $d(u,w)$  between these nodes, i.e., the minimum number of links that must be traversed in order to reach node  $w$  starting from node  $u$ ; the average path length of the node  $u$  is defined as  $d(u) = N^{-1} \sum_{w \in V} d(u,w)$  and the *average path length*  $d$  of the graph  $G$  is defined as  $d = N^{-1} \sum_{u \in V} d(u)$ .

However, there is a category of graphs for which the average path length is relatively small. These are the *random graphs*, defined by Erdős & Rényi (1959) as follows: Start with  $N$  nodes and connect every pair of them with probability  $p$ , creating a graph with approximately  $pN(N-1)/2$  links distributed randomly. In general, in a graph  $G$  of  $N$  nodes, the *degree*  $k_u$  of a node  $u$  is the number of all the links from  $u$  to all its adjacent nodes and the average degree of the graph  $G$  is  $\langle k \rangle = N^{-1} \sum_{u \in V} k_u$ . Now, for a random graph, one easily computes that the clustering coefficient is  $C^{rand} \approx \langle k \rangle / N$  and the average path length is  $d^{rand} \approx \ln N / \ln \langle k \rangle$ .

After this preparation, we are in the position to define ‘small-worldness’ formally according to Watts & Strogatz (1998): A graph is said to have the *small world property* if:

- its average path length  $d$  is close to  $d^{rand}$  but
- its clustering coefficient  $C$  is much larger than  $C^{rand}$ .

Perhaps the most popular manifestation of the ‘small-world’ property is the concept of the “six degrees of separation” due to the social psychologist Stanley Milgram (1967), who concluded that there was a path of acquaintances with average length about six between any two people in the United States (Kochen, 1989). Similarly, the actors in Hollywood have been found to be on average distance of three from each other (Watts & Strogatz, 1998), the chemicals in a cell are separated on average by three reactions (Wagner & Fell, 2000), web pages are on average nineteen clicks away (Albert, Jeong & Barabási, 1999) and so on.

**Degree Distribution:** We have already seen that in a graph the *degree* of a node is the number of all the links from this node to all its adjacent nodes. Since not all nodes in a network have the same number of links (unless the network is regular lattice), it is important to be able to study the spread in node degrees. The latter is characterized by a distribution function  $P(k)$ , which gives the probability that that a randomly selected node has exactly  $k$  links, i.e., its degree is exactly  $k$ .

If the network is a random graph, since its links are placed randomly, the majority of nodes have almost the same degree, close to the average degree  $\langle k \rangle$  of the network. Therefore the degree distribution of a random graph is a Poisson distribution with a peak at  $P(\langle k \rangle)$ . However, for many complex networks, which have been empirically investigated recently, it was found that the

degree distribution significantly deviates from a Poisson distribution expected for a random graph. In fact, it has been found that the degree distribution of many complex networks has a power-law tail of the form  $P(k) \sim k^{-\gamma}$ . After Barabási & Albert (1999), such networks are named *scale-free*. Examples of scale-free complex networks include, among others, the Web (Albert, Jeong & Barabási, 1999), as we are going to see next, Internet (Faloutsos, *et al.*, 1999), metabolic networks (Jeong *et al.*, 2000), etc.

Since the World-Wide Web at the level of web pages is a directed graph, the Web is characterized by two degree distributions:

- the distribution of out-going links,  $P_{out}(k)$ , which signifies the probability that a web page has  $k$  out-going links and
- the distribution of in-coming links,  $P_{in}(k)$ , which is the probability that  $k$  links point to a certain web page.

It has been empirically manifested that both  $P_{out}(k)$  and  $P_{in}(k)$  have power law tails of the forms  $P_{out}(k) \sim k^{-\gamma_{out}}$  and  $P_{in}(k) \sim k^{-\gamma_{in}}$ . Albert, Jeong & Barabási (1999) have found for a sample of the Web consisting of 325729 pages that  $\gamma_{out} = 2.45$  and  $\gamma_{in} = 2.1$ . Kumar *et al.* (1999), using a 40 million web pages crawl by Alexa Inc., found  $\gamma_{out} = 2.38$  and  $\gamma_{in} = 2.1$ . Broder *et al.* (2000), using two Alta Vista crawls containing in total 200 millions web pages, found  $\gamma_{out} = 2.72$  and  $\gamma_{in} = 2.1$ . Furthermore, for the Web at the level of sites (as a valued undirected graph), Adamic & Huberman (2000) found that  $\gamma = 1.94$ . In a simulation of the Web that we are presenting at Appendix I of the EICSTES Deliverable D8.1, we found that  $\gamma_{out} = 2.3637$  and  $\gamma_{in} = 2.3641$ .

## 6. Conclusions and Future Research

The most complicating factor in webometric research is that of data collection. Data collection on the web depends on the retrieval features of the various search engines and web robots. Lawrence and Giles (1998) provide a substantial contribution with respect to the commercial search engine coverage of the Web space by introducing the concept of 'indexable web'. The concept signifies the portion of the web that can be indexed by the engines. The engines do not cover the entire web, the overlap between them is not substantial and their retrieval functions too simplistic for webometric analyses.

One problem we will not have is that of *lack* of data. But besides the above-mentioned problem of information retrieval on the web, there are also problems related to methodology and theory. All this activity is taking place inside electronic networks and computers produce them automatically. But how do we make sense of data in the absence of any knowledge about the social practices underlying them? Most sociological methods of gathering information suppose the laboratory (in our case) as a site. It is likely that only sites present a suitable unit of analysis in terms of link structures. Universities and other large institutions are likely to have a very heterogeneous and un-codified linking pattern. On the other hand, individuals and small groups will have a unrepresentative linking pattern with respect to the higher level of aggregation (e.g. 'the footballclub). It seems that the new concept of a site; the presence of research groups, universities and other R&D related institutions on the Internet provides the best basis for analysis.

Scientometrics have now a whole new field of study, one that offers us the possibility to integrate different aspects of the socio-cognitive activity of science and its relationship with technology transfer. We must integrate the electronic network reality in our model for information flow among scientists, detect the links that consolidates collaborations, devise means of understanding the practices of these global scientific communities, establish relevant indicators for the socio-cognitive activity taking place on the internet and finally, embody our results into policy recommendations.

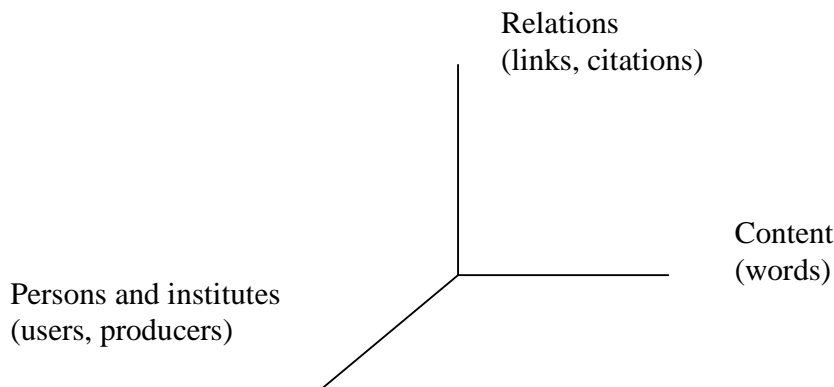
From the case studies carried out it could be concluded that that the electronic and print communication patterns can provide information about the changing nature of knowledge production. We were able to identify a heterogeneous set of relevant institutes (governmental, industrial and academic) that represent the users and producers context of mode 2 knowledge production. Interesting in this context is the presence of archives, databases, software producers and governmental institutions in the communicational environment of academic research institutes.

This raises some interesting issues on the development of scientific research and the quantitative methods of mapping this development. We argued before, that as a consequence of the information revolution in society at large, and science in particular, researchers are developing new information and communication patterns. Our thesis is that the sciences are in the midst of an informational revolution. Of course, information has always been central to scientific research, but the emergence of digital information and ICT has enabled a radical lowering of costs related to information dissemination, both in pure form and black boxed in technologies. In short, information is a resource, raw material and output in the process of knowledge production. This informational turn in the sciences coincides with the processes of commercialization and commodification of scientific knowledge, the new hybrid roles of academic research institutes and universities, the transformation of economic mechanisms by technology and innovation, a wider variety of types of research output and more attention to norms and values.

We were able to determine empirically what is the most appropriate level of analysis for mapping techno-scientific developments in the information society. It was concluded that at the level of the research group, in depth information can be obtained by linking architectures about the context of knowledge production. This level of analysis bears most resemblance with traditional print based analysis. Constructing a matrix of linking institutes seems to be comparable to journal-journal citation analysis. The communication are sufficiently codified within a scientific context to provide a meaningful overview of the context in which knowledge production takes place. It is also a sufficiently aggregated system at a level of abstraction such that one does not have to open the "black box" of each link in order to examine the information which is communicated, but that one analyzes only the number of messages.

This provides an interesting starting point for mapping University-Industry-Government Relations. Boolean search operators provided by the search engines will be used for organizing data in terms of domain names, linkages, title words, free-text words, and hyperlinks. These will be statistically analyzed to measure the Triple Helix at the Internet. In a second part, this information will be related to other resources such as the Patent databases, the Science Citation Index, and Medline data.

Scientometric methodologies rest on the assumption that traces of communicated information can be used to map the development of the science system in interaction with its (social, economical, political and technological) environment. These communications can be studied in three different dimensions. The structure of the communication networks can be mapped by quantifying the number and direction of the existing relations in the network. What is communicated can be represented (to a certain extent) by mapping title-words, key-words, etc. The role and behavior of the persons and institutes involved in the communication can be studied as a third dimension through surveys, log-files, browsing patterns etc.



**Figure 1.** The various methodological dimensions of scientometric and webometric analysis.

The actors-axis represents the dimensions of the relevant actors involved; the type of institution (governmental, industrial, academic), the geographical location (NL, UK, USA, etc), the role of these institutions in the process of knowledge production and new economy (users, producers, suppliers, commissioner).

The relations-axis represents the type (cited reference, hyperlink) and quantity of relations between actors.

The content-axis represents the 'cognitive' dimensions of communications (title-words, free-text words) and makes it possible to position the various actors in the communication network.

A number of topics will be selected that represent the various aspects of scientific-governmental-industrial interactions and their public interface. First, a list of relevant actors will be selected (based on expert opinions and relational analyses). A number of analyses will be carried out that inform us about the situation in one or more of the dimensions mentioned above. A multi-layered self-organizing maps approach will be used to combine the information. We know from previous studies that combining the various dimensions results in more reliable indicators. Possible topics that are selected include 'information science' and 'genetically modified food'.

We are also planning to carry out simulations of techno-scientific developments. We can use empirical data from scientometric and webometric sources to validate the relevant parameters in our simulation of electronic and print communications.

## 7. References

- Abraham, R.H. (1997): Webometry: measuring the complexity of the World Wide Web, *World Futures* 50 (1997): 785-791, <http://www.ralph-abraham.org/>, abgefragt am 12. 11 2001
- Adamic, L.A. & Huberman, B.A. (2000). Power-law distribution of the World Wide Web. *Science*, vol. 287, pp. 2115-2116.
- Aguillo, I. F. (1998). STM Information on the Web and the development of new Internet R&D databases and indicators. *Proc. Online 1998. Learned Information*, London
- Albert, R., & Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, vol. 74, no. 1 (in press). Available online at: <http://xxx.lanl.gov/abs/cond-mat/0106096>
- Albert, R., Jeong, H., & Barabási, A.-L. (1999). Diameter of the World-Wide Web. *Nature*, vol. 401, pp. 130-131.
- Almind, Tomas C., and Peter Ingwersen, Informetric analyses on the World Wide Web: A methodological approach to "webometrics", *Journal of Documentation*, September 1997
- Bak, P., Chen, K., 1991: Self-organized criticality. *Sci. Am.* , January, 26-33.
- Barabasi A. L. (2001) The physics of the Web, *Physics World*, 2001, vol. 14, Issue 7, available in <http://www.physicsweb.org/>
- Barabási, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, vol. 286, pp. 509-511.
- Barbut M. and Monjardet B. (1970) *Ordre et classification : algèbre et combinatoire*. Paris, Hachette.
- Berge C. (1987) *Hypergraphes*. Paris, Gauthier-Villars.
- Bharat, K., M. Henzinger (1998), improved algorithms for topic distillation in a hyperlinked environment. In: W. B. Croft et al (eds). *Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval*. ACM press, pp. 104-111
- Björneborn, L. (1999), Link patterns on the World Wide Web, <http://ix.db.dk/lb/linkpatterns/page.htm#2>, abgefragt am 19. 10. 2001
- Björneborn, L. and Ingwersen, P. (2000): *Perspectives of Webometrics*, abgefragt am 19. 11. 2001
- Björneborn, Lennart and Peter Ingwersen. *Perspectives of webometrics*. *Scientometrics*, 2001, 50(1), 65-82.
- Boudourides, M. A., Sigrist, B and Alevizos, P. D. (1999): *WEBOMETRICS AND THE SELF-ORGANIZATION OF THE EUROPEAN INFORMATION SOCIETY*, Draft Report, Task 2.1 of the SOEIS project; Rome Meeting, June 17-19, 1999, <http://hyperion.math.upatras.gr/webometrics/>, abgefragt am 19. 10. 2001

Boudourides, Moses A.; Sigrist, Beatrice & Alevizos, Philippos D.(1999). "Webometrics and the self-organization of the European Information Society". Draft Report presented during the Rome Meeting of the SOEIS project. June 17-19, 1999. <<http://hyperion.math.upatras.gr/webometrics>>

Braam, R. R., Moed, H. F. & van Raan, A. F. J. (1991). Mapping of science by combined co-citation and word analysis. II: Dynamical aspects. *Journal of the American Society for Information Science* 42(4): 252-266.

Brin, S. and L. Page (1998), the anatomy of large-scale hypertextual Web search engines, WWW7 conference. (<http://www-db.stanford.edu/~backrub/google.html>)

Broder A., Kumar R., Maghoul F., Raghavan P., Rajagopalan S., Stata R., Tomkins A., Wiener J.. (2001) Graph structure in the Web, available in <http://www.almaden.ibm.com/>

Broder, A., Kumar, R., Maghoull, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A. & Wiener, J. (2000). Graph structure in the web. *Computer Networks*, vol. 33, no. 1-6, pp. 309-320. Available online at: <http://www.almaden.ibm.com/cs/k53/www9.final>

Burt, R. (1992). *Structural Holes*. Chicago: Chicago University Press.

Callon M., Law J., and Rip A. eds (1986) *Mapping the Dynamics of Science and Technology*. London, Macmillan.

Callon, M. (1986). The sociology of an Actor Network: The case of the electric vehicle. In M. Callon, J. Law & A. Rip (eds.), *Mapping the Dynamics of Science and Technology: Sociology of Science in the Real World*, pp. 19-34. London: Macmillan.

Callon, Michel (1986a). *The Sociology of an Actor Network: The Case of the Electric Vehicle* .

Callon, Michel, Courtial, Jean-Pierre, Turner, William et Bauin, Serge, "From Translation to Problematic Networks: an Introduction to Co-word Analysis", *Social Science Information* , 1983, 22, 2, pp 191-235.

Callon, Michel, J. Law, A.Rip, : *Mapping the Dynamics of Science and Technology. Sociology of Science in the Real World* , 19-34. Macmillan, London.

Campanario J. M. Using Neural Networks To Study Networks of Scientific Journals, *Scientometrics*, 33 (1995) No. 1, p. 23-40.

Chakrabarti S. Dom B. E. Kumar S. R. Prabhakar R. Rajagopalan S. Tomkins A. Gibson D. (1999) "Mining the Web's Link Structure," *IEEE Computer*, p. 60-67.

Christensen, F.H. and Ingwersen, P. (1996): Online citation analysis: a methodological approach, *Scientometrics*, 37(1), 1996, pp. 39-62, <http://ix.db.dk/cis/texts/abstract10.htm>, abgefragt am 15. 10. 2001

Christensen, F.H., Ingwersen, P. and Wormell, I. (1997): Online determination of the Journal Impact Factor and its international properties, *Proceedings of the 5th Biennial Conference of the International Society for Scientometrics and Informetrics*, Jerusalem, June 1997. Ed. by B. Peritz et al., <http://ix.db.dk/cis/texts/abstract1.htm>, abgefragt am 15. 10. 2001

Courtial J. P. (1986) *Technical Issues and Developments in Methodology*, in Callon, Law, and Rip (1986), p. 189-210

Cozzens, S. E.. What do citations count? The rhetoricfirst model. *Scientometrics*, 15:437--447, 1989.

Cozzens, Susan E. and Loet Leydesdorff, *Journal Systems as Macro-Indicators of Structural Change in the Sciences*, in: A. F. J. Van Raan, R. E. de Bruin, H. F. Moed, A. J. Nederhof and R. W. J. Tijssen (eds.), *Science and Technology in a Policy Context* (Leiden: DSWO Press, 1993), 219-33.

Degenne A., Forsé M. (1994/2001) *Les réseaux sociaux*. Paris, Armand Colin

Duquenne V. (1992) *GLAD (General Lattice Analysis & Design): A Fortram program for a Glad user*. Paris, MSH-Maison Suger.

Duquenne V. (1993) *GLAD*. Paris, C.N.R.S

Egghe, L. (1997). Fractal and informetric aspects of hypertext systems. *Scientometrics*, vol. 40, no. 3, pp. 455-464.

Elkana Y., Lederberg J., Merton R. K., Thackray A., Zuckerman H., eds. (1978) *Toward a Metric of Science: The Advent of Science Indicators*. New York, Wiley

Erdős, P., & Rényi, A. (1959). On random graphs. I. *Publicationes Mathematicae (Debrecen)*, vol. 6, pp. 290-297.

Faloutsos, M., Faloutsos, P., & Faloutsos, C. (1999). On power-law relationships of the Internet topology. *ACM SIGGCOM 99, Comp. Comm. Rev.*, vol. 29, pp. 251-262.

Fosmire, M. and Yu, S. (2000): Free Scholarly Electronic Journals: How Good Are They?, *Issues in Science and Technology Librarianship*, 27, <http://www.library.ucsb.edu/istl/00-summer/refereed.html>, abgefragt am 19. 10 2001

Freeman L. (1992) La résurrection des cliques : applications du treillis de Galois, *Bulletin de Méthodologie Sociologique*, 37, p. 3-24

Fujigaki, Yuko & Loet Leydesdorff, "Quality Control and Validation Boundaries in a Triple Helix of University-Industry-Government Relations: 'Mode 2' and the Future of University Research," *Social Science Information* 39(4) (2000) 635-655.

Fujigaki, Yuko (1998) Filling the gap between discussions on science and scientists' everyday activities: applying the autopoiesis system theory to scientific knowledge. *Social Science Informatics* 37(1), pp 5-22

Gibbons, M.C., H. Limoges, S. Nowotny, P.S. Schwartzman and M. Trow (1994): *The New Production of Knowledge*, London, Sage.

Gibson, D., J. Kleinberg, P. Raghavan (1998). Inferring webcommunities from link topology, *Proceedings of the 9th ACM Conference on Hypertext and hypermedia*. (<http://www.cs.cornell.edu/home/kleinber/ht98.pdf>)

Ginsparg, P. *Winners and Losers in the Global Research Village*. Proceedings of the Joint ICSU Press/UNESCO Expert Conference on ELECTRONIC PUBLISHING IN SCIENCE. UNESCO, Paris, 19-23 February 1996

Glänzel, W. (2001): An introduction to principle differences between citations and citation links. A methodological and mathematical approach, Proc. 6th Nordic Workshop on Bibliometrics, 4.-5. 10. 2001, <http://www.umu.se/inforsk/6thNordicBibliometric.htm>, abgefragt am 17. 10. 2001

Gläser, J. (2001) Scientific Specialties as the (Currently Missing) Link between Scientometrics and the Sociology of Science. Proceedings ISSI 2001 Australia.

Gould, R.V., & Fernandez, R.M. (1989). Structures of mediation: A formal approach to brokerage in transaction networks. *Sociological Methodology*, vol. 19, pp. 89-126.

Granovetter, M. (1973). The strength of weak ties. *American Journal of Sociology*, vol. 78, pp. 1360-1380.

Granovetter, M. (1974). *Getting a Job*. Cambridge, MA: Harvard University Press.

Harter, S.P. (1996): The Impact of Electronic Journals on Scholarly Communication: A Citation Analysis, *The Public-Access Computer Systems Review* 7, no.5, <http://info.lib.uh.edu/pr/v7/n5/hart7n5.html>, abgefragt am 15. 10. 2001

Harter, S.P. and Kim, H.J. (1996): Electronic Journals and Scholarly Communication: A Citation and Reference Study, <http://ezinfo.ucs.indiana.edu/~harter/harter-asis96midyear.html>, abgefragt am 15. 10. 2001

Haythornthwaite, C. (1996). Social network analysis: An approach and technique for the study of information exchange. *Library and Information Science Research*, vol. 18, pp. 323-342.

Hecht-Nielsen R. (1990) *Neurocomputing*. New York, Addison-Wesley

Hernández-Borges, A.A., Macías-Cervi, P., Gaspar-Guardado, M.A., Torres-Álvarez de Arcaya, M.L., Ruiz-Rabaza, A. and Jiménez-Sosa, A. (1999): Can Examination of WWW Usage Statistics and other Indirect Quality Indicators Help to Distinguish the Relative Quality of Medical websites?, *Journal of Medical Internet Research* 1(1), e1, <http://www.jmir.org/1999/1/e1/>, abgefragt am 17. 10. 2001

Hulme EW (1923) *Stat Biblio Relation*

Ingwersen, P. and Christensen, F.H. (1997): Data set isolation for bibliometric online analysis of research publications: fundamental methodological issues, *Journal of the American Society for Information Science*, vol. 48 (1997) 3, pp. 205-217, <http://ix.db.dk/cis/texts/abstract2.htm>, abgefragt am 15. 10. 2001

Jeong, H., Tomber, B., Albert, R., Oltvai, Z.N., & Barabasi, A.-L. (2000). The large-scale organization of metabolic networks. *Nature*, vol. 407, pp. 651-654. Available online at: <http://www.nd.edu/~networks/cell/papers/metabolic.pdf>

Katz, J. S., The self-similar science system, *Research Policy* 28 (1999) 501-517

Katz, J.S. and D.Hicks,(1997) *Bibliometric Indicators for National Systems of Innovation*. Available <http://www.sussex.ac.uk/Users/sylvank/best/nsi/index.html>.

Kay, S.M. (1988). *Modern Spectral Estimation: Theory and Applications*. New Jersey: Prentice Hall.

Kessler, M. M. (1963): Bibliographic coupling between scientific papers, *American Documentation* 14, p. 10-25.

Kleinberg, J. M., 1998, Authorative sources in a hyperlinked environment, *Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms*, pp.668-677.

Knoke, D. (1990). *Political Networks: The Structural Perspective*. Cambridge, UK: Cambridge University Press.

Kochen, M. (ed.) (1989). *The Small World*. Norwood, NJ: Ablex Publishing Corporation.

Kohonen et al. (1999, p. 171-182)

Kohonen T. (1984) *Self-Organizations and Associative Memory*, Berlin: Springer-Verlag.

Kohonen T. (1997) *Self-organizing Maps*. Second Edition. Berlin, Springer Verlag.

Krugman, P. (1996) "Are currency crises self-fulfilling?", *NBER Macroeconomics Annual*

Kugiumtzis, D., & Boudourides, M.A. (1998). Chaotic analysis of Internet ping data: Just a random number generator? SOEIS project conference, Bielefeld, Germany, March 27-28, 1998. Available online at: <http://www.math.upatras.gr/~mboudour/articles/ping.ps>

Kumar, R., Raghavan, P., Rajagopalan, S., & Tomkins, A. (1999). Trawling the Web for emerging cyber-communities. *Computer Networks*, vol. 31, pp. 1481-1493.

Lamirel J-Ch. (1995) *Application d'une approche symbolico-connexionniste pour la conception d'un système documentaire hautement interactif*. Ph.D. Université de Nancy 1 Henri Poincaré.

Lamirel J-Ch., Ducloy J., and Oster G. (2000) Adaptive browsing for information discovery in an iconographic context. *Proceedings of the RIAO Conference*, vol. 2, p. 1657-1672

Larson, R.R (1996): *Bibliometrics of the World Wide Web: An Exploratory Analysis of the Intellectual Structure of Cyberspace*, <http://sherlock.berkeley.edu/asis96/asis96.html>, abgefragt am 15. 10. 2001

Latour B. (1987) *Science in Action*. Milton Keynes, Open University Press

Latour B. (1988) *The Pasteurization of France*. Cambridge, Mass., Harvard University Press.

Latour, B. (1987). *Science in Action*. Cambridge, MA: Harvard University Press.

Lebart L., Morineau A., Piron M. ( 1995) *Statistique exploratoire multidimensionnelle*. Paris, DUNOD.

Lei Cui, M. S. (1999) Rating Health Web sites using the principles of Citation Analysis: A Bibliometric Approach, *Journal of Medical Internet Research* 1999, 1 (1), e4, <http://www.jmir.org/1999/1/e4/index.htm>, abgefragt am 15. 10 2001

Lelu A. and François C. (1992a) Information retrieval based on a neural unsupervised extraction of thematic fuzzy clusters. *Les Réseau Neuro-Mimétiques et leurs Application Conference*, 2-6 November, Nîmes, France.

Lelu A. and François C. (1992b) Hypertext paradigm in the field of information retrieval: A neural approach. 4th ACM Conference on Hypertext, 30 November – 4 December, Milan, Italy.

Lelu, A.(1991) From data analysis to neural networks: New prospects for efficient browsing through databases, *Journal of Information Science*, vol. 17, p. 1-12.

Lelu, A.(1993) *Modèles Neuronaux pour l'Analyse de Données Documentaires et Textuelles*. Ph.D. Université de Paris 6.

Lepair C (1988) The citation gap of applicable science. In: *Hdb Quantitative Stu* P537

Leydesdorff, L and Curran, M. (2000): Mapping University-Industry-Government - Relations on the Internet: The Construction of Indicators for a Knowledge-Based Economy, *International Journal of Scientometrics, Informetrics and Bibliometrics* 4 (1), <http://www.cindoc.csic.es/cybermetrics/articles/v4i1p2.html>, abgefragt am 15. 10. 2001

Leydesdorff, Loet & Paul Wouters, *Between Texts and Contexts: Advances in Theories of Citation? (A Rejoinder)*, *Scientometrics* 44 (1999) 169-182.

Leydesdorff, Loet *The Non-linear Dynamics of Sociological Reflections*, *International Sociology* 12 (1997) 25-45.

Leydesdorff, Loet. *Indicators of Innovation in a Knowledge-based Economy*. *Cybermetrics*, 5 (Issue 1), Paper 2, at <http://www.cindoc.csic.es/cybermetrics/articles/v5i1p2.html> or <http://www.cindoc.csic.es/cybermetrics/articles/v5i1p2.pdf>

Leydesdorff, Loet. *Words and Co-Words as Indicators of Intellectual Organization*, *Research Policy* 18 (1989) 209-223.

Lin X., Soergel D., Marchionini G. (1991) A self-organizing semantic map for information retrieval; *Proceedings of the 4th International SIGIR-ACM Conference on R&D in Information Retrieval*, 13-16 October, Chicago, USA, p. 262-269.

Luger G. and Stubblefield W. A. (1999) *Artificial Intelligence*. Reading, Mass., Addison Wesley Longman.

Luukkonen, T. (1997). Why has Latour's theory of citations been ignored by the bibliometric community? Discussion of sociological interpretations of citation analysis. *Scientometrics* 38, 27-37.

Marsden, P., & Campbell, K.E. (1984). Measuring tie strength. *Social Forces*, vol. 63, pp. 482-501.

Maturana, H. R. & Varela, F. J. ( 1980), *Autopoiesis and Cognition: The Realization of the Living*, Vol. 42 of *Boston Studies in the Philosophy of Science*, D. Reidel Publishing Company, Dordrecht, Holland. With a preface to 'Autopoiesis' by Stafford Beer. Series editors: Robert S.Cohen and Marx W. Wartofsky.

McLachlan G. J. (1992) *Discriminant Analysis and Statistical Pattern Recognition*. New York, J. Wiley.

Milgram, S. (1967). The small-world problem. *Psychology Today*, vol. 1, pp. 60-67.

Newman, M.E.J., Strogatz, S.H., & Watts, D.J. (2001). Random graphs with arbitrary degree distributions and their applications. *Physics Review E*, vol. 64, 026118. Online available at: <http://xxx.lanl.gov/abs/cond-mat/0007235>

NIWI Research programme 2000-2004 ([http://www.niwi.knaw.nl/us/research/res\\_prog.pdf](http://www.niwi.knaw.nl/us/research/res_prog.pdf))

Okubo, Yoshiko *Bibliometric Indicators and Analysis of Research Systems: Methods and Examples*. STI working papers. OECD 1997.

Polanco X. (2001) *Clusters, Grafos, Redes*. Powerpoint presentation in Spanish. 5° Taller Iberoamericano / Interamericano de Indicadores de Ciencia y Tecnología, 15-18 de Octubre de 2001, Montevideo, Uruguay.

Polanco X. and François C. (2000b) Data clustering and cluster mapping or visualization in text processing and mining. *Proceedings of the 6th International ISKO Conference*, 10-13 July, Toronto, Canada, *Advances in Knowledge Organization*, 7, p. 359-365

Polanco X., Boudourides M. A., Besagni D., Roche I. (2001) Clustering and Mapping European University Web Sites Sample for Displaying Associations and Visualizing Networks, *ETK & NTTS 2001 Pre-Proceedings of the Conference*, Hersonissos (Crete) 18-22 June 2001, Volume 2, p. 941-944

Polanco X., François C., and Lamirel J-Ch. (2001) Using artificial neural networks for mapping of science and technology: A multi-self-organizing maps approach, *Scientometrics*, vol. 51, No 1, p. 267-292

Polanco X., François C., Keim J-P. (1998a) Artificial Neural Network Technology for the Classification and Cartography of Scientific and Technical Information, *Scientometrics*, vol. 41, Nos 1-2, p. 69-82

Polanco X., François C., Lamirel J-Ch. (2000a) Using artificial neural networks for mapping of science. *6th International Conference on Science and Technology Indicators*. 24-27 May, Leiden, The Netherlands. *Book of Abstracts*, p. 89.

Polanco X., François C., Louly M. A. O. (1998c) For visualization-based analysis tool in knowledge discovery process: A multilayer perceptron versus principal component analysis – a comparative study. In J. M. Zytkow and M. Quafafou (eds.) *Principles of Data Mining and Knowledge Discovery*. Berlin, Springer Verlag, p. 28-37

Polanco, François, Louly M. A. O. (1998b) An artificial neural networks perspective on cartography. *Proceedings of the 5th International ISKO Conference*, 25-29 August, Lille, France. *Advances in Knowledge Organization*, 6, p. 64-71

Price (1965), "Network of Scientific Papers," *Science*, vol. 149, p. 510-515

Price, D. de Solla 1963, *Little Science, Big Science*, Columbia Univ. Press, New York.

Priestley, M.B. (1988). *Non-linear and Non-stationary Time Series Analysis*. New York: Academic Press.

Pritchard, A. 1969, *Statistical bibliography or bibliometrics?* *Journal of Documentation* 24, 348-349.

- Rand Corporation. Approaches to evaluation of science. (<http://ms161u06.u-3mrs.fr/annexe/Cyber0-2804/PAGE/page352.html>)
- Rifkin, Jeremy ,(2000), THE AGE OF ACCESS: The New Culture of Hypercapitalism, Where All of Life Is a Paid-for Experience. Jeremy P. Tarcher/Putnam.
- Rogers, R. (ed) Preferred Placement. Jan van Eyck Editions Maastricht 2000.
- Rousseau R. (1997). Sitations. An exploratory study. Cybermetrics, 1 paper 1. ISSN: 1137-5019. (<http://www.cindoc.csic.es/cybermetrics/vol1iss1.html>)
- Smith, A. (1999, 1): ANZAC webometrics: exploring Australasian Web structures, Information online & on disc, Proceedings of the Ninth Australasian Information Online & On Disc Conference and Exhibition, Sydney Australia, 19–21 January 1999, <http://www.csu.edu.au/special/online99/proceedings99/203b.htm>, abgefragt am 22. 10. 2001
- Smith, A. (1999, 2): The Impact of Web sites: a comparison between Australasia and Latin America, presented at INFO'99, Congreso Internacional de Informacion, Havana, 4-8 October 1999.
- Sowa J. F. ed (1991) Principles of Semantic Networks. San Mateo, CA., Morgan Kaufmann
- Tsonis, A.A. (1992). Chaos: From Theory to Applications. New York: Plenum Press.
- Turnbull, D. (2000): Bibliometrics and the World-Wide Web, <http://donturn.fis.utoronto.ca/research/bibweb.html>, abgefragt am 17. 10. 2001
- Van den Besselaar, Peter & Loet Leydesdorff, Mapping Change in Scientific Specialties: A Scientometric Reconstruction of the Development of Artificial Intelligence, Journal of the American Society for Information Science 47 (1996) 415-36.
- Van den Besselaar, Peter and Gaston Heimeriks, Deliverables for the European Commission: Codification and Self-Organization in the European STI System Deliverable Task 4 SOEIS
- Van Raan A. F. J. ed. (1987) Handbook of Quantitative Studies of Science and Technology. Amsterdam, North-Holland, Elsevier Science Publishers.
- van Raan, A.F.J. (Ed.) 1988, Handbook of Quantitative Studies of Science and Technology, North-Holland, Amsterdam.
- van Raan, A.F.J. Bibliometrics and internet: Some observations and expectations. Scientometrics 50 (1):59-63, January 2001."
- Wagner, A., & Fell, D. (2001). The small world inside large metabolic networks. Proc. Roy. Soc. London Ser. B. (in press). Available online at: <http://samba.unm.edu/~wagnera/wagnerfell2000.pdf>
- Wasserman, S., & Faust, K. (1994). Social Network Analysis: Methods and Applications. Cambridge, UK: Cambridge University Press.
- Watts, D.J. (1999). Small Worlds: The Dynamics of Networks Between Order and Randomness. Princeton, NJ: Princeton University Press.
- Watts, D.J., & Strogatz, S.H. (1998). Collective dynamics of 'small-world' networks. Nature, vol. 393 (June 4), pp. 440-442.

- Wellman, B. (1988). Structural analysis: From method and metaphor to theory and substance. In B. Wellman & S.D. Berkowitz (eds.), *Social Structures: A Network Approach*, pp. 19-61. Cambridge, UK: Cambridge University Press.
- Wellman, B., & Potter, S. (1999). The elements of personal communities. In B. Wellman (ed.), *Networks in the Global Village: Life in Contemporary Communities*, pp. 49-81. Boulder: Westview Press.
- Wellman, B., Salaff, J., Dimitrova, D., Garton, L., Gulia, M., & Haythornthwaite, C. (1996). Computer networks as social networks: Collaborative work, telework, and virtual community. *Annual Review of Sociology*, vol. 22, pp. 213-238.
- White H. D., Lin X., McCain K.W. (1998) Two Modes of Automated Domain Analysis: Multidimensional Scaling vs Kohonen Feature Mapping of Information Science Authors. in *Proceedings of the 5th International ISKO Conference, 25-29 August, Lille, France. Advances in Knowledge Organization*, 6, p. 57-63
- Wilson, W.J. (1997). *When Work Disappears: The World of the New Urban Poor*. New York: Alfred A. Knopf.
- Winston P. H. (1977) *Artificial Intelligence*. Reading, Mass., Addison Wesley
- Wormell, I. (1998): Online searching is like gold-washing, presentation at the Online Information Scandinavia 98. May 12-14, 1998, Stockholm International Fairs, <http://ix.db.dk/cis/texts/artikel3.htm>, abgefragt am 19. 11. 2001
- Wouters, P. 1999, *The citation culture*, PhD Thesis, Private edition.
- Wouters, Paul (1998). The signs of science. *Scientometrics* 41, 225-241.
- Ziman, J. (1994): *Prometheus Bound: Science in a Dynamic Steady State*, Cambridge, Cambridge University Press.
- Ziman, J. M. (1984). *An introduction to science studies: The philosophical and social aspects of science and technology*. Cambridge: Cambridge University Press.
- Zurada J. M. (1992) *Introduction to Artificial Intelligence*. New York: West Publishing.