

**REFEREE’S REPORT BY WOLFGANG GLÄNZEL ON THE EUROPEAN
PROJECT ENTITLED “EUROPEAN INDICATORS, CYBERSPACE AND THE
SCIENCE–TECHNOLOGY–ECONOMY SYSTEM”
(IST-1999-20350)**

[MID-TERM REVIEW]

Content	page
1. Executive summary	1
2. Overview of the project	2
3. Comments on the published deliverables	3
4. Suggestions for the pending work	8
5. Curriculum vitae	9

1. EXECUTIVE SUMMARY

EICSTES is a large multidisciplinary project combining a variety of different approaches and developing new methodologies necessary to achieve the objectives of the project. The studies and results provided in the deliverables of the workpackages range from state-of-the-art reports being of WP 1 and presenting an excellent overview of the groundwork on which the project is methodologically based, over physical Internet statistics (WP 2) that also helps to understand how physical Internet data can be used as indicators of the complex dynamics occurring in the Information Society and the New Economy, non-web data collection and analysis (WP 5), the development of new webindicators (WP 8), to the visualisation of cyberspace data (WP 9) and a user-friendly presentation of results on the Web. Most results of the deliverables of the workpackages are in keeping with the proposed objective. Some of the objectives have been modified or could not yet been achieved and one of the initial hypotheses of WP 6, namely, that intermediaries become one of the major economic growth points of the web is still under discussion.

The project combines state-of-the-art techniques derived or adapted from related fields, such as bibliometrics and technometrics, with novel approaches appropriate for analysing, describing and measuring the complex phenomenon of the *New Economy*. Several workpackages, above all, WP 5 are also designed as empirical validation of new indicators.

The integration of the workpackages into the whole project is good. However, there are redundancies since most reports are using results of others. This should be reduced to an acceptable minimum. The English of several deliverables could be brushed up, typographic errors and errors in several equations as well as missing or incomplete references should be corrected for the final version.

In all, EICSTES is a very ambitious project the objectives of which are a truly innovative quantitative approach to the description and analysis of the New Economy. The results achieved at the time of mid-term evaluation promise a successful finalisation of the project.

2. OVERVIEW OF THE PROJECT

The EICSTES project which is funded by the Fifth Framework Program of R&D of the European Commission aims at developing indicators and offering statistics on the European Indicators, Cyberspace and the Science – Technology – Economy System in Internet. From the scientific and technological point of view, this is basically done in the five main steps.

- First, new technologies are developed to recover and collect data about the Internet from the Web itself and from external sources.
- In a second step, obtained data are organised in statistical databases.
- New indicators are developed, advanced mathematical and statistical techniques are used to analyse, describe and measure phenomena of the emerging information society and the so-called *New Economy*.
- New models are created for the dynamics of the *New Economy*.
- Finally, the indicators of the *New Economy* are presented in a universal user-friendly environment using new web visualisation techniques.

In order to achieve the proposed objectives, cases studies are prepared to apply quantitative methods, the web indicators are tested and validated (also from the viewpoint of system dynamics) and the role of intermediaries in the cyberspace are analysed in the framework of this project.

The project is organised and developed as a series of 10 workpackages the deliverables of which will be discussed in the following section. Seven institutions from six countries are participating in the EICSTES project.

EICSTES also contributes to other European programmes and key actions, in particular, to Key Actions 1 – 3 of the Information Society Technologies Programme (IST) and the Statistical Indicators for the New Economy (SINE), as well as eEurope 2002 initiatives. There is a strong concordance between the objectives of EICSTES and the needs outlined in the eEurope 2002 initiative.

For the mid-term evaluation, all necessary deliverables are available. However, only six deliverables are intended to serve as a basis of this evaluation, namely, D.2.1., D.3.1., D.5.1. - 5.2., D.6.1. and D.8.1. Nevertheless, I will also make some comments on the State-of-the-Art papers that are part of WP 1 (deliverables D.1.4.A - 1.4.C) and on WP 9 (deliverable D.9.1.).

3. COMMENTS ON THE PUBLISHED DELIVERABLES

D.1.4A–C State of the Art

These deliverables report on the state of the art of the topics of WP 6, 8 and 9. Especially, D.1.4B and D.1.4C present a profound introduction into the background of biblio-/webometrics and the visualisation of information. They also present an excellent overview of the existing methodology and literature. Remarkable in deliverable D.1.4C is the organisation and the excellent integration of the components of the study. By contrast, deliverable B.1.4B is rather focussed on bibliometrics and the common in bibliometrics and webometrics. Here, a more critical view at principle differences is needed. At present, a considerable part of webometric methodologies is still derived from their bibliometric “equivalents”. Some methodological questions in bibliometrics that might possibly highly relevant in webometrics, too, are discussed only in part. In this context, the role of co-citation analyses is somewhat overestimated. Limitations, especially applying to young subjects, should be mentioned (see, for instance, *Diana Hicks, Limitations of co-citation analysis as a tool for science policy, Social Studies of Science*, 17, 1987, 295-316). These limitations are a consequence of ageing characteristics of scientific literature and lacking “critical mass”. Because of completely different ageing concept in webometrics, especially, these properties might be crucial in webometric application. This might serve just as an example for problematic cross-field application.

Revising the text, removing typographic errors and correcting incorrect equations could improve the quality of this report.

D.2.1. Physical Internet Statistics

The main objectives of WP 2 are to collect and study Internet physical data, to analyse and describe the data and their value, and the methodologies and tools used. In order to achieve these objectives, four forms of Internet traffic data have been described: network topology (data about the European networks), workload measurements (measurement of traffic information from a point within the network), end-to-end characteristics (through measuring the latency of the paths, the percentage of packets lost in them and the changes of the paths) and web caching analyses. The results also give an insight into the use of physical Internet data as global indicators of *New Economy*.

This report is a profound, well-organised and polished paper that is in complete accordance with the aims and tasks proposed. Profound and useful background information is given; all methods and tools are properly described. There are no further deliverables planned in the framework of this workpackage and no further deliverables are needed for WP 2, either.

D.3.1. Directory of EU R&D public system

One of the objectives of the EICSTES project is to identify the European R&D public system in Internet at the level of websites, and to discover and describe the new economy patterns using academic and research institutional websites as source of data. This deliverable describes approach, methodology, problems to be solved and presents and discusses the results.

End 2001, the database has covered more than 24.000 University and non-University websites so far, what corresponds to about 20% of the population of the proposal. For the final value, the author expects more than 26.000 sites with over 40.000.000 web objects.

Impressive effort has been made to achieve these results. However, the objectives proposed for this deliverable have not been completely achieved. Moreover, the report seems to be written a bit hastily; the English should be improved.

D.5.1. “Triple helix” case study

Content and organisation of workpackage 5 have been slightly modified. Originally planned as “Triple helix” case study, it is now focussing on the “Meaning of links“. D.5.1 is designed also to contribute to WP 8. The case studies in D.5.1 are based on data collected from the web. Problems in connection with the use of “link data” from the web are briefly discussed in the *Methodological Reflections*.

The main objective of the first case study is to analyse what is communicated by cyber-links on the basis of the academic linking structure at four different levels of aggregation (between EU countries, within the Netherlands, the linkage structure of a research group, particularly, of SWI and of individual researchers). At the macro and meso level MDS and factor analysis has been applied to cluster outlook and inlink patterns. The micro level analysis is based on software from ARCS, that is, from one of the partners contributing to EICSTES, however, the software has not been specified. Nevertheless, methodology is sound and data collected from the Internet itself provide an appropriate basis for the study. The outcomes are promising; they are (especially at higher levels of aggregation) large and by in keeping with those obtained from similar bibliometrics studies. The ‘research group’ proved to be the most meaningful unit of analysis.

The second case study is devoted to the analysis of the Triple Helix network of University – Industry – Government relations. Most significant clusters of institutions have been identified by factor analysis of the linkage matrix. The outcomes are interesting as they in part contradict common belief: Link structures on the web proved to be strongly dominated by universities. The hypothesis that lower levels of aggregation are the main source of information about the complex dynamics of knowledge production and new economy could be confirmed in this case study, too.

(Scientific) communication patterns of a research group have been analysed in the third case study. Besides the inlinks and outlinks, also email communication has been analysed. A statistic on scientific co-publications and a survey have been added. The study results in the conclusion that measures developed on these data are potential indicators for academic environment at lower levels of aggregation.

Although the objectives have been modified, D.5.1. is a methodologically sound and important paper that contributes to the issue of web-indicators both in terms of validation and meaning of internet-based indicators. It is thus contributing also to achieve the objectives of WP 8, too. At the same time, it gives insight into the role of electronic communication in the development of science and technology.

D.5.2. Non-web data collection and analysis

This deliverable is devoted to the analysis of non-web data in the context of web. According to the new design of WP 5 (cf. previous section), science and social-science domains are delineated in the JCR, the institutes are obtained from the (S)SCI and mapped on the CORDIS database, Miri@d (cf. D.4.2.) and the web. The non-web data are used to study techno-scientific evolving communication, later on, the relationship of the print based networks with electronic networks of scientific communications is analysed. The authors proceed from the assumption, that at the boundaries of “disciplinary“ research fields, multi-disciplinary and interdisciplinary fields emerge. Three typical mode 2 science fields, in particular, *artificial intelligence*, *biotechnology* and *information science* have been selected and studied at different stages of their evolution. The first task was to delineate the 'identity' of 'interdisciplinary' research fields. This has been done starting with core journals; the results of an analysis of journal cross-citation matrices have then been used to delineate the fields.

BibTechMon from ARCS has been used to visualise data in knowledge maps. Maps are based on both, keywords and cited references (bibliographic coupling). In addition, co-author networks have been analysed. In the fourth chapter, CORDIS and SCI databases have been matches at the lowest possible institutional level, and similar techniques already used in the previous chapter have been applied to the actors in the intersection. The analysis of data from the Miri@d database has been left for future steps.

Although this deliverable should play an important part also for other workpackages, it has at its present stage still several shortcomings. First, it contains a lot of redundancies. It forms certainly a continuation of D.5.1., so that a better harmonisation with that deliverable (but also with others) might help to overcome this problem. However, redundancies within D.5.2. could be avoided, too. Especially, chapter 1-3 can be shortened.

The methodology of this deliverable is sound; nevertheless, the bibliometric background of data collection and methods applied (e.g., in the context of bibliographic coupling, co-publication links) is not always sufficiently explained. At least the field *information science*, for instance, is rather covered by the SSCI than by the SCI. The authors should use the standard bibliometric parlance. Also this shortcoming could be overcome by proper harmonisation with other deliverables. All data should be carefully interpreted in their statistical context. I am afraid the strong German-Hungarian co-authorship link in *information science* in 1996, for instance, might prove to be an artefact.

Besides the missing analysis of Miri@d data, I would like to mention that also analyses of data from patent databases and Medline were originally planned.

D.6.1. Intermediaries' functions, operations and types – a taxonomy

The objective of this deliverable of workpackage 6 was twofold: 1. to develop a socio-economically relevant taxonomy of web intermediaries that is 2. amenable to automatic recognition by software agents.

The first objective has been achieved through analysing different approaches to classification. The study is based on three existing approaches, particularly, the web as market place, the web as navigation place, and hybrid typologies representing the web both as market place and navigation space. The authors have developed an own typology of *transit sites* that is also designed as a hybrid typology.

Taxonomy of infomediaries (defined as intermediary dealing with information in the link chain between author and information user) is given in the Annex.

The main unsolved problem remains the question of automation. The authors had to conclude that practically none of the discussed typologies proved to be appropriate for automatic recognition by software agents. This might be a serious obstacle to proving the initial hypothesis, namely, that intermediaries will become one of the major economic growth points of the web.

Nonetheless, this report (D.6.1.) remains an important contribution to the whole project.

D.8.1. Development of webindicators

This is a very good and comprehensive compilation of definition, methodology, presentation and interpretation of webindicators. D.8.1. is practically the conceptual, theoretical and methodological core of the project. It can serve as a good example of a well integrated, well harmonised deliverable. However, integration and harmonisation can still be improved (see below). WP 8 is using data from WP 2, 3, 4, and 5. On the other hand, users are expected to provide feedback in WP 7 on the indicators and visualisations are developed in WP 9 basing on data and indicators from WP 3, WP 4 and WP 5.

The first chapter is devoted to terminology and definitions. Here, a set of 23 basic definitions is given. This helps to avoid undesired misunderstandings as, for instance, unfortunately occurred due to a “lax” terminology, for instance, in early stages of bibliometric research.

In the second chapter, the web is considered a graph. This allows applying well-known mathematical models to phenomena such as the small-world-phenomenon. Other indicators based on the same concept are developed in the third chapter. The first set of indicators is based on *density*. In the second part, *centrality* and *centralization* specifying important structural characteristics of communication networks (*degree*, *betweenness* and *closeness*) are introduced.

In Chapter 3, the use of Miri@d data for indicator engineering is described. In this context, I would like to mention that the term ‘*Impact Factor*’ may be used *only* for Garfield’s Impact Factor (published in the annual editions of the *Journal Citation Report*). There are many versions and extensions of this measure in bibliometrics (not only that by *Egghe*). I suggest

using “impact measure” for all generalisations of the ISI Impact Factor. Outlines of usage and co-usage analyses on the basis of *Salton’s* vector-space approach and the definition of corresponding indicators are given in the last section. All Miri@d indicators and their definition are listed in an Annex.

Reorganising this chapter, modifying its logical structure and brushing up its text might help to improve the comprehensibility.

The last chapter is devoted to mapping and clustering web sites. This is done applying techniques of co-word analysis, originally developed for use in information retrieval and bibliometrics, to the analysis of web-site structure. This results in a *co-site analysis* that consists of four steps. The co-word analysis programme SDOC has been adapted to perform this co-site analysis. The study is based on data collected the Computer Technology Institute of Patras (Greece) that is member of the EICSTES consortium. The original data set has been reduced the 15 countries of the European Union. Besides the analysis of clusters, a two-dimensional mapping of internal and external associations is given. Finally, several indicators are proposed to measure cluster characteristics.

The methodology is sound; SDOC is already a classic tool in co-word analysis. The terminology could, however, be more “modest”. The declared aim is *web mining*; web-site mining would be more realistic. The “definition” of *information society* is in any case a step to far.

The indicators derived from the analysis of clusters can easily be confused with indicators defined from the *web as a graph* (chapter 2) bearing the same or at least similar names. In this context, a more detailed explanation is needed.

This deliverable is concluded by an exhausting list of web indicators. Moreover, detailed explanation on definition, characteristics, validity, interpretation, visualisation and related indicators is presented for all indicators discussed in this deliverable.

The objectives of WP 8, proposed till mid-term evaluation, seem by and large to be achieved. Especially, the set of indicators defined and described in this report is quite impressive.

D.9.1. and D.9.1b Progress Report on Visualisation Application Test

Interconnectivity of the great number of web objects is by far to complex to represent multidimensional indicators numerically for interpretation.

A large scale of data such as linkage, co-linkage, log-file, content, co-citation, co-sitation and collaboration data is visualised. In a second step accessibility of visualisation and presentation of results to a user is solved through designing a web site on which all kind of data and indicators are offered to the visitor.

Using appropriate bibliometric methods, the collected data about relationship between websites can be used to calculate similarities between websites and to visualise interrelations through networks. For visualisation, *BibTechMon* software originally developed at ARCS (one of the members of EICSTES consortium) for Bibliometric Technology Monitoring has

been adapted to webometric needs; in additions REMA (*Read Matrix*) software application for BibTechMon, the description of which is given in the Appendix, has been developed.

Among others, data from other workpackages (WP 3, 4 and 5) have been used for visualisation.

The last chapter is devoted to the prototype for a visualtion concept including the concept for representing the results according to the second proposed aim.

In all, the authors present a well organised, professionally written report. The results are in complete keeping with the proposed objectives.

4. SUGGESTIONS FOR THE PENDING WORK

Although most of the objectives proposed for the time from launching the project till mid-term evaluation has been achieved, several tasks had to be modified or could even not be performed. This work has still to be done after this evaluation.

The members of the consortium have attempted to avoid the initial errors and the “teething troubles” as criticised in the context of the evolution of the field of bibliometrics (see dedicated issue of *Scientometrics*, 30, 2-3, 1994, and the proceeding of the “Workshop on standardisation in bibliometric research and technology”, *Scientometrics*, 35, 2, 1996) by clarifying and unifying terminology and definitions.

Concerning future steps and possible improvement, I would like to make the following suggestion.

1. Models, methodology and software are to a large extent based on “external sources” such as bibliometrics. Cross-field applications might have undesired consequences. Some analyses on principle differences between bibliometric and webometric phenomena might help to avoid invalid conclusions and interpretations. I will just mention two examples.

Authorship and citations in printed scientific literature and thus in bibliometric models have relatively strict rules, although the practice has sometimes been criticised (e.g., authorship without really having contributed to the paper, authors cite literature without having read the cited literature, establishing co-called citation-cliques etc.). In printed scientific literature, information on funding or sponsorship and on research staff and the project, if such information is at all available, is given in different sections (acknowledgement) or in text accompanying the paper. There are no rules like these on the Web. There is, however, an even more crucial problem. In bibliometrics, authorship, co-authorship, citation and co-citation links are “cumulative” (like the number of visitors or downloads in webometrics) in the sense that once a link has been established it cannot be removed anymore. (There are a few retracted papers each year – but this from the statistical viewpoint a negligible phenomenon.) Besides subject matter, social status and document type, ageing is the main factor influencing when a system evolves to its asymptotically stable stationary distribution. Site-related web links can be removed or re-established anytime. While ageing in bibliometrics can be measured by changing activity or citation impact in time and time-windows can thus be chosen accordingly, ageing and stability (in terms of stationary solution) have a completely different meaning in webometrics. And while retractions in bibliometrics are a marginal phenomenon, research on changes of web links might contribute to the validation of webindicators defined in analogy to bibliometric measures.

2. From the part of non-web data, patent indicators and technometric techniques might be discussed and utilised in webometrics, too. This should be part of WP 5 (as originally planned).
3. The authors of deliverable D.8.1. have defined and discussed a plenty of webindicators. An “excess” of indicators has already frustrated users of bibliometrics, above all, in science policy. Besides validation, efficiency and interpretation of indicators might be a focus in further steps of WP 5 and 8.
4. Deliverables of WP 4 are obviously designed for internal use in EICSTES. However, more polished and comprehensible reports could improve the overall impression of the project deliverables.
5. Several problems concerning harmonisation and integration of workpackages has already been mentioned. Although all deliverables are stand-alone reports, redundancies could in future be avoided through proper cross-references.

In spite of several critical remarks made above, the deliverables available for mid-term evaluation allow the conclusion that the project will be successfully finalised and that the results will beyond doubt be of use for other European programmes and key actions, too.

5. CURRICULUM VITAE

Personal, education and professional experience

- Wolfgang Glänzel was born on 13 April, 1955 in Frankfurt (Oder), East-Germany.
- He studied *Mathematics* at the Eötvös University Budapest, Hungary (1974-1979) (field of specialisation: *Probability Theory and Mathematical Statistics*).
- He worked as assistant lecturer at the Dept. Mathematics & Informatics of the University of Applied Sciences Mittweida (Saxonia, East-Germany) in 1979-1980.
- Since 1980, he is a member of the Bibliometrics research unit at the Library, from 2002 on, at the Institute for Research Organisation of the Hungarian Academy of Sciences. He doing research in Bibliometrics and mathematics (applied probability theory).
- He received his Dr. rer. nat. from the Faculty of Science of the Eötvös University Budapest (Hungary) in 1984.
- He was Research Fellow of the Alexander von Humboldt Foundation at several German institutions (1990-1991) and 1995-1996).
- Head of a research group conducting a large bibliometric study on scientific research in Europe (1996-1997)
- He received his PhD from the Faculty of Social Sciences of University Leiden (Netherlands) in 1997 (nostrificated at the Faculty of Economic and Social Sciences of the University for Technology and Economics Budapest in 2000)
- Also mentor for mathematics at the Distance Education Centre (Budapest) of the Fernuniversität Hagen (Germany) 2000-2001
- Senior research fellow at the Dept. Applied Economics of the Katholieke Universiteit Leuven (Belgium)

Publications

Wolfgang Glänzel is author/co-author of more than 100 publications and guest editor of several dedicated issues of journals in science studies.

Membership and awards

- President of the *Research Association for Science Communication and Information e. V.* (RASCI) in Berlin
- Secretary-Treasurer of the *International Society for Scientometrics and Informetrics* (ISSI)
- Honorary member of the *Society for Science Studies e. V.* (GeWiF) in Berlin
- Member of the *Hungarian Humboldt Association*
- Co-Editor of the international journal *Scientometrics*
- Prize of the Hungarian Academy of Sciences awarded to young scientists (1986)
- International *Derek deSolla Price Award* for outstanding contributions to the quantitative studies of science (1999)