

SCIENTIFIC PUBLICATION LIFE-CYCLE MODEL (SPLC)

BO-CHRISTER BJÖRK ; TURID HEDLUND

Department of Management and Organisation
Swedish School of Economics and Business Administration
P.O Box 479, 00101 Helsinki, Finland

e-mail: bjork@shh.fi ; hedlund@shh.fi

The scientific publishing process has during the past few years undergone considerable changes, due to the possibilities offered by the Internet for fast delivery and inter-linking of publications which refer to each other. The socio-economic structures have, however, not changed much, and many academics and librarians view the current situation as sub-optimal and highly unsatisfactory. This has triggered a number of initiatives to set up e-print repositories and electronic peer reviewed journals, which usually offer the full text for free on the web. The label *Open Access* is gaining popularity for describing such efforts, in line with the Open Source term used in the software development community. Despite the obvious advantages it offers, Open Access hasn't become as widespread as expected, and there is a need for both successful demonstrator projects and serious research studying the way the scholarly communication system is affected by the Internet. As a part of the EU funded SciX project the cost implications of different publishing models are being studied. A basis for the cost studies is the formal process model of the scientific publishing process, which is presented in this paper. The model has been developed using the IDEF0 modelling method, a method which allows the breaking up of a process into its parts connected by output and input arrows. The *Scientific Publication Life-Cycle Model* models the life cycle of the single publication, in particular the refereed journal article, from the research leading to it and writing it into it being read by other researchers years later or used as an impulse for practical implementation. The paper presents the 22 hierarchical diagrams of the model including 64 separate activities. Conclusions are drawn about the usefulness of the modelling methodology for this particular purpose as well as of future use of the model itself. In addition to providing a basis for cost studies the model could function as a road map for different types of Open Access initiatives.

Keywords: Scientific Publication, Open Access, IDEF0, Process modelling

1 INTRODUCTION

A breakdown of the costs of producing and delivering a typical refereed journal paper indicates that perhaps as much as 90 % of the cost consists of the actual research work preceding the writing of the paper. The research work is usually financed by public bodies and the costs are in no way recuperated through the sales of the publications (as would be the case for commercial products such as books, music CDs, movies etc). Even if we look only at just the costs of preparing, reviewing, distributing, archiving and retrieving scientific articles, thus excluding the actual production costs of the knowledge reported on, almost all the costs are in the end born by the universities, public libraries etc. Because of the commercial interests of one group of stakeholders, the journal publishers, which incur a very small fraction of the total life-cycle cost, the access to scientific publications is highly restricted and expensive and the process as a whole is highly inefficient. For a recent overview cf.[1]. Publishing parallel electronic versions restricted to subscribers only partly remedies the situation and leaves much of the potential of the Internet untapped.

The dilemma is that it would be in the interest of the researchers and the public to have all this information published for free on the Internet, facilitating global access and hyper-linking of research publications. Nevertheless it is in the legitimate interest of the publishers to make a profit from selling this information, which leads to restricted circulation, pass-word protection schemes for digital versions of traditional journals, bundling of electronic subscriptions to so called "big deals" with library consortia.

Pioneering work to try out new methods of scientific publishing has already been started, usually by enthusiasts from subgroups in the scientific community. Examples include preprint repositories, open access scholarly journals and journals offering a review process where the readers perform the peer review after a manuscript has been posted on the web. There are, however, many psychological, legal and institutional barriers to change the process and these have been underestimated by the pioneers and enthusiasts. Consequently only a small part of the overall volume of the scientific communication process has so far been affected by isolated efforts involving e-journals and preprint archives. A recent study made at the Swedish School of Economics and business studies indicates that only about 0,7 % of peer reviewed journals offer open access on the web.

The SciX (Open, self organising repository for scientific information exchange) project, which is financed through the IST program of the European Commission, aims at demonstrating that the Internet enables new business models for the scientific publishing process which are much more cost and time efficient to the scientific community than the current practice [2]. The SciX project group will create services on the Web that will enable scientists as well as practitioners from the fields of architecture and construction easy and free access to relevant research publications. In addition existing publishing practices will be analyzed systematically and business models for re-engineering the scientific publishing process will be developed, taking into account also the legal, social and psychological barriers to change. The model presented in this paper is one of the deliverables of the theoretical work done within the SciX project.

2 AIM AND SCOPE OF THE MODELLING EFFORT

The aim of the modelling is to help us understand the scientific publishing process and how it is 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 of scholarly communication. The model explicitly includes the activities of all the stakeholders involved in this system, including the activities of the:

- **Researchers** who perform the research and write the publications
- **Publishers** who manage and carry out the actual publication process
- **Academics** who participate in the process as editors and reviewers
- **Libraries** that archive the publications and provide access to the them
- **Bibliographic services** which facilitate the identification and retrieval of publications
- **Readers** who search for, retrieve and read publications
- **Practitioners** who implement the research results directly or indirectly

In the model the 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 viewpoint taken is life-cycle cost per publication. Thus at later stages all cost and time data which is collected will be translated to a per publication basis.

The model depicts publishing and value added services using both paper and electronic formats. 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 current version of the model has some limitations, which should be kept in mind. It only includes 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, sharing of data sets and models, which all are essential parts of the scientific communication process, are out of scope. These could be added at a later stage, but would also add to the complexity of the model.

The model includes some activities, which 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.

How easier access to scientific publications might influence the quality of later research and industrial practice, which use these publications as input, is clearly also an important issue, but such qualitative effects of a more efficient process are very difficult to model and even more difficult to measure, and have not been attempted in this model. The same applies for the effect of the publishing on the careers of the authors, which is an important aspect for the choice of where to publish and has created a very strong barrier for change to pure electronic journals from established “brand name” journals.

3 OVERALL ORGANIZATION OF THE MODEL

The current version of the SPLC-model includes 22 separate diagrams, arranged in a hierarchy up to seven levels deep. There are typically three activity boxes on each diagram, although there are a couple of diagrams with more activities and some with only two. The overall hierarchical breakdown of the model is shown below in table 1. Only the separate diagrams are shown. Some diagrams are further broken down into separate activities. In the following each diagram is explained separately. The diagrams are numbered using the standard IDEF0 numbering scheme [2], which helps keeping track of the hierarchical position of each diagram. The complete model is available at the SciX website [3].

TABLE 1. HIERARCHICAL BREAKDOWN OF THE MODEL, ONLY THE DIAGRAMS ARE SHOWN

- A-0 Context Diagram
 - A0 Do Research, Publish, Study and Exploit the Results
 - A1 Perform the Research
 - A2 Publish the Results
 - A21 Write Manuscript
 - A22 Perform Publishing Activities
 - A221 Publish as Monograph
 - A222 Publish as Conference Paper
 - A223 Publish as Scholarly Journal Article
 - A2231 Do General Publisher's Activities
 - A2232 Do Journal Specific Activities
 - A2233 Do Article and Issue Specific Activities
 - A22331 Article Specific Activities
 - A22332 Prepare Issue
 - A22333 Publish Article
 - A224 Publish in Miscellaneous Form
 - A23 Archive and Index
 - A231 Make Publication Available
 - A2311 Secure Access Rights and Subscription
 - A2312 Make Paper Publication Available
 - A2313 Make Electronic Copy Available
 - A2314 Integrate Meta Data into Search Services
 - A232 Perform Value-Adding Services
 - A233 Archive Securely
 - A3 Study the Results
 - A31 Find out about Publication
 - A311 Search for Publication
 - A312 Be Alerted to Publication
 - A32 Retrieve Publication
 - A33 Read Publication
 - A4 Implement the Results

4 MODEL WALK-THROUGH

A0 Do Research, Publish, Study and Implement the Results - breakdown

This diagram is crucial for understanding the life-cycle view adopted in this modelling effort. The whole life-cycle is seen as consisting of four separate stages. The *do the research* stage is probably the most expensive part, usually consisting of several man-months of work effort per resulting publication, but the one least affected by the reengineering efforts facilitated by the Internet (at least directly, indirectly the effect can be substantial in terms of better quality of the research). The *publish the results* and *study the results* stages constitute the main object of study in this project. This part of the model tries to clarify the dual nature of the publication process. From the perspective of the public bodies that to a large part finance research it is the efficiency of the total process, including both the production and “consumption” of publications, that should be optimized. The important thing is that in a life cycle analysis, the cost and efficiency of both the publish the results activity and the study the results activity are important. Optimizing only one of these may lead to a sub optimal solution for the total process. Here Internet has changed the situation dramatically, as it has for any information goods that can be delivered in a digital format.

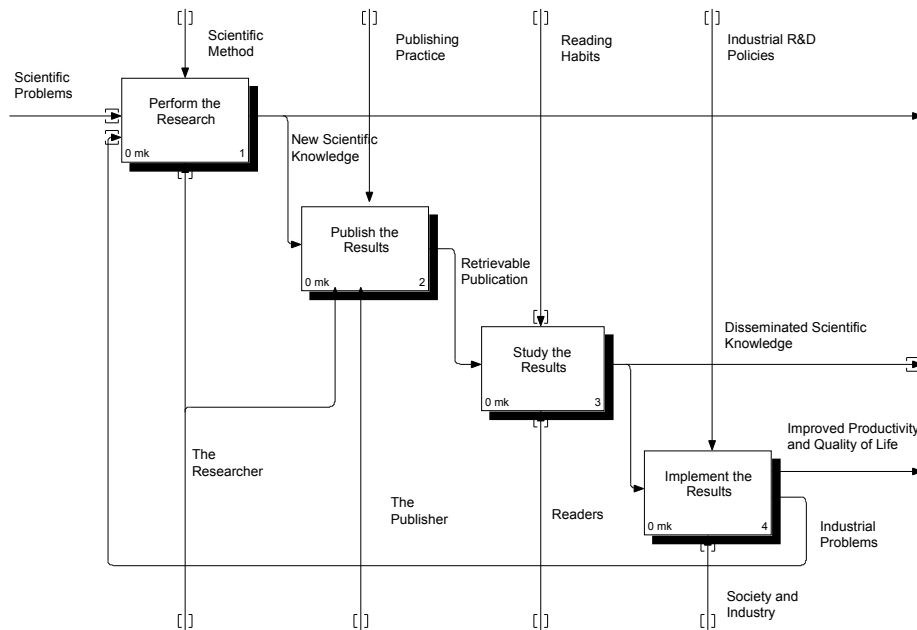


Diagram AO Do Research, Publish, Study and Apply the Results, Breakdown

The end result of these two stages is called *disseminated scientific knowledge*, reflecting the viewpoint that scientific results which have been published, but which are not read by the intended readers are rather useless. In the last stage, *implement the results*, the *disseminated scientific knowledge* is transformed to an improved performance of society and industry.

A2 Publish the Results

This part of the model has been split up into three distinct activities, which to a large extent are carried out by different stakeholders. Based on the results of the research, the researcher writes a manuscript, which then in the next stage through a number of transformations is changed into a *publication* (on paper or electronic). The last activity is extremely important from a life-cycle viewpoint and involves the activities of libraries, bibliographic services etc. to make the publication easily available to researchers and practitioners world-wide.

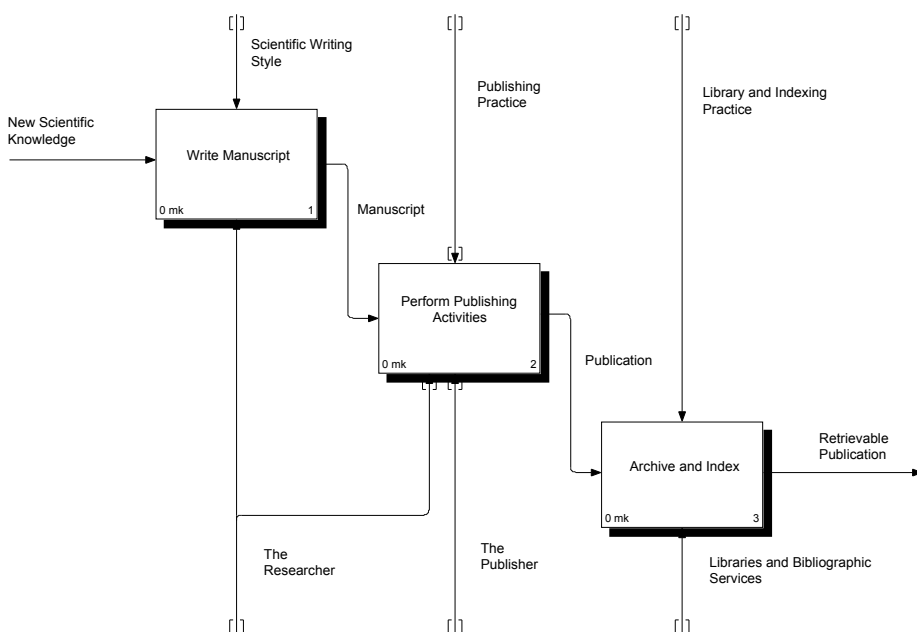


Diagram A21 Publish the Results

A22 Perform Publishing Activities

At this stage the model is split into four parallel tracks which all take the generic “*Manuscript*” as input.

A223 Publish as Scholarly Journal Article

This diagram may at first sight be difficult to understand. The idea is to show all the activities which are carried out by the publishing organization, and thus have a direct cost implication for them. This is the reason for separating activities such as *do general publisher’s activities*, *do journal specific activities*. Both of these demand resources, which cause overhead costs, which then are added on top of the basic variable costs caused by the processing of each individual article (in the activity *do article specific activities*). For instance setting up and maintaining the IT-technical infrastructure for a portfolio of journals could be such an overhead causing item. The main pipeline in the model is, however, the input arrow *manuscript*, which directly enters the activity *do article and issue specific activities*.

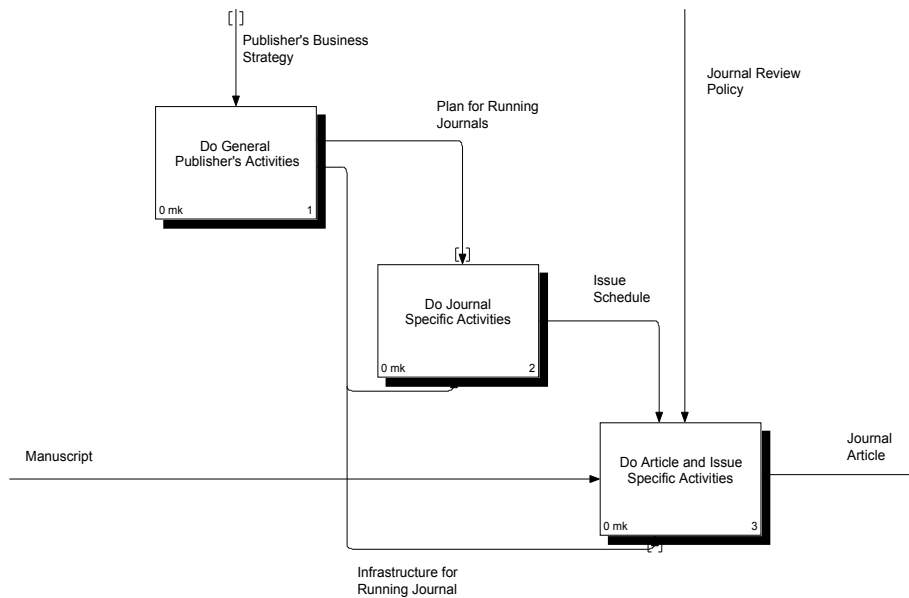


Diagram A223 Publish as Scholarly Journal Article

A2232 Do Journal-Specific Activities

Like many of the diagrams in this model, this model represents a choice of viewpoint. Here an important aspect is that commercial journals may spend a lot of money on marketing, and also on the management of subscribers (invoicing, setting up ways of checking access to electronic versions). For open access electronic journals, the latter activity is almost non-existent. Note the output arrow *issue schedule*, which is later used as a control of issue-specific activities.

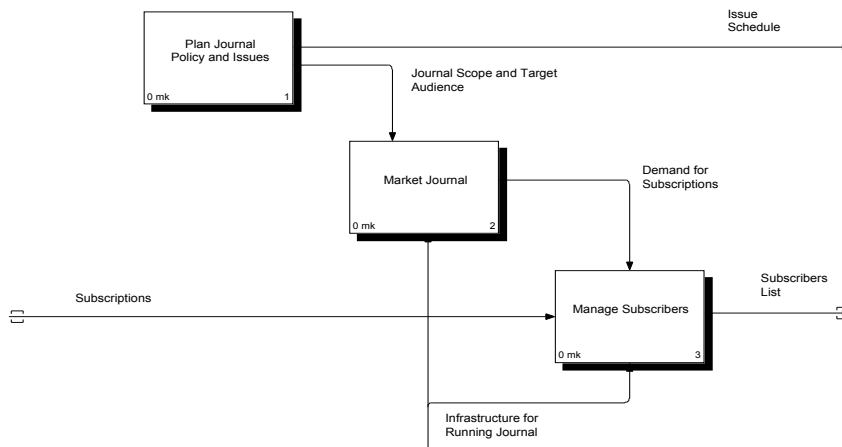


Diagram A2232 Do Journal-Specific Activities

A 2233 Do Article and Issue Specific Activities

This diagram shows the two major modes for publishing scientific journals. In the paper-based world prior to 1990 articles were as a rule bundled into issues and had to wait for publishing until the whole issue was ready. Electronic publishing does however provide the possibility to publish each article as soon as it is ready. Today many journals are printed in both print and electronic formats but still retain the issue-based structure.

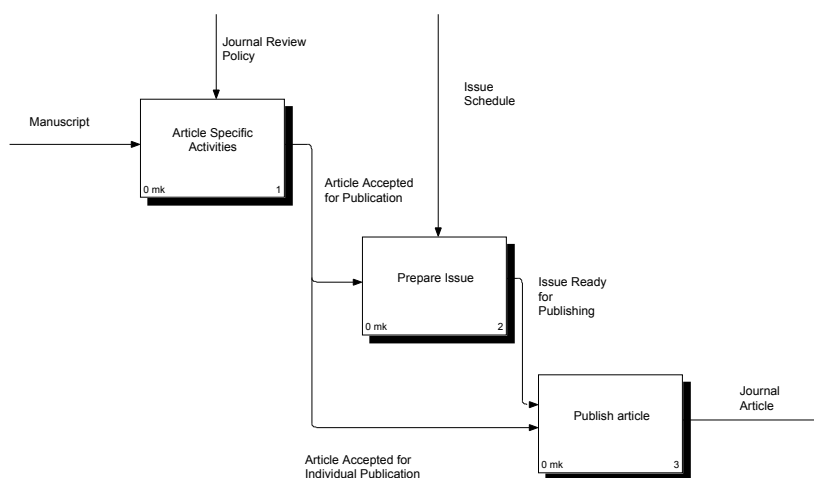


Diagram A2233 Do Article and Issue Specific Activities

A 22331 Do Article Specific Activities

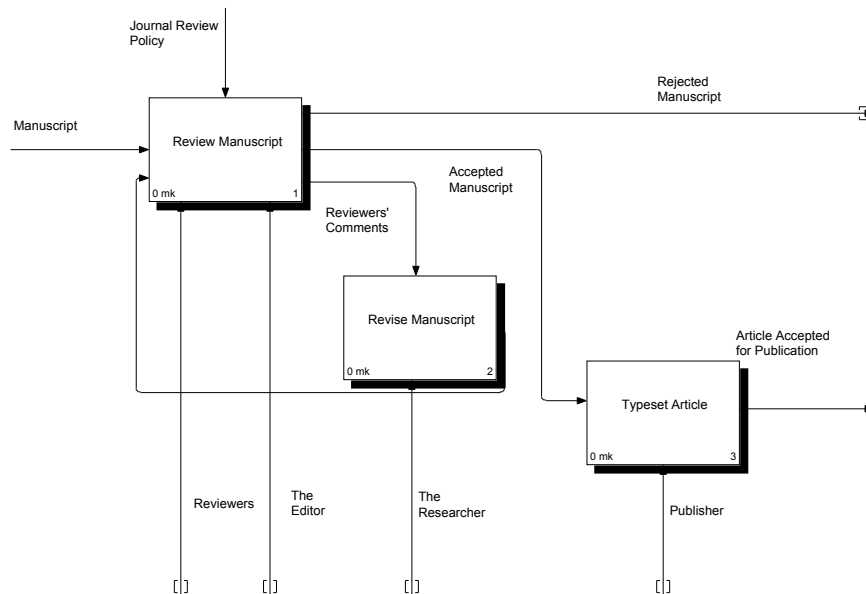


Diagram A22331 Do Article Specific Activities

A22332 Prepare Issue

This diagram includes a very significant activity which might escape modelling in many modelling exercises, which focus solely on cost or the publisher's activities. Once an article is accepted for publishing, it enters an activity called *queue for publishing*, which typically takes from half a year to a year for traditional issue & paper-based journals.

A23 Archive and Index

This is the part of the overall process, which traditionally to a large part has been handled by research libraries, with public funding. Note also that from a cost viewpoint, hundreds of libraries from all over the world have been performing the same archiving function for each paper version of an article. The primary activity is here *make publication available*, which secures that a publication is available either in print or electronically within a particular organization (such as a university), as well as that the publication can be found in different bibliographical search services. In the *perform value-adding services* a third party analyzes the data to calculate citation indexes, impact factors etc., or writes news bulletins about research results that practitioners can digest more easily. The *archive securely* activity is currently receiving increasing attention, since the archiving of electronic versions of journals for decades implies a number of difficult problems.

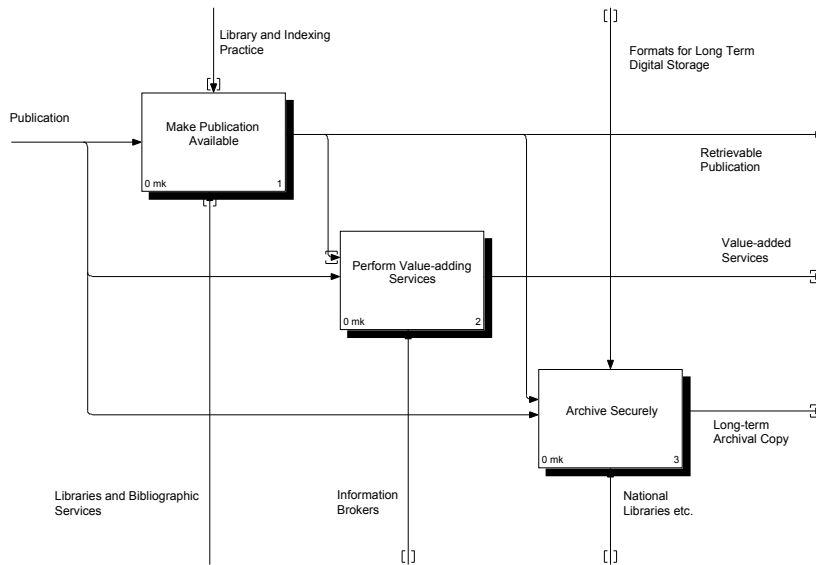


Diagram A213 Archive and Index

A231 Make Publication available

This diagram includes both the activity of making the paper publication available (placing it in the shelves of the library) and making the electronic version available. In both cases this is preceded by the longer term activity of securing subscriptions and access rights to the material, an activity which is even more visible today due to the large library consortia that negotiate “bit deals” with the large publishers. An additional value adding activity is the integration of the meta data about the publication in data bases which facilitates finding out about the existence of the publication.

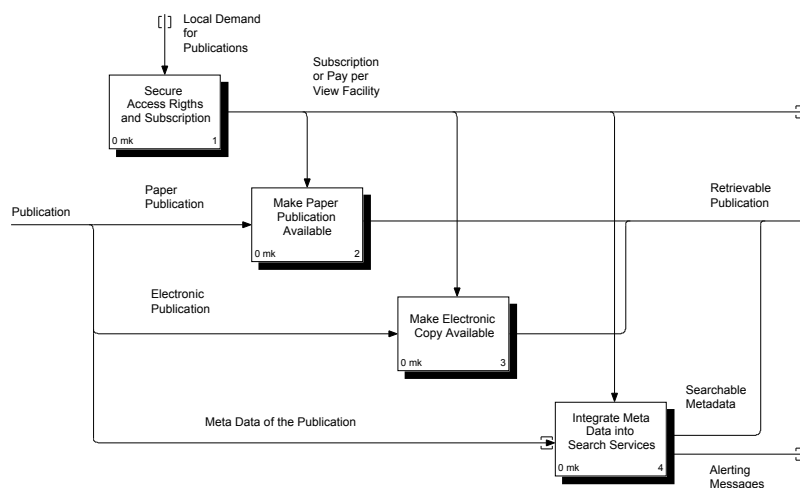


Diagram A231 Make Publication Available

A2314 Integrate Meta Data into Search Services

Traditionally commercial indexing services have dominated this function and libraries have paid subscriptions to them. Over the past years researchers have increasingly started to use general web search engines for trying to identify interesting publications. An effort to overcome the quality problems related to this 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 of readers to subscribe to services, which based on the interest profiles they define, can send them alerting email messages when something they might be interested in is published.

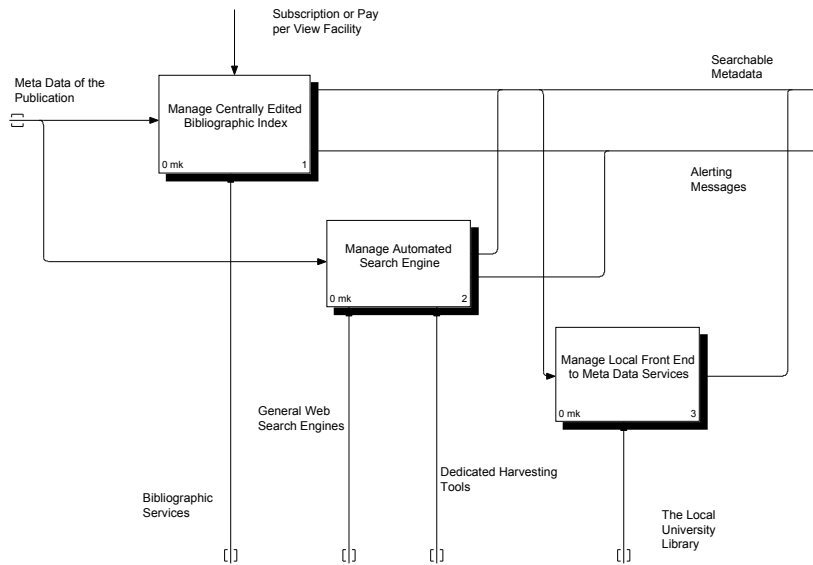


Diagram 2314 Integrate Meta Data into Search Services

A3 Study the Results

This diagram structures the activities of the readers of scientific activities. Note again that from a cost per publication 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 *meta-data 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 has been disseminated. Note that as a rule researchers self-archive interesting publications they have read either as paper copies or today increasingly as bookmarks or in a data base.

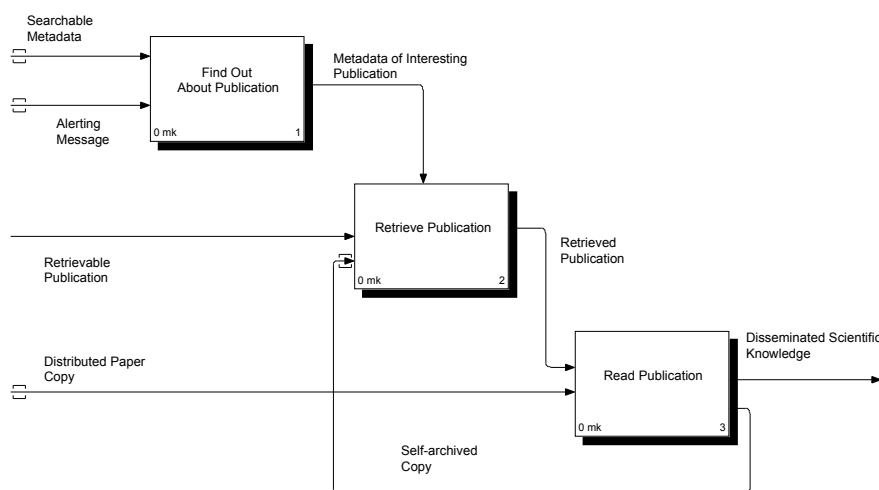


Diagram A3 Study the Results

A31 Find Out about Publication

This activity is rather difficult to split up into alternative parallel options (for each identified article one of these is used). On this first level there is a split into two generic categories. The first one is active search where the reader is “pulling” for information. The other option is push, where the reader receives a notification through some mechanism that something interesting has been published.

A311 Search for Publication

The first modelled option of the pull variety is a traditional bibliographic data base search, for instance using key words. Other possibilities include more unstructured web searches using a general search engine or just browsing from one hyperlink to the other. Less and less the traditional method of physically browsing in library bookshelves is used.

A312 Be alerted to publication

Receiving a hint from a third party could be a hint from a colleague or a supervisor, or in today’s world a hint from a software based alerting service. The important distinction is that the bibliographic search is triggered by the researcher himself (pull) whereas the hint is coming from the outside (push).

A32 Retrieve Publication

The basic split has here been made between the retrieval of a copy of a paper publication and one in digital form. Although the two activities modelled here may look straightforward, they might become rather complex in reality. If for instance the organization that the researcher belongs to doesn’t subscribe to the journal in question, it might take quite a long time to obtain a copy of the article through some add on service for interlibrary loans or through a service for buying individual articles over the web. Many younger researchers have become lazy for retrieving paper-based copies at all, since so much is there on the web.

5 CONCLUSIONS

The use of a formal process modelling language for a purpose such as this was motivated by the personal experience of the main researcher, having earlier used this type of methodology for the modelling of the construction process as well as having led an international project investigating innovative process modelling methodologies. The initial experiences have been very positive. The studied process is by its very nature rather linear (contrary to for instance architectural design), which makes the modelling easier than for processes involving a lot of networking or iterative procedures. Also colleagues to whom the model has been shown have quite easily grasped the fundamentals of the IDEF0-notation and have been able to follow the logic of the model.

The model in its current shape is not yet validated in its details, but has been discussed with a number of domain experts (publishers, librarians) with encouraging feedback. Based on these discussions and on the extensive literature review done as part of the SciX project it is the conclusion of the authors that this is the first time a formal process modelling methodology is used in this comprehensive way to model the system of scholarly communication. Publishers employ methods of a similar nature to study the workflows within their organizations, but the point here is to study the whole process, including the activities of libraries and readers, and to use the process model as a basis for determining the activities which will be studied more in detail as a part of the cost modelling.

The cost modeling will be a synthesis task of data from several different sources. To some extent web surveys will be used, in particular concerning the economics of open access journals and repositories and reader behavior. Research work of other researchers concerning the cost of certain activities in the model will also be used [4],[5]. The common denominator will be the to study costs per publication flowing through the system. In addition to the cost modelling the model could also 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.

ACKNOWLEDGEMENTS:

The presented work has been conducted in the context of the SciX project, funded by the European Commission under the contract IST-2001-33127. The opinions expressed in this paper are that of the authors and do not necessarily represent the opinions of their employers, of the SciX Consortium or of the European Commission. Note that the version of the model presented in this paper is the second first draft (version 2.0) and that the model is continuously evolving based on the feedback we receive. The latest updated version of the model can be found on the SciX website (<http://www.scix.net/>), which interested readers are advised to consult.

REFERENCES:

- 1 Guédon, J.,C. In Oldenburg's Long Shadow: Librarians, Research Scientists, Publishers, and the Control of Scientific Publishing. ARL Proceedings 138, Washington D.C.: Association of Research Libraries, 2001. <http://www.arl.org/arl/proceedings/138/guedon.html>
- 2 NIST Integration Definition for Function Modeling (IDEF0). Draft Federal Information Processing Standards Publication 183, 1993 December 21, Federal Information Processing Standards Publications No 183. Gaithersburg, Md, USA : National Institute for Standards and Technology, 1993. <http://www.itl.nist.gov/fipspubs/by-num.htm>
- 3 SciX 2002 Open, self organising repository for scientific information exchange project home page <http://www.scix.net/>
- 4 Odlyzko, A. The Economics of Electronic Journals. Journal of Electronic Publishing, Vol 4/1 1998 <http://www.press.umich.edu/jep/04-01/odlyzko.html>
- 5 Tenopir, C. and King, D. Designing Electronic Journals With 30 Years of Lessons from Print The Journal of Electronic Publishing, Vol 4/2 1998 <http://www.press.umich.edu/jep/04-02/king.htm>