Learned Publishing (2005) **18**, 165–176

A lifecycle model of the scientific communication process

Bo-Christer Björk
Swedish school of Economics and Business Administration, Helsinki

ABSTRACT: A formal model of the scientific communication process is presented using the IDEF0 notation. The model provides a roadmap for policy discussions and further research. In comparison to earlier models it is more detailed and hierarchical, and includes more modelling constructs. It includes the whole communication value chain, from initial research to the assimilation of research results to everyday practice. Although the model treats both informal and formal communication, as well as the publishing of data, its focus is on modelling the publishing and indexing of traditional peer-reviewed journal articles, and finding and retrieving them. New developments enabled by the Internet, such as open access journals and e-print repositories, are also included.

Bo-Christer Björk

Introduction

The speed of progress in science has always been strongly dependent on how efficiently scientists can communicate their results to their peers and to lay-persons willing to implement these results in new technologies and practices. For centuries the communication chain was very slow, relying, for instance, on tedious copying of scientific texts by hand. Communication was to a large extent local, taking place orally in the few universities then existing. The invention of the printed press was a major step forward and enabled the cost-effective reproduction of monographs, as well as the establishment of more systematic forms of communication, in the form of regularly appearing scholarly journals.

During the 20th century science became recognized as the major driver for economic development and the number of scientists increased dramatically. In addition to journals and monographs, conferences became an important form of communication, due to the increased possibilities for travel. During the latter half of the 20th century information technology (IT) has had a profound impact on the scientific publishing process. First it enabled the setting up of electronic databases of bibliographic data, which greatly facilitated the search for relevant publications. Secondly word processing has meant increased efficiency in both the writing of manuscripts and in the handling of them during the printing process.

But the most dramatic effects on the overall process have occurred during the last 15 years or so through the Internet and the World Wide Web in particular. It is perhaps no coincidence that scientists have been among the pioneers in taking these technologies into use. Science is by its nature both global and collaborative and the sorts of networking capabilities now offered are perfectly aligned with the *many of the players have strong vested interests in preserving the status quo*

open knowledge-sharing goals of the academic community.

A large part of this communication process takes place as a so-called peer production process. Scientists usually demand no monetary rewards for sharing their results (in contrast to producers of music or popular literature). What they are interested in, in addition to advancing science itself, is building up relationships with other scientists, or building a reputation which enables them to advance in their careers, get better grants, etc. The more others read their publications and cite them, the better. Unfortunately the legal, economic and behavioural infrastructure that underpins this communication process was shaped in the decades before the era of the Internet and has now become something of a straitjacket that hinders progress.[1]

A substantial part of the scientific communication process has traditionally taken place within the academic community itself. Some parts, however – mainly in the print world – have been handled by commercial intermediaries, because they have been able to do a more cost-effective job or because the academic community (primarily the learned societies) has been unable or unwilling to organize the work. What has happened recently is that these intermediaries have migrated on a large scale to electronic delivery, but have not essentially changed their subscription-based business models. This means that the overall process from a cost and communications perspective is rather sub-optimal. This change has come about because of the peculiarities of this market, which very strongly favours established players and monopolies, and poses high barriers to entry for new types of actors.[2]

Due to the currently rapidly changing environment (i.e. the move from paper to electronic delivery, new open access business models for journals and parallel publishing in e-print repositories) there is a demand for scientifically researched knowledge about the status and attributes of the scientific publishing process. In the discussion in the scientific and popular press, as well as on email listservers, highly varied opinions on the costs of journal publishing, as well as on the effects of different strategies concerning open access, are put forward,[3] and the situation today might be characterized as confused. Also many of the players have strong vested interests in preserving the status quo, or are fervent advocates of new ideologically flavoured strategies, and this colours the discussion.

Although a number of empirical studies of the effects of switching to electronic/and or open access have been made, it is difficult to compare the results of such studies since they are often measuring different aspects of the overall process. Thus there is a clear need for models that structure the overall scientific communication process and can be used as a basis for comparing and integrating the results of different studies.

Scientific communication viewed as a global information system

One interesting aspect of the scientific communication process is that it is a global, interconnected information system. The academic discipline which studies the development and use of information systems in companies and organizations is usually called information systems science. Another related but separate discipline is information studies.

Information systems science typically studies IT systems that commercial organizations build to support activities such as production, sales, logistics, accounting and management. The systems can also span different organizations or be interfaced to customers (i.e. e-commerce systems). These systems are usually purposefully planned and built in a top-down fashion. A good example is provided by the comprehensive enterprise resource planning (ERP) systems that large companies build for themselves. In this respect the scientific communication system, and the IT support it uses, is different, because it has grown in an organic way over decades, through the integration of tools produced by a large number of independent players in a non-hierarchical fashion. Nobody owns or has control of the scientific communication system, just like the Internet. Nevertheless it can be perceived as one global special-purpose information system.

An important aspect of large integrated ITsystems in corporations is that they fulfil multiple functions. Firstly they support transactions, such as registering and controlling sales in an e-commerce setting. Secondly they provide management with a basis for decision support by providing aggregate information based on often huge amounts of low-level transactions.[4] The quarterly accounts of large companies are a good example. The scientific communication process also includes two kinds of functions. The primary one is, of course, to help in communicating interesting research results to interested recipients. The second is to provide decision support to research administrations to help in deciding about research grants, professorial appointments, etc.

There are several stages in the development of information systems, including requirements analysis, design, programming and implementation. In the early stages formal modelling methods, usually supported by graphical tools, are typically used. Methods used include data flow models, semantic data models and object models.[5] In his book on scientific publishing and knowledge sharing, Hars[6] includes example diagrams using both flowcharts and object models. A significant benefit of using some of these techniques is that they are supported by IT-based modelling tools, and can be used as a basis for more detailed design and programming in an integrated fashion.

Despite the fact that the scientific communication process has not been designed as a whole but rather has evolved, it might be useful to model it using some suitable formal process modelling technique. Using a formalized tool helps in particular in communicating about the process and in organizing data and knowledge about it. The technique chosen in this work is called IDEF0. The traditional use of IDEF0 models has been in illuminating current and alternative processes in business process re-engineering projects, typically focusing on the design and manufacture of industrial products such as submarines or buildings. The choice of IDEF0 was partly a matter of convenience; the author is well familiar with the method from previous business process re-engineering research.[7]

The IDEF0 modelling methodology

The main concepts of the method are the activity and the flow.[8] Activities are shown as rectangles and their names start with verbs. Flows are represented by arrows and the names are nouns. A flow can be either an input, output, control or mechanism. An input represents something that is consumed in an activity to produce an output. Typical inputs could be raw materials, energy, human labour, but also information when the purpose of the activity is to transform the information. Outputs can be reused as inputs to further activities, and feedback loops are possible. The execution of activities is guided by controls. Outputs that take the form of information can also be used as controls. Mechanisms that point at activities from below are persons, organizations, machines, software, etc., which carry out the activities. The presentation of the IDEF0 diagrams is hierarchical such that diagrams on lower levels are more detailed than those at higher ones. For this modelling exercise a particular tool called BPwin has been used for making and editing the IDEF0 model. Compared to a simple drafting tool, BPwin enhances the speed and consistency of the modelling work, especially for larger models and when changes are needed.

the scientific communication process has not been designed as a whole but rather has evolved

Scope of the SCLC-Model

The overall aim of the modelling work undertaken here was to understand the scientific communication process and how it has been affected by the Internet, in order to provide a basis for a cost and performance analysis of various alternative ways of organizing it. The model can also work as a roadmap for positioning various new initiatives, such as e-print repositories and harvesting tools, within the overall system. Although the current model also includes scientific communication in general, the emphasis is on publishing of traditional journal articles. The model explicitly includes the activities of all the major stakeholders in the overall process, including the activities of the *researchers* who perform the research and write the publications, the *publishers* who manage and carry out the actual publication process, the *editors* and *reviewers* who assure

Table 1 Hierarchical breakdown of the SCLC-model

- Context diagram
  - Do research
    - Perform the research
    - Communicate the knowledge
      - Communicate results informally
      - Publish results
        - Publish textual account of results
          - Write manuscript
          - Publish the results
            - Publish as book
            - Publish as monograph
            - Publish as conference paper
            - Publish as scholarly journal article
              - Do general publisher's activities
              - Do general journal activities
              - Process article
                - Review the manuscript
                - Copyedit manuscript
                - Queue for publishing
                - Negotiate copyright and/or author charges
                - Publish article
        - Publish data and models
      - Facilitate dissemination and archiving
        - Facilitate retrieval of publication
          - Make manuscript or copy of publication available openly on the web
          - Bundle publications from different sources into electronic services
          - Integrate metadata into search services
        - Facilitate retrieval inside reader's organization
        - Preserve publication
      - Study the results
        - Find out about publication
          - Search for publication
          - Be alerted to publication
        - Retrieve publication
        - Read publication
    - Apply the knowledge
    - Evaluate the research or researcher

the quality, the *libraries* that help archiving and in providing access to the publications, the *bibliographic services* that facilitate the identification and retrieval of publications, and the *readers* who search for and retrieve the publications.

The current version of the model has some limitations, which should be kept in mind. Its main emphasis is on the publication and dissemination of research results in the form of publications that in the end can be printed out and studied on paper (irrespective of whether the publications are distributed on paper or electronically). Thus forms of communication such as oral communication, unstructured use of email and multimedia, which all are essential parts of the scientific knowledge management process, as well as publishing of data and models, are only shown on a high level of abstraction in the model. Details could be added at a later stage, but would also add to the complexity of the model.

In the model the central unit of observation is the single publication, how it is written, edited, printed, distributed, archived, retrieved and read, and how eventually it may affect practice. The scope is thus the full lifecycle of the publication and the activities of reading it, which also is reflected in the name chosen for the model. This means in practice that most of the activities take place 5–10 years after the initial writing of the manuscript, but in some cases there will be a demand to access the results even decades later.

The model depicts publishing and value-added services using both paper and electronic formats in an integrated way. Pure electronic or pure paper-based publishing could be described by subsets of the model. The same goes for free publishing on the web ('open access'), which resembles traditional publishing, but where certain activities such as negotiating, keeping track of and invoicing subscriptions can be almost entirely left out.

The model includes some activities that would be typical for a scientific publisher publishing several journals, allowing for economies of scale. The activities of single-journal publishers could be described by a subset. The reason for including activities such as the general activities of a publisher is that these significantly influence the cost of running individual journals in the form of the general overhead costs that publishers add to the subscription prices.

Analysing the whole process in this way should help in highlighting how different actors provide added value to the end customers at each stage. It is therefore close in spirit to the concept of value-chain analysis as defined by Porter.[9] In the long run the customers (authors and readers) will decide on which business models prevail based on how much added value different intermediaries, such as OA journals, provide them.

Overall organization of the model

The current version of the SCLC-model includes 26 separate diagrams, arranged in a hierarchy up to seven levels deep. There are typically three or four activity boxes on each diagram, although there are a couple of diagrams with more activities and some with only two. Altogether, there are 80 activity boxes and around 250 labelled arrows. The overall hierarchical breakdown of the model is shown in Table 1.

Only the separate diagrams are shown in Table 1. Some diagrams are further broken down into separate activities, but this is not shown. Note that this version of the model is the third draft and that the model is continuously evolving based on the feedback received. Due to the enlarged scope, the model has been renamed the Scientific *Communication* Life Cycle Model. The earlier published version was called the Scientific *Publishing* Life Cycle Model. A working report describing the earlier model has been posted on the website of the SciX project.[10] In addition, a conference paper[11] and a journal article[12] discussing parts of it have been published.

Model walk-through

Due to the space restrictions of this article only some central parts of the model can be explained here and the diagrams shown. Interested readers are strongly adviced to download the full model description (a 35-page PDF file) from the OACS project website.[13]

In the following the diagrams that are shown will be identified by their hierarchical codes according to the IDEF0 standard notation, rather than consecutive figure numbers. The diagram A0, *Do research, communicate, study and implement the results*, is crucial for understanding the lifecycle view adopted in this modelling effort (Figure 1). The whole life-cycle is seen as consisting of four separate stages. The *Perform the research* stage is probably the most expensive part, usually consisting of several hundred hours of work effort per resulting publication, but the one least directly affected by the effects of the Internet (at least directly; indirectly the effect can be substantial in terms of better quality of the research due to better access to existing knowledge). The *Communicate the knowledge* activity has been

in the long run the customers will decide on which business models prevail

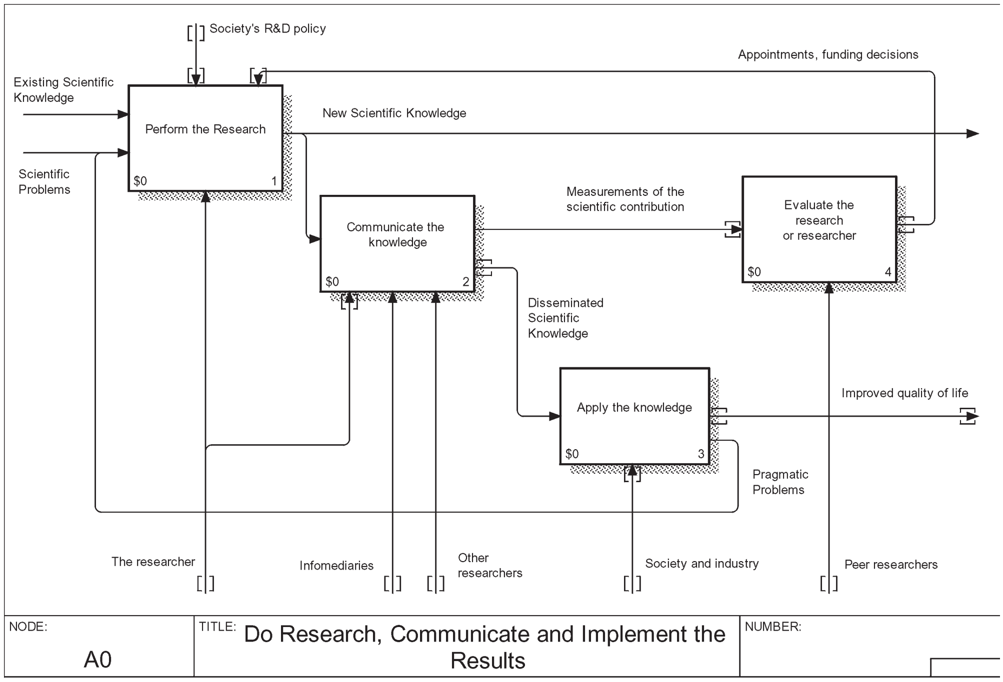

Figure 1. Diagram A0: do research, communicate and implement the results.

profoundly affected by the Internet and is the main subject of this modelling effort. The downstream activity *Apply the knowledge* is important in order to achieve the goal of improved quality of life, but is here mainly included for illustrative purposes.

a decision support system for university administrations, granting agencies, etc

One major change compared to earlier versions of the model is that a separate activity, *Evaluate the research or researcher*, has been included. This is because the system fulfils two functions; one is to communicate the knowledge as efficiently as possible. The other is to act as a decision support system for university administrations, granting agencies, etc. This latter function has important repercussions on how researchers communicate, in particular in which journals they publish.

The *Communicate the knowledge* activity, diagram A2, is split up into an informal and a formal process. Informal communication is carried out in the form of oral presentations of all sorts (person-to-person meetings, conference presentations) as well as email messages, whereas formal communication (*Publish results*) relies on written texts and on quality control by peers (Figure 2). The *Facilitate dissemination and archiving* activity describes activities carried out by a large number of organizations, IT tools, etc., that help readers find out about and retrieve publications of interest. This is in contrast to the earlier informal communication where the author usually is directly communicating with the recipients of the information.

The last part of the communication chain is carried out by the recipients of the information. In any lifecycle studies this part is extremely important, and has also been profoundly affected by the Internet.

Publishing consists of two separate activities, the writing of the manuscript, which the researcher carries out alone or in a small group, usually taking into account feedback from colleagues, and the more formal publishing process, in which outside persons, such as conference organizers, journal editors and staff, etc., participate.

The sub-process *Publish the results* is split into four parallel tracks which all take the generic 'Manuscript' as input: books, monographs, conference papers and peer-reviewed journal articles. Of these four only the part *publish as scholarly journal article* has at this stage been further detailed in the

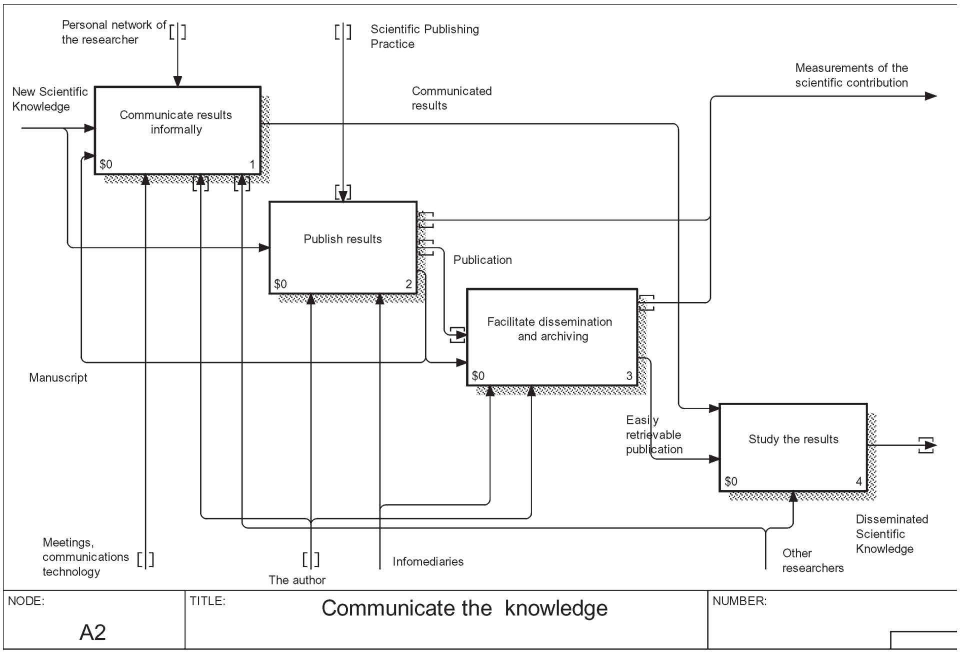

Figure 2. Diagram A2: communicate the knowledge.

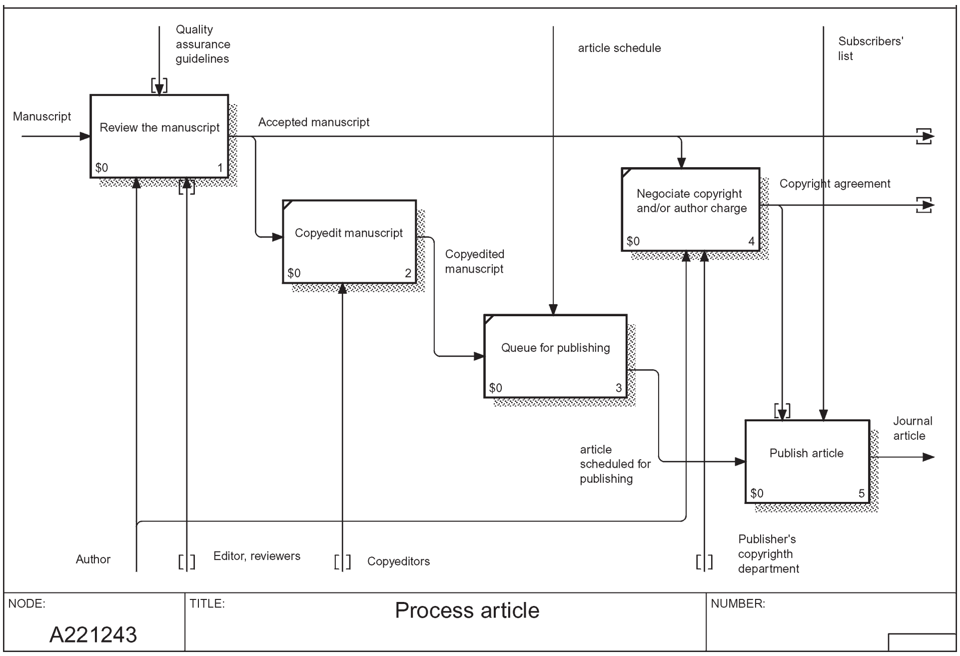

Figure 3. Diagram A221243: process article.

model. This is because of its relative importance in scientific publishing. The main pipeline of this part of the model is the input arrow Manuscript, which directly enters the activity *Process article* (see Figure 3).

The *Process article* diagram (A221243) starts out with the review activities carried

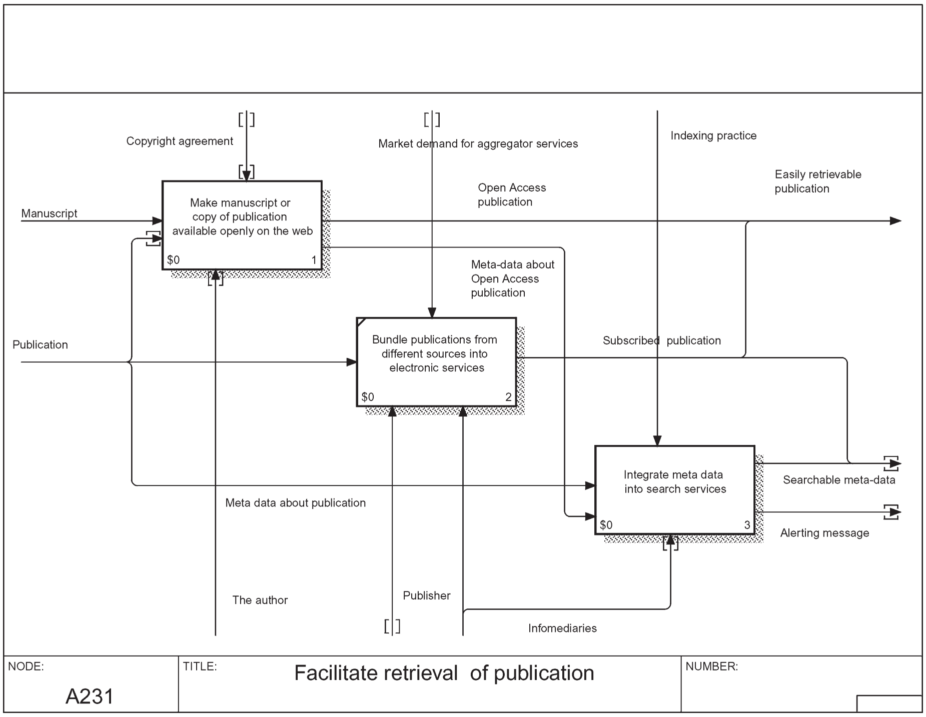

Figure 4. Diagram A231: facilitate retrieval of publication.

out as a co-operation between the editor, the researcher and anonymous peer academics. This activity demands resources but is usually not a cost item of significance for the publishers since reviewers usually work for free (from society's viewpoint this is, however, a significant cost item). A new item compared to the earlier version of the model is the *Negotiate copyright and author charge activity*. The rest of this diagram is a straightforward workflow of a typical article. Note in particular the value-decreasing activity of *queue for publishing*, where fully processed articles have to wait for several months due to the issue scheduling of the journal. Waiting does not imply a direct cost, but there may be an important opportunity cost involved from the viewpoint of the researcher and society, since the results are poorly spread before the actual publishing. This opportunity cost is different for different domains of science. It might be low for the humanities but is usually higher in the STM (science, technology and medicine) domain. In particular this is the case for IT research, where developments are extremely fast. This has been a strong motivator for the founding of both e-print archives and electronic open access journals.

The part of the model dealing with how intermediaries facilitate finding out about and retrieving material includes a distinction between open access material, which can be either in the form of manuscripts or copies of formally published papers posted in e-print archives, and in 'toll-gated' material (Figure 4). The process has been split into two sub-activities in which the first models activities carried out by different intermediaries, typically only once for the whole world market (*Facilitate retrieval of publication*, A231), and the second the activities carried out in the local organizations of the readers (*Facilitate retrieval inside reader's organization*, A232), thus typically thousands of times for each article. From a cost viewpoint, hundreds or even thousands of libraries from all over the world have been performing the same archiving function for each paper version of an article.

hundreds or even thousands of libraries from all over the world have been performing the same archiving function

For subscribed material a further split is made into secondary publishers who bundle full-text material from several different sources (an example is EBSCO) and sell it to

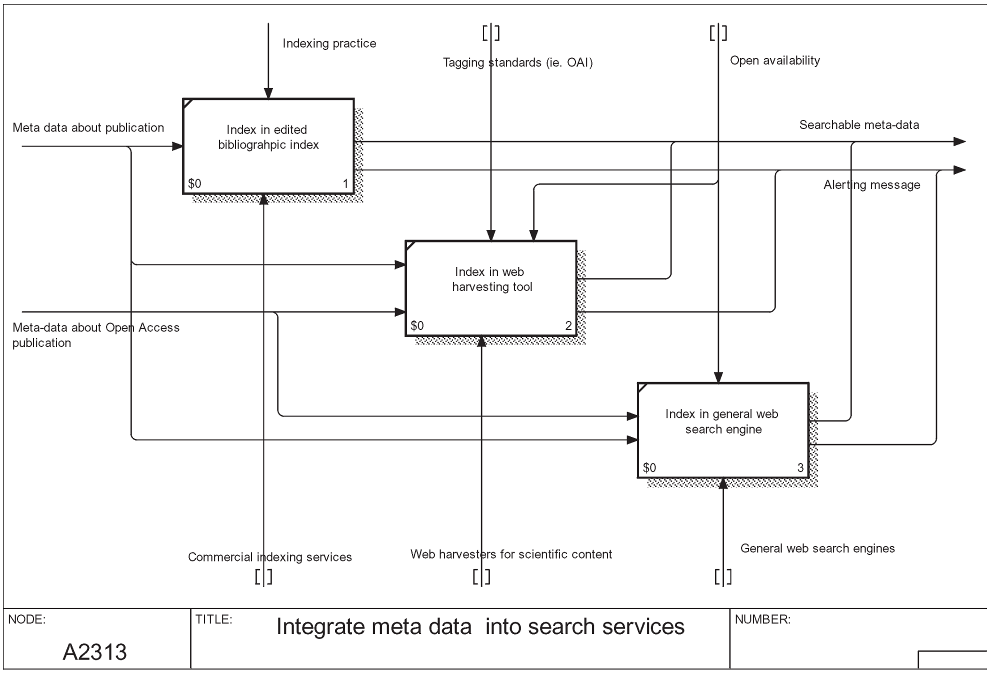

Figure 5. Diagram A2313: integrate metadata into search services.

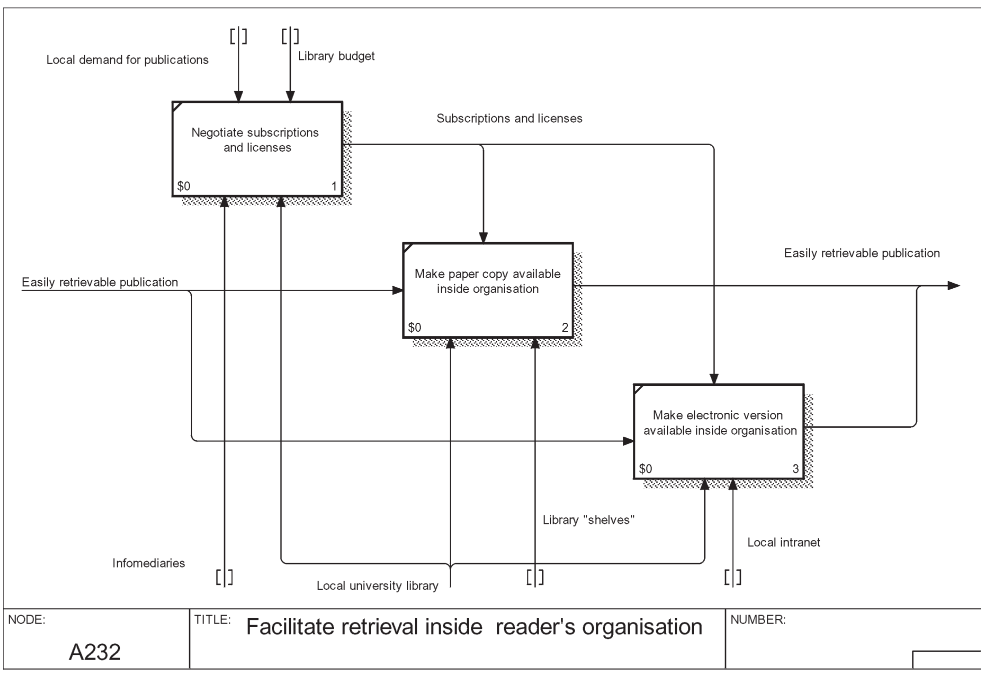

Figure 6. Diagram A232: facilitate retrieval inside reader's orgnanization.

libraries, and indexing services, which help in the retrieval function only.

Subscription-based indexing services have traditionally dominated the *Integrate metadata into search services* (A2313) function (see Figure 5). Over the past years researchers

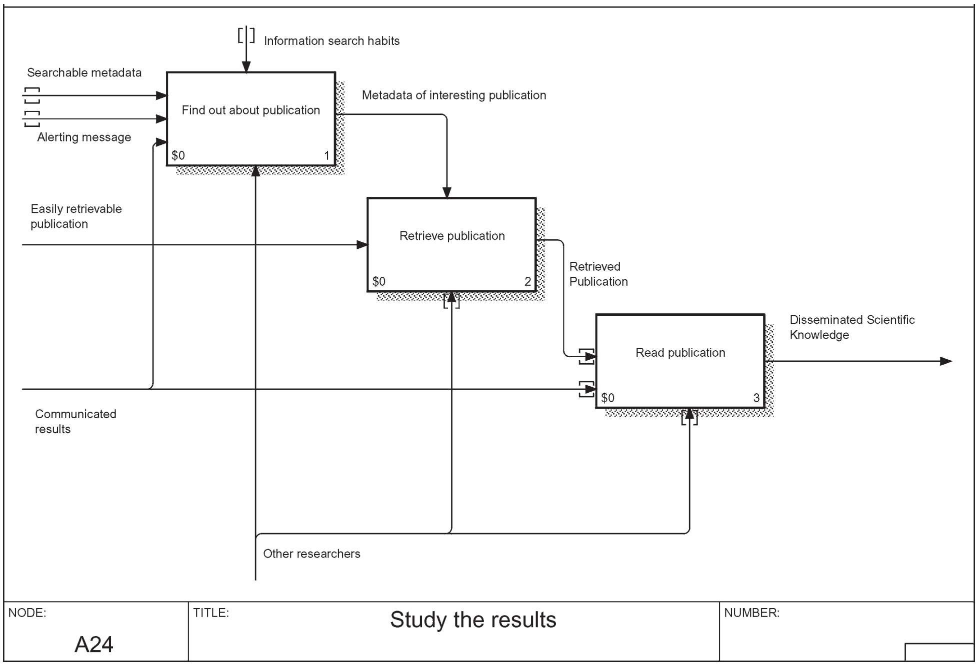

Figure 7. Diagram A24: study the results.

researchers often self-archive interesting publications they have read

have increasingly started to use general web search engines to try to identify interesting publications. An effort to overcome the quality problems related to using general search engines is the definition of the Open Archives Initiative standard for tagging scientific content material on the web, which will enable dedicated harvesting search engines to maintain a much more focused database of links to relevant publications.

A by-product of the heavy use of IT for these purposes is the possibility that readers can subscribe to services that, based on the interest profiles they define, can send them alerting email messages when something they might be interested in is published.

The *Facilitate retrieval inside reader's organization* (A232) sub-process shows the activities carried out by the organization in which the reader works to facilitate access, e.g. the university library of the reader (Figure 6). Note the inclusion of a separate activity for the negotiations that the library carries out in order to obtain the necessary licenses (the activities by library consortia could be included here as well as a sort of overhead cost). One of the biggest changes that electronic publishing has brought is the dramatic reduction in the activities to make paper publications available inside the organizations.

The *Study the results* (A24) diagram structures the activities of the readers of scientific publications. Note that from a cost per publication viewpoint the activities of individual readers all over the world and in different time periods should be summed up. The *Find out about publication* activity results in the output *metadata of interesting publication* (including the location from which a paper or electronic version can be retrieved). This output is used as the control of the retrieve publication activity. Finally the publication is read and the scientific information in question disseminated (Figure 7). Note that researchers often self-archive interesting publications they have read either as paper copies, or today increasingly as bookmarks or in a database.

Earlier models of the scientific communication process

Some earlier models or studies of the scientific communication process have been presented in the scientific literature. Garvey and his colleagues at the Johns Hopkins

University presented a model in the early 1970s, based to some part on empirical observation of scientists in the domain of psychology.[14] The Garvey–Griffith model was a good description of how the communication process functioned at a time when IT support was still lacking. The modelling consisted of verbal descriptions supplemented by graphical diagrams. A central aspect was the integration of both formal and informal communication of research results and also the inclusion of the research into the body of scientific knowledge in its domain through citations in other publications, inclusion in review articles, etc.

In the mid 1990s Julie Hurd re-examined the status of the scientific communication process and took explicit account of the emerging effects of the Internet (i.e. email and listservers and electronic publications).[15] She has recently written on the subject, taking into account recent developments such as self-publishing on the web and institutional repositories.[16]

The book by Carol Tenopir and Donald King,[17] *Towards Electronic Journals – Realities for Scientists, Librarians and Publishers*, contains a comprehensive discussion of the scientific publication process from a lifecycle perspective, and in particular synthesizes a large body of empirical evidence concerning the costs of different phases.

Discussion

The model in its current shape has not been validated in its details, but has been discussed with several colleagues with encouraging feedback. Based on these discussions and on a literature review it is the conclusion of the author that this is the first time a formal process-modelling methodology has been used to model the system of scholarly communication in this comprehensive way. Publishers employ methods of a similar nature to study the workflows within their organizations, but the whole point here is to study the whole system, including the activities of authors, libraries and readers.

Compared to the earlier models presented by Garvey and Griffith, and Hurd the main differences are:

- Hierarchical structure of the model.
- More modelling constructs, i.e. controls and mechanisms.
- Much more detailed modelling of many of the functions.
- Disaggregation of inputs and outputs on more detailed levels.
- Modelling of many of the new system functions that have emerged as a result of the Internet (OA repositories, harvesters).

It is hoped that the model could prove useful in providing a roadmap showing the place of a number of different initiatives for increasing access to scientific publications, within the overall system of scholarly communication. It could also be used a basis for empirical cost studies (e.g. Morris[18]) and a framework for integrating the results from a wide variety of different studies, focusing on specific parts of the overall process.

References

1. Guédon, J.C. In Oldenburg's long shadow: librarians, research scientists, publishers, and the control of scientific publishing. *ARL Proceedings*, 2001:138, Association of Research Libraries, Washington DC. URL: http://www.arl.org/arl/proceedings/138/guedon.html
2. Björk, B.-C. Open access to scientific publications – an analysis of the barriers to change? *Information Research*, 2004:9(2), paper 170. URL: http://InformationR.net/ir/9–2/paper170.html
3. Goodman, D. and Foster, C. Editorial: Special focus on open access: issues, ideas, and impact. *Serials Review*, 2004:30(4), 257–381.
4. Turban, E. and Aronson, J. *Decision Support Systems and Intelligent Systems*, 5th edn. Upper Saddle River, NJ: Prentice Hall (1998).
5. Sommerville, I. *Software Engineering*, 5th edn. Harlow: Addison-Wesley (1995).
6. Hars, A. *From Publishing to Knowledge Networks – Reinventing On-line Knowledge Networks*. Berlin: Springer Verlag (2003).
7. Karstila, K., Karhu, V., Björk, B.-C. Development of process modelling tools for the construction industry. In Goncalves, R. *et al* (eds), *Product and Process Modelling in Building and Construction. ECPPM 2000, Lisbon*. Rotterdam: Balkema (2000).
8. NIST. Integration Definition for Function Modelling (IDEF0) Draft Federal Information Processing Standards Publication 183, 1993 December 21, National Institute for Standards and Technology, Gaithersburg, MD. URL: http://www.itl.nist.gov/fipspubs/by-num.htm
9. Porter, M. *Competitive Advantage – Creating and Sustaining Superior Performance*. New York: The Free Press (1985).
10. Björk, B.-C., Hedlund, T. Scientific Publication Life-Cycle Model (SPLC). A deliverable of the project Open, self organising repository for scientific information exchange (SciX) EC fifth framework

programme, Information Society Technologies (IST) (2002). URL: http://www.scix.net/
11. Björk, B.-C. and Hedlund, T. Scientific Publication Life Cycle Model (SPLC). In *Proceedings of the Elpub2003 Conference: From Information to Knowledge, Guimaraes, Portugal* (2003). URL: http://elpub.scix.net/cgi-bin/works/Search?search=Bjork&paint=1&_form=AdvancedSearchForm
12. Björk, B.-C. and Hedlund, T. A formalised model of the scientific publication process online. *Information Review*, 2004:28(1), 8–21.
13. Björk, B.-C. Scientific communication life-cycle model, version 3.0, 10.2.2005. URL: http://www.oacs.shh.fi/publications/Model35explanation2.pdf
14. Garvey, W. and Griffith, B. Scientific communication: its role in the conduct of research and creation of knowledge. *American Psychologist*, 1971:26(4). Reprinted in *Key Papers in Information Science*, ed. Griffith, B. White Plains, NY, Knowledge Industry Publications, 1980, pp. 38–51
15. Hurd, J.M. Models of scientific communications systems. In Crawford, S., Hurd, and Weller, C. (eds), *From Print to Electronic: The Transformation of Scientific Communication*. Medford, NJ: Information Today, Inc., pp. 9–33 (1996).
16. Hurd, J. Scientific communication: new roles and new players. *Science & Technology Libraries*, 2004:25(1–2), 5–22.
17. Tenopir, C. and King, D. *Towards Electronic Journals – Realities for Scientists, Librarians and Publishers*. Washington, DC: Special Libraries Association (2000).
18. Morris, S. The true costs of scholarly journal publishing. *Learned Publishing*, 2005:18(2), 115–26. URL: http://www.ingentaselect.com/rpsv/catchword/alpsp/09531513/v18n2/s6/p115.

Bo-Christer Björk
Swedish School of Economics and Business Administration
PO Box 479
00101 Helsinki, Finland
Email: Bo-Christer.Bjork@hanken.fi