

Mapping communication and collaboration in heterogeneous research networks

Gaston Heimeriks*, Marianne Hörlesberger**, Peter Van den Besselaar*

**NIWI-KNAW, Royal Netherlands Academy of Arts and Sciences, Amsterdam (The Netherlands)*

*** ARCS, Austrian Research Centers, Seibersdorf (Austria)*

Address for correspondence:

Gaston Heimeriks

Social Sciences Department, NIWI-KNAW, Royal Netherlands Academy of Arts and Sciences

PO BOX 95110, 1090 HC Amsterdam

Gaston.Heimeriks@niwi.knaw.nl

Abstract

The aim of this mainly methodological paper is to present an approach for researching the triple helix of university-industry-government relations as a heterogeneous and multi-layered communication network. The layers included are: the formal scholarly communication in academic journals, the communication network based on project collaborations, and finally the communication of information over the 'virtual' network of web links. The approach is applied on typical 'Mode 2' fields such as biotechnology, while using a variety of data sources. We present some of the initial findings, which indicate the different structures and functions of the three layers of communication.

Introduction

Two major and related changes can be observed in the knowledge production system. Firstly, information and communication technologies (ICT) are increasingly important in knowledge production and knowledge use. On the one hand, the new communication technologies form additional means of communication between organizations in what can be called the Science-Technology-Economy system. On the other hand, the new modes of communication occupy the empty space that previously existed between the interpersonal communication of researchers and academic journals. Although information exchange has always been central to scientific research, the emergence of digital information, of online accessible databases, and of CMC have enabled a radical lowering of the costs related to collaboration, communication and information dissemination within the science system and between knowledge producers and users. As a consequence of this ICT revolution, new patterns of communication and collaboration are emerging.

Secondly, this so called informational turn in science is associated with a new mode of knowledge production, with hybrid roles of academic, commercial and governmental institutes, with a more prominent orientation on the social and economic role of science and technology, and consequently with a wider variety of types of research output such as instruments and patents, but also norms and values. It is not only the output of the knowledge production that is changing, the same holds for the operation and boundaries of the system, and for the patterns of communication and collaboration. Gibbons et. al. have introduced the concept of *Mode 2 knowledge production* to describe this change [7]. Using a different perspective, one may say that Mode 2 research fields function in a network of University-Industry-Government relations that constitutes an infrastructure which enables the innovative process of techno-scientific development: the *Triple Helix of university-industry-government relations* [5].

In other words, the knowledge production system is increasingly becoming a hybrid network, characterized by heterogeneous collaborations between different actors, and by heterogeneous communications using an increasing number of different media. Therefore, the knowledge production system can be conceptualized as an 'evolving communication system' in which computer mediated communication technologies do play an increasing role. This development can be studied from at least two different perspectives. First, research on the communicational behavior of knowledge producers has recently started to include scholarly communication using CMC technologies [13]. The organizational aspects of using CMC and the Internet for scholarly communication and information sharing is also increasingly studied [13, 14]. A second line of research builds upon the bibliometric and scientometric tradition in mapping the cognitive structures of scientific landscapes and in the study of scholarly communication [2, 3]. Here the focus is less on behavioral aspects of scholarly communication and more on the resulting (changing) communication networks. These methods are increasingly applied to the link structures of the WWW, and this has resulted in a new research field called webometrics [1, 18]. One of the drawbacks of much of this research is that it generally focuses on one medium only, such as on *printed media* within bibliometrics and on the *WorldWideWeb* within webometrics, and therefore does not clarify the different functions

and mutual relations of the various media within scholarly communication and collaboration. In this paper we begin to fill this gap by studying the interaction between various communication media, as this may enable us to account for the increasing complexity of communication in knowledge production. In this paper we focus on the following three levels of communication:

- Firstly, communication as it plays a role in the collective production of knowledge within scientific specialties [8, 9]. On this level, researchers publish smaller or larger knowledge claims in scholarly journals by using, modifying or rejecting previous knowledge claims published in the journals.¹
- Secondly, communication networks in terms of collaboration at the level of research projects. On this level, memberships of projects are relevant indicators for the emerging patterns of collaboration.
- Thirdly, communication of research related information using the WWW, using links patterns between web pages as indicators for these communication networks.

Senders as well as receivers may be producers and users of knowledge, and may be within the knowledge production system, or located elsewhere in society. The communication networks will be compared in terms of the actors involved (the nodes), the structure of the network (the relations), and finally in terms of what is communicated (content). As we do not restrict ourselves to the traditional forms of knowledge communication (the scholarly journals) and of knowledge application (patents), we must explore the possibilities of other data sources in order to improve our understanding of the dynamics of the new mode of knowledge production, in particular WWW data and data about collaborations in European projects.

We selected Biotechnology as an empirical field and used two other fields for comparison: Artificial Intelligence and Information Science. These fields belong to the *techno-sciences*, typical Mode 2 fields characterized by a strong emphasis on application, and subject to regularly reformulated S&T policies [7].

The main aim of this paper is to provide a framework for studying the mutual interaction of information and communication technologies (ICT) and knowledge production and dissemination. Therefore the focus is on the methodology we use to map and compare different levels of scholarly communication networks. In the next section we describe our general approach. In the third section we explain the methods used for each of the levels, and give some preliminary findings. Section four compares the findings between the three communication levels (and media), and in the fifth and last section we summarize the findings and formulate questions for further research.

¹ Of course, apart from paper journals other media also play a role here, such as e-journals, (electronic) preprints, and conferences. However, the 'traditional' journal is still the primary medium.

Approach, methods, and data

If one wants to study communication using different media, the first problem is the determination of the boundaries of the system under study. This is not trivial, as the actors, the relations, and the communicated content in the various communication networks (in our case journal, project, and website based networks) are expected to be different. The second problem is that the number of nodes in the network quickly becomes very large. Therefore we will not analyze the complete network, but select a core set of actors, and create the ego-network of this core using the three selected media.

Determining the core of the network

How to select the core set of actors? In which medium should they be core nodes? It is well known that for the traditional 'Mode 1' research field, one can use journal-journal citations to determine the boundaries of the research field in terms of a set of core journals [4]. The organizations frequently publishing in those journals can be considered as a core set of actors in the research field [22].

Underlying this method is the idea that researchers in a field share a common knowledge base, and that this is reflected in the references. Through local citing behavior, researchers reproduce the identity of the field on an aggregated level – and journal-journal citations can be used to map this identity in terms of sets of journals: a scientific discipline can be defined as a network of journals dealing with similar research questions and methodologies and referring to a largely overlapping set of literature. As a consequence of this last characteristic, we expect journals belonging to the same field to exhibit similar aggregated citation patterns. If that is the case, the analysis of journal-journal citations may result in an operational definition of a scientific specialty in terms of a set of journals with a similar citation pattern. The method has been developed and corroborated for delineating disciplinary fields [4], but in this study the focus is on interdisciplinary 'Mode 2' fields, which are not solely based on the traditional academic output of journal publications. Elsewhere we showed that the method is also well suited for delineating these interdisciplinary research fields [23].

The method is described in detail in another paper [24], therefore we summarize it here. The analysis begins with the selection of the entrance journal: the main journal in the specialty under study. Step two is the determination of the journals belonging to the citation environment of the entrance journal. These are all journals that cite or are cited by the entrance journal. The data are obtained from the *Journal Citation Reports (JCR)* of the *Science Citation Index (SCI)* and the *Social Science Citation Index (SSCI)*. As many of the journals identified in this way have only a very weak citation relation with the entrance journal, we remove all journals (step three) below a certain threshold. The remaining journals are used in step four to construct the journal-journal citation matrix. The matrices are factor-analyzed in the 'citing' dimension, and this generally results in clearly identifiable clusters of journals representing scientific disciplines. The factor including the entrance journal represents the specialty under study, and the other factors represent research fields obviously relevant for this specialty. By repeating this procedure using

data from different years, one can map the change in the specialty and in its relevant environment (step 5).

The knowledge production network

The resulting set of journals can be used for a further description of three dimension of the knowledge production network as mentioned above: nodes, relations and content. We downloaded the records of all articles, reviews, letters, and notes from the journals identified from the SCI and the SSCI. This gives us the required information to map the network of knowledge co-production: institutional addresses, cited references, titles, and abstracts.

The nodes in the network are all organizations that are mentioned as the corporate address of the authors of the publications. As far as possible we homogenized the addresses to a low but more relevant level of aggregation: for example we replaced universities by the relevant departments.

The structure of the network is based on the technique of *bibliographic coupling*. We identify a relation between two institutes if both refer to the same literature. And the more references two institutions share, the stronger is the relation. Why do we use bibliographic coupling? Specialization occurs within a research field, and co-production of knowledge takes place within the smaller scientific community working on this specialty. Knowledge production proceeds over small new knowledge claims which position themselves in terms of how they differ from earlier knowledge claims [6]. We therefore expect that within a specialty researchers refer to and build upon a relatively small set of previous results. This can be traced by ‘bibliographic coupling’.

Finally, the content of the communication is based on an analysis of words in titles and abstracts of the papers. By comparing key word occurrences, we will analyze the different contents communicated in the three networks. Using co-word analysis [17], we attempt to obtain a more fine grained picture of the communicated contents.

The collaboration network

Mode 2 knowledge production is characterized by networked and heterogeneous collaboration. However, the journal output is generally produced by the academic partners only. In order to determine the wider collaborative environment, we decided to take the collaboration within EU funded research as the empirical base. Of course not all externally funded biotechnological research is included in this database, but it gives us the opportunity to study a European communication and collaboration network in which many organizations are involved. In other words, the Cordis database creates the opportunity to analyze the ‘Mode 2 context’.

We will study this ‘project collaboration network’ on two levels. Firstly, the network structure of the EU projects can be studied as such. Secondly, we can start with the institutions in a field which publish in the scholarly journals and determine the wider project collaboration environment of those institutions. The resulting network can be considered as an aggregation of so called ‘personal networks’ of all ‘persons’ in the set: the European research institutes that publish in the relevant academic journals. Of course,

a remaining question is how this research network relates to the rest of biotechnological research.

The nodes of the collaboration networks are the organizations involved in the projects. In the case of biotechnology, these are the projects under the BEP, BAP, Biotech1 and Biotech2 programs. We define the links straightforwardly as project collaborations. Finally, to map the content of the collaboration we use the title, project description, and results (when available) as included in the database.

However, this was not an easy operation, as firstly the database had to be made accessible for analysis, and secondly a lot of homogenizing work had to be done. In order to identify multiple partnerships by the same organization the data underwent an extensive procedure of data standardization². For our purposes, it was necessary to split large organizations (universities, multinationals etc) into smaller entities, in most cases departments. This was necessary not only to correctly identify the organizations and organizational entities that had participated in the projects, but also to be able to link the Cordis data with SCI data on the organizational or departmental level. The data had to be further standardized to eliminate distortions due to differences in spellings, abbreviations and synonyms.³

Studying complex and large networks like the one we are interested in also requires visual inspection. The BibTechMon⁴ software is a flexible tool for analyzing and visualizing large networks in various dimensions. The standardization of the data was also necessary for the use of this software tool. In the analysis presented here the *Jaccard Index* was used to normalize the co-occurrence matrix, and the visualization method was based on a ‘mechanical spring model’ [15]. This enables a transformation into an intuitively readable 2-dimensional map.

The virtual network

Finally, we will investigate the WWW-link structure of the institutions detected in the previous two networks. This results in a third communication network showing the information exchange with a different and most probably larger environment of research and application. The World Wide Web (WWW) is a systematic body of ‘pages’ that contain information, and hyperlinking is the capability that links a specific WWW page to another. Hyperlinks can be analyzed in a similar way to citations [1, 18, 19] and many recent publications have adopted these methods. Despite the fact that university websites

² We had to standardise different versions of the names of organizations (TU Wien, TU Vienna, Technische Universität Wien, Technical University of Vienna, etc), and of places (Copenhagen is spelled in more than ten different ways in Cordis).

³ In cases where there was no information available on the departmental affiliation of participants, the organization was split on geographical criteria such as the region, city or street of location. In order to identify who was actually taking part in the projects, an attempt was made to take the history of the organisation into account specifying as far as possible the changes in the structure of the organisations (mergers, organizational change, etc.).

⁴ The analyses were performed with the bibliometric software BibTechMonTM developed at the Department of Technology Management of the Austrian Research Center Seibersdorf.

are very heterogeneous, and consist of institutional information, teaching related information, individual information, and relatively little research related information, link count metrics for universities and measures of institutional research seem to correlate [20]. The modern, multi-faceted, multi-genre, and partly unregulated university Web site is a challenging new object of study [21, 25]. Nevertheless, hyperlinks are much more varied in use than citations, and existing citation techniques are difficult to generalize to the new medium. Several methodological questions emerge in this context.

First, on which organizational level should we define the node in the hyperlink network: the university, the faculty, or the research group? And how should one decide which WWW site (institution) belongs to the network? The main conclusion of our previous research is that one always finds networks, but that unless the set of included nodes of the network is carefully selected the interpretation generally remains difficult [11]. It also shows that the department and research group level creates the most meaningful networks, as higher level networks are often unclear mixtures of lower level networks [10]. Second, the question has to be resolved of how to operationalize a department on the web. Is this the home page? The whole (sub)domain, including for example the personal pages of the staff? Or something in between? For the moment, we decided to take the home page, plus all pages two levels deep. In this way we avoid focusing on relatively empty homepages, consisting only of links and address information, and at the same time we hope to avoid including personal pages with information about family and hobbies.

The identification of the university sites included in the analysis comes from the University Sites Database, created within the context of the EICSTES project by the CINDOC InternetLab. We have built a crawler to find all the related websites, and the crawler also provided us with the textual content of the various webpages.

In other words, the nodes of the networks are the university department domains identified using the SCI data, and the InternetLab data, as well as the project partners identified using Cordis data, and the web environment consisting of organizations linking to university sites and project partners. The inlinks constitute the relations of the network (which has directed links, in contrast to the previous two networks), and the content of the communication is the textual content of the websites.

Summary

Figure 1 summarizes the approach: we aim at identifying three different communication systems (the systems of co-production, collaboration, and dissemination) in three dimensions (actors, network, content). In this paper we will present some of the first results of the analysis for the biotechnology field, including some comparison with the artificial intelligence field, and information science. The substantial results are intended to demonstrate the usefulness of the method developed to study triple helix as network of networks (projects, journals, web).

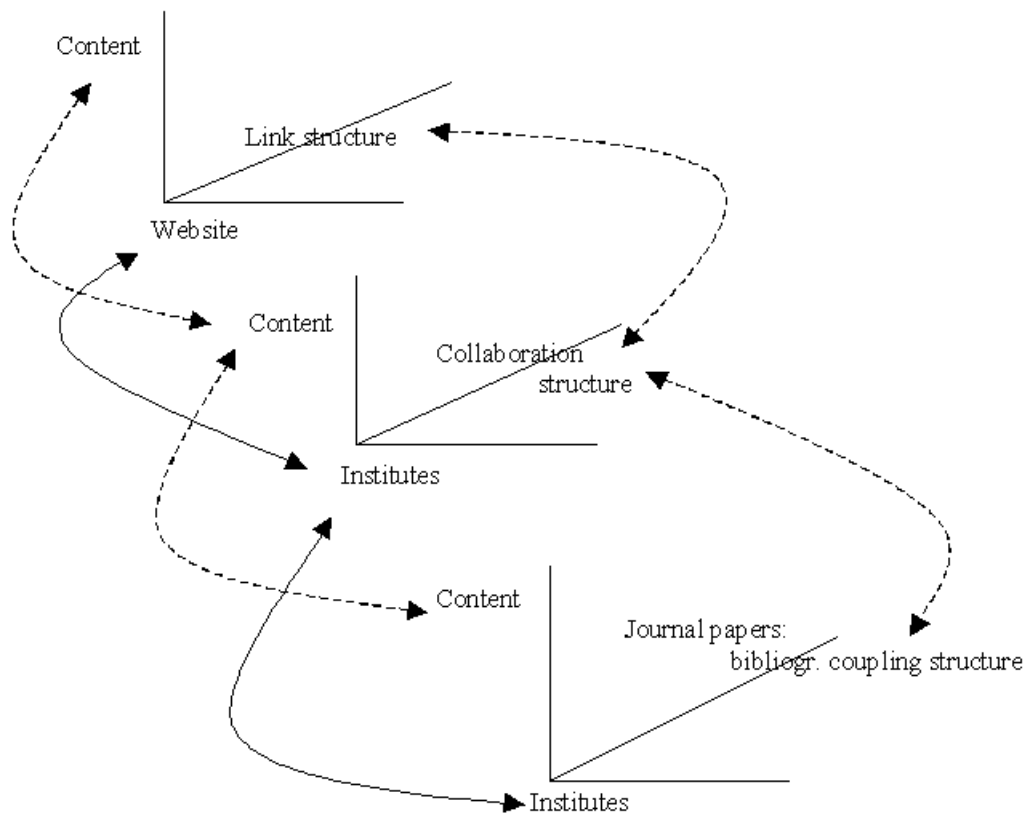


Figure 1. The design of the study

Results

The boundaries of the system

The first field we studied within this project is biotechnology for the year 1998, and we have chosen the journal *Biotechnology and Bioengineering* as the entrance journal. This choice was based on the fact that of all journals with ‘Biotechnology’ in their titles, this one has the highest ‘impact factor’ [17]. Using the method indicated in section 2.1, a journal-journal citation matrix was constructed, with a threshold of 1%. Factor analysis of the citation matrix resulted in seven separate clusters (scientific specialties). The results show *Biotechnology and Bioengineering* was the appropriate choice: the journal has the highest loading on its own factor and therefore it can be considered as the ‘central tendency journal’ in the citing dimension [4]. Analyzing the data for the years 1986, 1992, 1996 and 1998 reveals that the cluster of journals becomes somewhat larger, but rather stable. Information science and artificial intelligence were delineated in the same way, using *JASIST* and *Artificial Intelligence Journal* as entrance journals. The results

were also similar the biotechnology case, also with respect to the fairly stable results over the years [23].

Co-production of knowledge: actors, network, content

The co-production network can be described and analyzed on various levels, such as the research field level (e.g., biotechnology as a whole) and on lower levels of sub-fields. Using the SCI data for the 1998-2000 period on the field level, we found 481 organizations with an EU-address contributing to biotechnology, whereas for AI and information science the respective figures were 240 and 84. The relationships between the organizations were defined in terms of bibliographic coupling, because when two organizations refer to the same literature, they are expected to operate in the same sub-field. Most organizations are of academic origin and most of them contributed to only one or two publications. As an illustration we visualize the biotechnology network in figure 2.

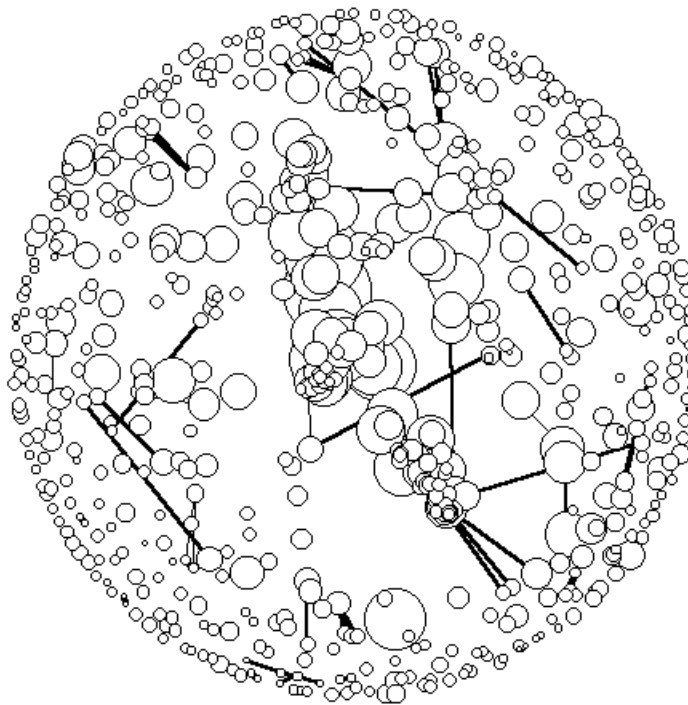


Figure 2. The network of European institutes publishing in the selected Biotechnology journals 1998-2000

Each node (a circle in figure 2) represents an organizational unit that published in one or more of the biotechnology journals. The size of the node represents the number of times the organizational unit is bibliographically coupled through cited references, and this in average increases with the number of publications. Nodes that are in each others vicinity

refer to the same publications. The lines between the nodes represent the links between the nodes, and for readability we plotted only the 200 strongest relations. So nodes which are not linked in the figure may in reality be linked through bibliographical coupling. The number of nodes is 633, and the number of links is 6768. In terms of network analysis the density is low (0.0034) and on average there are 10 links per node. We have reported related analysis of the networks elsewhere [12]. The largest component in biotechnology contains 83% of the nodes, and the ‘longest shortest’ path is 16. In other words, the biotechnology co-production network is linked, but not dense. The results for the two other fields are similar [12]. Although the overall networks are fairly sparse, several parts of the network are much more dense. Further analysis of the denser parts of the network may therefore be useful, as we expect the dense parts to represent specialties within the larger research fields, as is also suggested in figure 3.

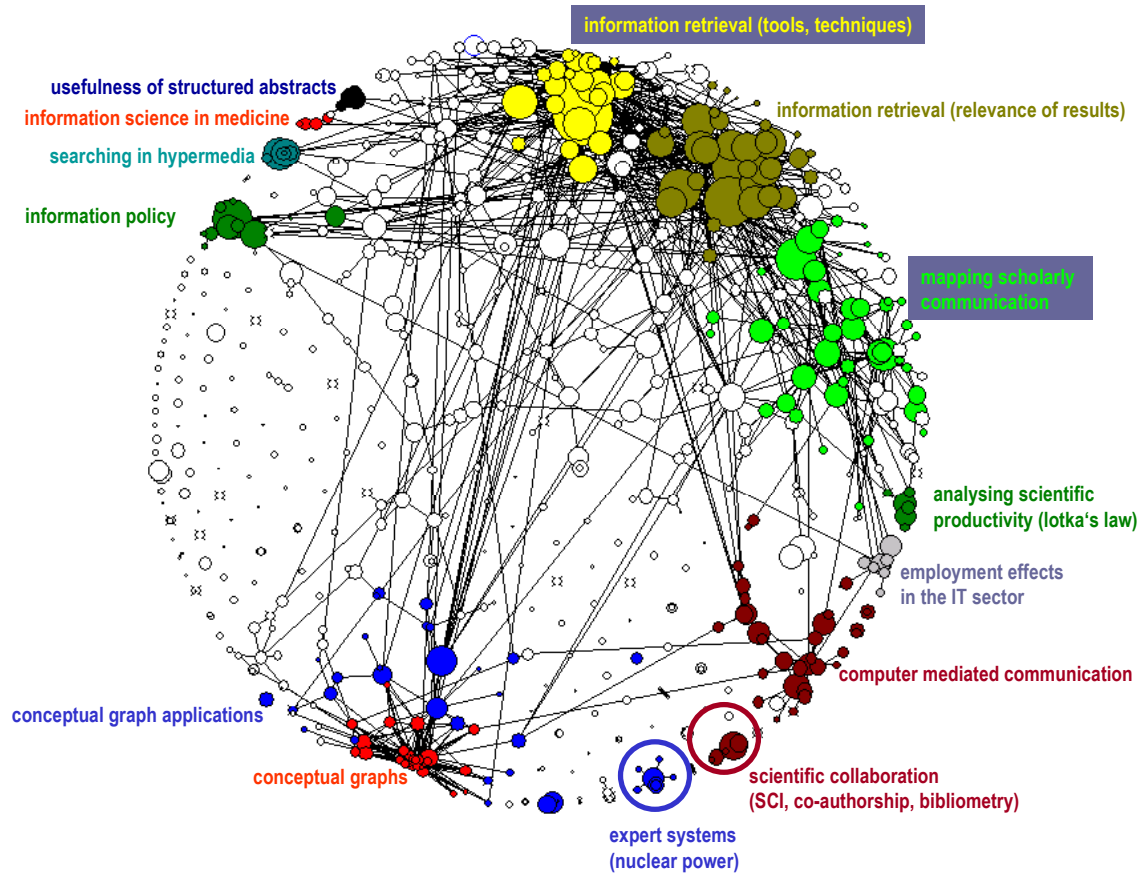


Figure 3: The knowledge map of Information Science

The content of the communication is represented by key words obtained from the abstracts and titles. We generated them using a procedure based on the context sensitive longest-match principle and a phrase recognition algorithm [26]. This automatic indexing

module was applied to the abstract of each record. In this way, a word frequency list is obtained which can be compared with the Cordis data and the Web data. The SCI-based word list contains mainly technical terms. Using co-word occurrences in the biotechnology papers, we tried to cluster the papers. The first analysis indicates that in this way a research field can be split into thematic sub-fields in a meaningful way. Figure 3 gives a map of a network of papers based on similarity (Jaccard) in terms of co-word patterns, in this case for information science. The closer the nodes, the stronger the relation in terms of co-word occurrence. The larger the nodes, the more relevant words are used in the paper. The lines represent the links between more remote nodes. The labels in the figure indicate the topics of the clusters of papers we found.

Collaboration networks

The lists of institutes derived from the SCI-data constitute the core of the scientific communities in the three fields we study in this paper. This provides the starting point for investigating the cooperation between universities, companies and governmental organizations. The relevant domain within the Cordis database is defined as the list of (commercial, governmental and research) organizations that have participated in projects with the organizations with a publication in the SCI domain. This enables us to define the economical, scientific and political context of knowledge production. Here we have also disaggregated the institutions as much as possible.

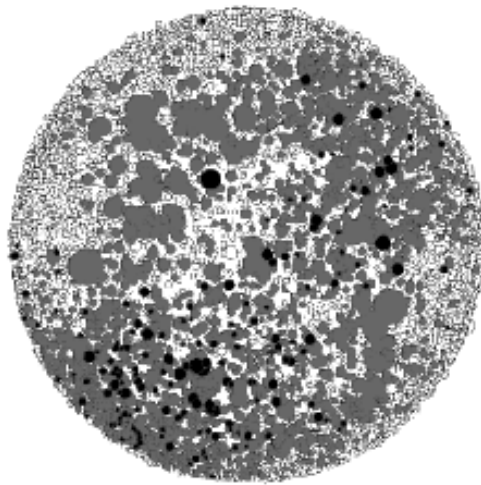
The database covers a period from 1984 to 2000, the period of the first five Framework Programs, with 26968 projects good for 97590 partnerships. Correcting the data for multiple partnerships by the same organization(al unit) results in 31773 unique participants. Out of these we identified organizations and organizational entities that were involved in projects associated to the three disciplines Biotechnology, AI and Information Science.⁵ For Biotechnology we identified 2063 projects in the five FW programs until December 2000, for Information Science 5276 and for Artificial Intelligence 253 projects. In AI, these projects involve 1015 participants on the level of organizations and organizational entities, in Information Science 11095 and in Biotechnology 4544. For these three fields we mapped the standardized organization names from Cordis on the organizations in the SCI. For biotechnology, for example, this method found some 481 organizations publishing in biotechnology journals that also participated in Cordis biotechnology projects.

In figure 4 we visualize the network of the core set of actors identified in Cordis and SCI, using BibTechMon. The figure shows all participants in biotechnology projects of the five FW programs. A participant is represented by a node. The bigger the node, the more projects the participant was involved in. The more frequently participants collaborate in projects, the smaller the distance between the corresponding nodes. The

⁵ It should be noted that we selected the relatively restricted field of information science in the SCI, but that the projects in Cordis cover the much larger field of all information technology related projects. However, as explained earlier, we define the fields under study starting from the core actors identified in the journals, and therefore we include only a small part of the projects in our analysis (see table 1).

organizations in the journal system of Biotechnology that also participate in the EU projects are colored red. Their project partners who were not in the journal system are colored in blue. The rest of the nodes represent participants in all other biotechnology projects in the five Framework programs.

The figure suggests that the number of scholarly publishing organizations is small compared to the number of non-publishers. Furthermore, participation of ‘publishing’ organizations is generally limited to a small number of projects (as indicated by the small size of the nodes). If we look at the Biotech2 program only, we have in total 1645 different participating organizations, with 8907 links. This results in a low density of the network (0.0001), and on average every node links to eight other nodes.



Black = SCI; Grey = environment in Cordis; White = Other Cordis

Figure 4. Biotechnology Organizations

Finally, the communicated content in the European research projects was represented by the same key word analysis described above, but now using project titles and descriptions in the Cordis database. The word frequency list of the European research projects reflects the policy dimension, the technical content, and also the application context of the research. We will discuss this further in section 4.

Information dissemination in electronic networks

The identified universities, companies and governmental organizations provide the starting point for link and content analysis on the web. The first step is finding whether these institutions are represented on the WWW. From the 765 core organizations operating in Biotechnology, AI and Information Science, 601 homepages could be obtained from the EICSTES database maintained at CINDOC. The remaining 164 organizations had no (functioning) homepages or no longer existed. The hyperlinks

constitute the relations between the nodes, and we take the textual parts of the websites as the communicated content. In other words, other objects such as images, databases and so on, are not included in this phase of the analysis.

Using our crawler, we downloaded the content and the hyperlinks of web pages. Each organization was crawled two levels deep, starting from the homepage. In other words, each organization is represented by the aggregated set of web pages, and all links to all these pages are included in the mapping of the network architecture. The same holds for the content: all words of all the pages were downloaded for the analysis. Finally, co-link analysis was used to determine the set of organizations in the direct environment of this core list: all websites of the organizations that receive inlinks from more than one of the core organizations were included.

The first observation is the very weak structure of the network. Many of the core institutions are not linked to each other at all. Figure 5 shows this for the biotechnology field. A node represents a website and the size of the circle is an indicator of the amount of hyperlinks to different web pages within the domain. The hyperlink network of outlink relations shows an almost unrelated set of organizations. The white dots are the core biotechnology institutions as defined above. Only sixteen of these organizations are related to each other through hyperlinks. Close inspection shows that these links are based on nationality and language.

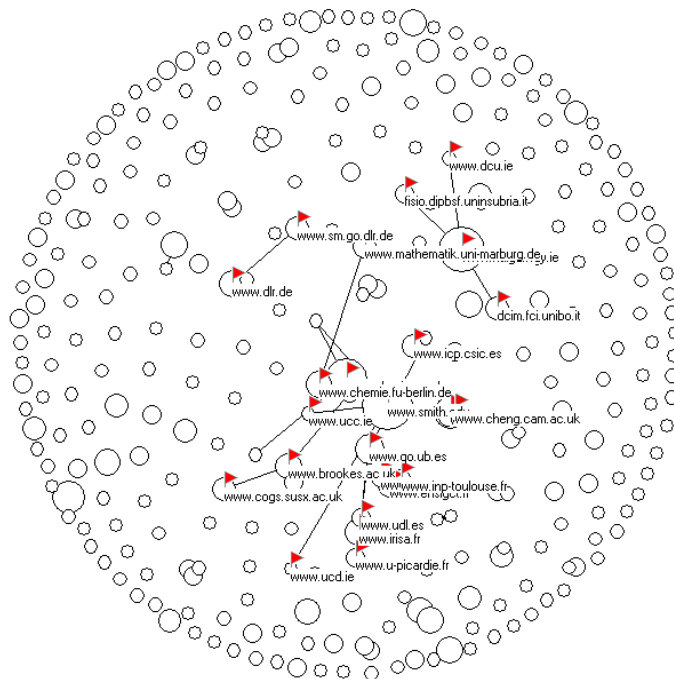


Figure 5. Network of hyperlinks in Biotechnology

In addition, we did a co-link analysis on the set of organizations in the three fields together. The few (out)linked organizations do not represent research collaborations, as the emerging clusters are small and heterogeneous. Furthermore, they are not interlinked.

This low level of research content is also clear in the analysis of the words used on the websites. We analyzed the content of the websites in a manner similar to the content analysis of the paper abstracts and project descriptions. The results of this analysis indicate that the websites are very much oriented at general and educational information. Among the words occurring most often on the web pages of the core organizations are the words ‘welcome’, ‘staff’, ‘information’, ‘search’, ‘school’ and ‘department’. The frequently used words on the web indicate a different audience than the communication in the journal system and in the research projects.

There is also no indication of a content based thematic clustering of organizational websites. How can we explain this? We have the impression that the links to institutions’ homepages have a more ‘strategic’ function for the linking organizations, demonstrating that they consider themselves in the relevant environment of the linked organization. On the other hand, out-links pointing to deeper-level pages probably indicate a common topic of interest, that is, a thematic relation. In other words, analyzing the websites two levels deep may be insufficient for finding this type of thematic relation. In further research we will test this, through a comparison of link patterns and word patterns that emerge from crawling sites at different depths.

Table 1: Overlap between the co-production network and the collaboration network

Research field:	European organisations/departments in		
	SCI	Cordis	Both
Artificial Intelligence	1179	1015	240
Information science	624	11095	84
Biotechnology	1843	4544	481

The Network of Networks: Comparing the communication systems.

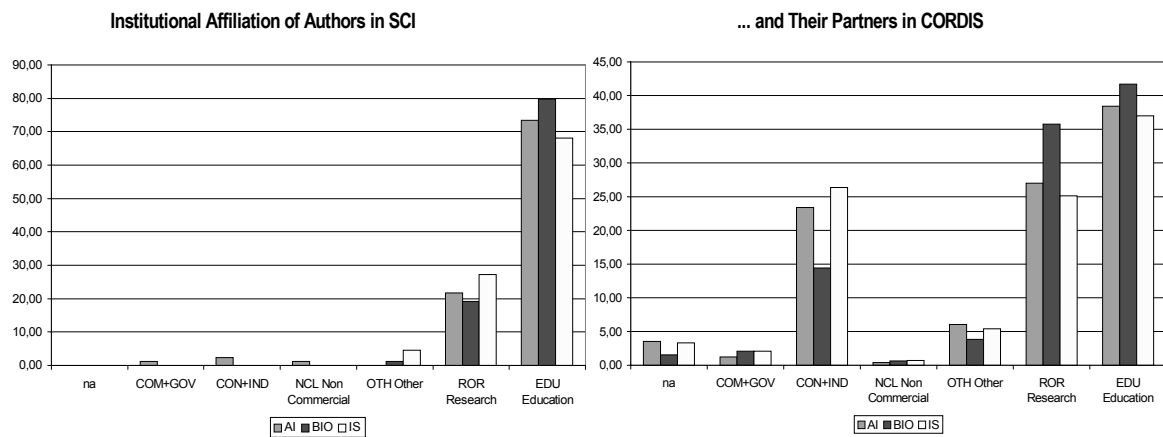
In the previous sections we analyzed the three communication networks separately. However, it is also necessary to compare the networks, as they may have complementary but different actors, network architectures and content.

The actors

Are the actors in the three networks different, and if so, what does this teaches us? At this moment we will only compare co-production and collaboration networks. First of all, in all the three fields, the overlap between organizations in the two networks is relatively small, indicating that the two networks are substantially different. Table 1 shows the

numbers, and the overlap represents 26.1% of the organizations in the SCI, and 10.6% of the organizations in Cordis. For information science and AI, the figures are in the same order of magnitude.

We also compared the networks in terms of the type of organizations involved, and we distinguished between five categories: European and national governmental organizations; consultancy and industry; non commercial organizations; contract research organizations; and universities. Figure 6 presents the distribution of organizations in the SCI and of their partners in Cordis over the categories mentioned. As expected, the communication system of scientific publications is almost exclusively the domain of universities and contract research organizations. The research projects carried out within the context of European research programs show a larger variety of participants, and besides universities and contract research organizations, industrial and governmental organizations also play a significant role. Nevertheless, the universities and contract research organizations are also the largest categories here. Finally, the participation of industrial partners is quite different in the three disciplines: in Artificial Intelligence and Information Science it is substantially higher than in Biotechnology.



na = Not yet identified; COM+GOV = European and national governmental organizations; CON+IND = Consultancy and industry; NCL = Non commercial; ROR = Contract research organizations; EDU = Universities ; * = Only for participants in more than three projects.

Figure 6: The distribution of institutional affiliations*

The relational networks

Neither the scientific network based on the bibliographic coupling nor the triple helix network of project co-operations are very dense. Nevertheless the analysis of the resulting graphs of the networks does show dense spots within the otherwise sparsely connected networks. This suggests that a further breakdown of the networks into specialized sub-fields is needed in order to obtain better insight in the networks of the triple helix systems. The map of the ‘cognitive fine structure’ of information science indicates that the data can be used for this.

The network of outgoing hyperlinks shows an unrelated set of organizations. The clusters of websites are small and heterogeneous and are not interlinked. This suggests that hyperlink networks function in the context of knowledge dissemination that is only loosely related to the co-production of knowledge (in scientific fields) and the collaboration networks in research and application (in research projects). The Internet seems to be used merely for communications with users of the knowledge resources in a predominantly local context. This is in line with other research [10].

What is communicated?

The differences between actors and network architecture already suggest that the various communication media have different functions. The analysis of the communicated content supports this further. A first comparison of the content communicated in the three media mentioned provides some interesting results. The content of the communicated messages can be operationalized in terms of word-frequencies. The analysis of word patterns is far from straightforward [16]. However, comparison of key word frequencies between large sets of documents may be used to compare the overlap and differences in content [22]. We use this method to test whether the different media have a different functions, and therefore communicate different content.

The following procedure was followed. Using BibTechMon (for the publications, and for the project descriptions) and our Crawler (for the WWW pages), we produced the three lists of words and phrases. These were combined to a gross list of 4902 words. In order to focus on the most relevant words, we first reduced the words to a single form, and then introduced a threshold of an occurrence of at least 10 times in one of the three media. In this way a set of 385 words remained, but unevenly distributed over the three media. The journal papers account for the largest share of this word set, as the number of papers exceeds the number of projects and exceeds the number of websites much further.⁶ Therefore we raised the threshold for the journal papers and lowered the threshold for WWW sites to have similar numbers (128) of most frequent words for the three media, and this reduced the gross list to 257 words. In order to compare the content in the three media we calculated the correlations between the word-frequencies of these words (385 and 257 respectively) in the three media. The following table shows the results, which are hardly influenced by changing the thresholds.

Table 2: Correlations between word frequencies

based on:	Biotechnology: correlation between		
	Articles & Projects	Articles & Web-pages	Projects & Web-pages
257 words	0.24	-0.07	0.29
385 words	0.28	-0.02	0.32

⁶ The frequency distribution of the words from the websites is the flattest distribution, and the frequency distribution of the journal article words is the most skewed. These differences indicate the different levels of codification in the three media.

The results clearly show the differences between the media. The difference between the journal papers and the websites is the largest, and the content of the projects is situated in between the websites and the journal articles. Whereas the correlation between the projects and the two other media is moderate (around 0.28), there is no correlation between the word patterns in the web-pages and the articles. This underlines the hypothesis that the three media have different functions in the maintenance of the triple helix networks. Inspection of the word lists shows that the journal papers contain many technical scientific words, which are less evident in the projects descriptions and absent on the websites. The projects contain less technical terms, and more words used for the description of research in non-technical language, and for the explanation of the social and economic relevance, as one would expect in project proposals and descriptions. This type of words is also more common on the web-pages. The content downloaded from the web indicates that much information is presented for educational purposes.

5 Conclusion and discussion

In this paper we proposed a strategy for researching the complex communication network that characterizes the triple helix of university-industry-government relations. To understand these complex, multilevel and heterogeneous communication networks we need to have the appropriate data, preferably micro data about the position and activity of organizations within the various networks. As we focused on formal communication using scholarly journals, on communication and collaboration in research projects, and on information dissemination over the world wide web, there was a need for creating an integrated dataset. Therefore we used ISI data, Cordis data and Webdata, the latter collected with web crawlers. The main effort in the creation of the dataset was in the homogenizing of the names and addresses of the organizations in the SCI and in Cordis. The overlapping set constituted the population under study. This set was determined for three so called Mode 2 research fields. In the processing of the organization names, the lowest possible organizational unit was identified, rather than including entire universities, for example, as these consist of many labs active in completely different fields [11]. After this was done, we identified the websites of the organizations involved. The whole process proved to be a time consuming effort, covering the whole Cordis database, and part of the ISI data (1998-2000), both for the three fields studied in this paper.

From a methodological point of view, the strategy of mapping various databases by mapping the actors is promising. We could identify the network structures, the types of actors involved, and the communicated content in the three networks. Comparing the networks, differences in structure and function became obvious. The analysis shows that the three networks consist of partly overlapping nodes, indicating that different audiences are served by the different networks. The analysis of the content communicated in the three networks underlined these findings. We found that even Mode 2 knowledge production networks remain to a large extent academic. The collaboration networks of the EU funded research consists of a more heterogeneous set of actors. Not only universities

and other research institutes participate in these projects, but also companies and governmental bodies. Finally, the structure of the hyperlink networks seems to be most diverse, and this suggests that the Internet serves as the *social and public interface* for research organizations. Word use, and the fact that many of the hyperlinks are between sites with the same language, suggest that the Internet is used for information dissemination in a predominantly local context [10].

Nevertheless, many questions remain unanswered. For example, we analyzed the networks at the level of broad research fields, such as biotechnology. In reality knowledge production and project collaborations are situated within specific sub-specialties. It would be interesting to analyze these more small scale communication networks, using the same method. What type of organizations (governmental, industrial, academic) play what kind of role (users, producers, suppliers, commissioners) in the various networks, and is this different between sectors and countries, for example? Another question is whether the type of research is different between the more complex Mode 2 networks and more traditional Mode 1 networks in the same research field. And, in this study each organization was crawled two levels deep, starting from the homepage. The results suggest that more specific thematic hyperlink structures may be found on deeper levels of websites. More generally, several methodological questions still have to be answered, such as the appropriate way of measuring and mapping hyperlink networks between knowledge producing (and using) organizations.

Acknowledgements

The research underlying this paper has been partially funded by the European Commission, grant IST-1999-20350, the EICSTES project. Partners in the project are CINDOC (Spain), ARCS (Austria), DTI (Denmark), INIST-CNRS (France), NIWI-KNAW (Netherlands), and the University of Surrey (UK). This paper is partly based on a deliverable from the project [12]. The authors would like to thank Alexander Kopcsa and Manolis Mavrikakis for their technical support, Isidro Aguillo, Mar Ananos, and Xavier Polanco for their support in gathering and organizing the data, and Bernhard Dachs for some of the analysis. We also would like to thank two anonymous reviewers, Loet Leydesdorff and Eleftheria Vasileiadou for their useful comments on an earlier draft of this paper. Part of the research was carried out within the former Social Informatics Group, University of Amsterdam, Department SWI.

References

1. Björneborn, Lennart and Peter Ingwersen. Perspectives of webometrics. *Scientometrics*, 2001, 50(1), 65-82.
2. Borgman, C.L. (ed.), *Scholarly communication and bibliometrics*. Newbury Park (CA): Sage, 1990

3. Borgman, C.L., J. Furner, Scholarly communication and bibliometrics. In B. Cronis (ed.), *Annual Review of Information Science and technology* 36 (2002) pp 3-72
4. Cozzens, S.E., & Leydesdorff, L. (1993). 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, & R.W.J. Tijssen (Eds.), *Science and technology in a policy context* (pp. 219-233). Leiden: DSWO/Leiden University Press
5. Etzkowitz, Henry & Loet Leydesdorff, *The Dynamics of Innovation: From National Systems and 'Mode 2' to a Triple Helix of University-Industry-Government Relations*, *Research Policy* 29(2) (2000) 109-123.
6. Gibbons, M., Limoges, C., Nowotny, H., Schwartzman, S., Scott, P., and Trow, M. (1994), *"The New Production of Knowledge"*, Sage, London
7. Gläser, Jochen, *Scientific specialties as the missing link between scientometrics and the sociology of science*. In 8th ISSI proceedings, Sydney: UNSW, 2001, 191-210.
8. Gläser, Jochen, & Grit Laudel, *Integrating Scientometric Indicators into Sociological Studies: Methodical and Methodological Problems* *Scientometrics* 52 (2001) 3, p. 411-434
9. Fujigaki, Yuko, *Filling the gap between discussions on science and scientists' everyday activity: applying the autopoiesis system theory to scientific knowledge*. *Social Science Information* 37(1) pp 5-22.
10. Heimeriks, Gaston & Peter Van den Besselaar, (2002) *The role of electronic communications in research - a case study*. Paper presented at the EASST conference York, 2002.
11. Heimeriks, Gaston & Peter Van den Besselaar, (forthcoming) *The level of aggregation in weblink studies*.
12. Heimeriks, Gaston, Peter Van den Besselaar, Doris Schartinger, Bernhard Dachs, Alexander Kopcsa, Maria del Mar Ananos Sanchez, *European Indicators, Cyberspace and the Science-Technology-Economy System*. ECSTES deliverable 5.2. Amsterdam 2002.
13. Hurd, Julie M., (2001), (ed.) *Perspectives issue on the changing communication system of science*. *Journal of the American Society for Information Science* 50 (2001), 1276-1337.
14. Kling, R. & G. McKim, *Not just a matter of time: Field differences and the shaping of electronic media in supporting scientific communication*. *Journal of the American Society for Information Science* 50 (2001), 1306-1320.
15. Kopcsa A., Schiebel E. (1998): *Science and technology mapping: A new iteration model for representing multidimensional relationships*, *Journal of the American Society for Information Science*, Vol. 49 (2000) pp 7-17.
16. Leydesdorff, L. and Heimeriks, G. (2001) *The Self-Organization of the European Information Society: The case of "Biotechnology"*, *Journal of the American Society for Information Science and Technology* 52 pp.1262-1274.
17. Rip, A., J.P. Courtial, *Co-Word Maps of Biotechnology: An Example of Cognitive Scientometrics*, *Scientometrics* 6 (1984) pp. 381-400.
18. Rousseau, Ronald (1997) *Sitations: an exploratory study*. In: *Cybermetrics*, 1 (1997) 1. www.cindoc.csic.es/cybermetrics/articles/v1i1p1.html

19. Thelwall, M. (2002a). Evidence for the existence of geographic trends in university Web site interlinking. *Journal of Documentation* , 58, 563 – 574
20. Thelwall, M. (2002b). A comparison of sources of links for academic Web Impact Factor calculations. *Journal of Documentation* , 58, 60-72.
21. Thelwall, M. (2002c). Conceptualizing documentation on the Web: an evaluation of different heuristic-based models for counting links between university web sites, *Journal of the American Society for Information Science and Technology*, 53(12), 995-1005.
22. Van den Besselaar, Peter, The cognitive and the social structure of Science and Technology Studies. *Scientometrics* **51** (2001) pp. 441-460.
23. Van den Besselaar, P. and Heimeriks, G. Disciplinary and interdisciplinary identities. Forthcoming
24. Van den Besselaar, P. and Leydesdorff, L. (1996) Mapping change in scientific specialties; a scientometric case study of the development of Artificial Intelligence. *Journal of the American Society for Information Science* **47**, 5.
25. Vaughan, L. & Thelwall, M. (2003). Scholarly use of the Web: What are the key inducers of links to journal Web sites? *Journal of the American Society for Information Science and Technology*, 54(1), 29-38.
26. Widhalm C., et al. (1999): Konzeptive Entwicklung eines Einlesesystems und einer Strategie zur automatischen Schlagwortgenerierung, OEFZS-S-0051. Seibersdorff: ARCS, 1999.