Reclaiming academic output through university archive servers

van Bentum, M.; Brandsma, R.; Place, T.; Roes, H.

Publication date
2001
Document Version
Final published version
Published in
New Review of Information Networking

Reclaiming academic output through university archive servers

Maarten van Bentum
University Library, University of Twente, The Netherlands
e-mail: M.vanBentum@dinkel.utwente.nl

Renze Brandsma
University of Amsterdam, Spui 21, 012 WX Amsterdam, The Netherlands
e-mail: brandsma@uba.uva.nl

Thomas Place
Tilburg University Library, P.O. Box 90153, 5000 LE Tilburg, The Netherlands
e-mail: T.W.Place@kub.nl

and

Hans Roes
Tilburg University Library, P.O. Box 90153, 5000 LE Tilburg, The Netherlands
e-mail: H.Roes@kub.nl

INTRODUCTION

In September 2000 the Dutch ARNO (Academic Research in the Netherlands Online) project was started. This project is funded by the Dutch equivalent of the UK eLib program, IWI (1), and is carried out mainly by library staff of the University of Twente, the University of Amsterdam and Tilburg University. The project is scheduled to run up until September 2002. The project's intermediate goal is to design and establish university archive servers which store each university's academic output. This output can range from educational materials to research reports, theses and dissertations, as well as articles published in regular scholarly journals. The ultimate objective is to make this output freely accessible to the academic community in support of research and education. ARNO archive servers will be interoperable by complying with the Open Archives Initiative standard (2).

This article first gives some background information on the project. The following section reports intermediate results of one of the core work packages, which identifies conditions for author participation and aims to design strategies to convince academic staff to make their output available through archive servers. Next, the technical development of the ARNO archive servers is described. Finally, research on possibilities for connecting peer review with archive servers and interoperability with publishing workflow processes will be described. It should be stressed that work on ARNO is still in progress and that this article reports on intermediate results only. Future developments can be monitored on the project's website (3).

BACKGROUND

The past few years have shown the beginning of change in relations between universities and publishers. The most important characteristic of this change has been an increase in universities' awareness of their role as suppliers of publications, rather than being just passive clients of publishers. IWI has stimulated this development by supporting research for new models for publication conditions between academic staff, publishers and universities - which employ the academic staff. In this work, it is proposed to make arrangements regarding the copyright of electronic publications of faculty in employment contracts. In this way, faculty can be obliged to reserve the right to (possibly after a limited period) freely use electronic versions of their publications for research and education. Other examples of this development have been IWI projects in which e-print servers and electronic journals were developed.

Anther important development at the time of writing the project proposal (early 2000) was consortium negotiations between the Dutch Association of University Libraries and the Royal Library with publishers. These concern pricing and access to digital versions of journals, but also point in the direction of research into new business models. By redesigning the production and distribution of scientific information, important gains in efficiency seem possible. These gains are necessary in order to achieve an economically sustainable information chain. In this redesign, the role of the various actors in the present information chain will be redefined.

In order to let these developments gain momentum, it is necessary, as

IWI pointed out in its strategic plan 2000 - 2002, that an infrastructure becomes available based on the need of faculty, and which offers them easy solutions for publishing their output electronically. This infrastructure needs to build on the achievements of the present publication process in which quality control, and the recognition which comes with it, play such important roles. At the same time there is a need to take into account the differences between academic disciplines with regard to publishing practices.

The three partnering institutions in the ARNO project are researching how such an infrastructure can be developed. In the past few years, technologies have become available which offer advanced possibilities for distributed storage and indexing of publications. By combining these technologies, an infrastructure can be built with which several objectives can be realised:

- electronic availability of the academic output of an institution
- subject oriented interoperability of distributed digital archives, as well as interoperability with the national library infrastructure
- connecting this infrastructure with the production processes of commercial and non- commercial publishers, especially in the submission and review of manuscripts
- connecting university archive servers with digital learning environments

At the same time, it should be remembered that a technical approach by itself is not enough. The crucial point is to convince academic staff of the imperative to change their attitude towards the present publication process. They need to be aware that electronic publishing of their output can be simple, is allowed and will also benefit much needed changes in the information chain, while at the same time preserving the advantages of the present system. Relevant motives for faculty in this respect are visibility, reputation and ease of use.

The project also keeps in mind cultural differences between academic disciplines. This implies that the basic infrastructure needs to be as open and generic as possible. On such an open base it is possible to cope with different needs, wishes and customs among the various disciplines. The project partners' mix of a classical, a technical and a university for humanities and social sciences implies a distinct advantage in this respect.

The results of the project should set an example for other Dutch universities and institutions abroad which would like to set up a similar infrastructure. Experiences gained in this project will therefore be widely and freely disseminated.

ATTITUDES AND PERCEPTIONS REGARDING E-PUBLISHING AMONG ACADEMIC STAFF

Partly based on a literature study, a survey was designed and carried out among academic staff and research managers of the three universities participating in the ARNO project to investigate perceptions and attitudes regarding electronic publishing and the use of an archive server as a parallel publication channel. A prime objective of the survey was to generate input for the subsequent development of strategies focussed at authors and research managers, of various disciplines, which optimise contributions to the archive servers. Another objective was to inform academic staff of the participating universities about the ARNO project.

The survey consisted of structured interviews with research managers, and focus group interviews or structured interviews with individual authors. The table below shows the number of interviews held at the various faculties and departments of the participating universities, grouped by main disciplines.

	research managers	authors
social sciences	5	11
economics	2	6
law	1	4
humanities	3	10
mathematics	4	4
biology	2	-
physics / astronomy	4	3
chemistry	2	-
technical sciences (electrical engineering, mechanical engineering, civil engineering, computer science)	3	7

Research managers were asked to reflect on the desirability as well as probability of realising the generally supposed advantages of electronic publishing. The presented advantages were: quick availability of research results; more control on the author's own output; faster

reviewing process; lower costs; new presentation possibilities; and faster academic interaction due to free access to academic output. The wish for quicker availability of results, faster reviewing and lower costs of publication is general among the research managers of all disciplines. More control on output is only desired among research managers in social sciences, economics, law and humanities. New presentation possibilities are generally desired by research managers in social sciences, economics, law, biology and chemistry, whereas the humanities, and, more surprising, mathematics, physics / astronomy and the technical sciences managers are neutral on this.

The belief that all these supposed advantages will indeed be realised by means of electronic publishing is generally less strong than the desire. Nearly all managers are most positive with regard to the probability that electronic publishing will lead to faster release of research results. Most doubts are found among the biology and physics / astronomy managers. Also, the belief that electronic publishing offers new presentation possibilities is generally recognised, although the humanities, physics / astronomy and technical sciences managers are slightly less positive. There is no strong belief among research managers that electronic publishing will make the reviewing process faster. Social sciences and economics managers are the most positive on this aspect, law, humanities and biology managers are the most negative. Social sciences, economics, mathematics and technical sciences managers are rather optimistic about the possibility that electronic publishing might be cheaper, whereas biology, physics / astronomy and chemistry managers are rather pessimistic. Finally, with respect to the supposed advantage of faster academic debate the economics and law managers are clearly positive, social sciences, humanities, mathematics, chemistry and technical sciences managers are divided, whereas the biology and physics / astronomy managers are negative.

As for the attitude of research managers towards a university archive server as a parallel publication channel, the conclusion is that the majority supports the idea that the university should play an important role in distributing its academic output, although among the sciences, and especially the technical sciences, managers disagree most. For social sciences, economics, law and humanities managers the main obstacle for the use of an archive server is the possible impediments to traditional publishing (problems are expected with publishers). The

managers in sciences and the technical sciences mention the unclear costs and financing of the use of an archive server as the most important obstacle. It comes as no surprise then that the first mentioned group of managers see good and clear arrangements with publishers as a first condition for use of an archive server whereas for the second group of managers low cost is the most important condition.

The survey investigated the following opinions among authors about, and attitudes towards, e-publishing: preference for traditional publishing (regardless whether electronic or print) or electronic publishing; the (perceived) main advantages of traditional and electronic publishing; authors' attitudes towards a university archive server as a parallel publication channel.

Authors from the social sciences and humanities prefer traditional publishing mainly because of the guaranteed quality control. Among authors from economics, mathematics, physics / astronomy and the technical sciences there are different preferences. Some authors distinguish between types of document. They think that journal articles should be published in traditional ways, whereas congress papers are regarded as suitable for electronic publishing. Concerning dissertations and books, the opinions are divided. The main advantages of traditional publishing are considered to be quality control and contribution to reputation, whereas electronic publishing is thought to be much faster as well as offering better accessibility. Most authors from law departments have no preference for either form of publishing. They also regard quality control and contribution to status as the most important advantages of traditional publishing and see speed as the main advantage of electronic publishing.

Authors from the humanities and technical sciences are generally negative about the idea of an archive server as a parallel publishing channel. In any case, conditions for contributing to archive servers are quality control and easy document delivery procedures. All the authors in mathematics and physics / astronomy are positive, but the conditions are manifold: good search facilities, easy document delivery procedures and the possibility for document removal. A majority of the authors in social sciences is also positive, provided that copyright issues are settled and clear arrangements are made with publishers. A minority does not see any advantage of an archive server. Authors working in

economics and law faculties are generally positive about the idea of an archive server. Economics authors mention as conditions clear arrangements with publishers and easy document delivery procedures. Authors in law departments also put forward clear arrangements with publishers as a condition and good facilities both for delivery and retrieval.

In summary, perceptions of research managers and authors regarding electronic publishing and the use of an archive server as a parallel publication channel vary among the disciplines. It is hard to conclude that all or most of the scientific managers or authors are positive or negative about electronic publishing, and such a conclusion can neither be drawn about their attitude towards a university archive server. This implies that programs and materials for encouraging the use of university archive servers should be tailored to accommodate these differences between disciplines. Concerns about copyright confirm the need to seek optimal convergence with related developments in crafting new and more equitable relations between universities and publishers. The importance of a quality label for academic output underlines that efforts to implement such mechanisms in archive servers are necessary.

DEVELOPMENT OF ARNO ARCHIVE SERVERS

With an ARNO archive server an institution can maintain a Harvestable Open Archive as defined by the Open Archives Initiative (OAI). By implementing the OAI protocol, metadata records that refer to electronic documents in the archive can be collected (harvested) by an external application, a harvester.

In the OAI framework, there is a distinction between data providers and service providers. Archives as defined by the OAI are data providers. Service providers have access to the metadata that describe the documents via the OAI Metadata Harvesting Protocol, the official name of the OAI protocol. This allows service providers to develop added value services. An example of such an added value service is a search interface that allows users to search the metadata and/or full text documents of one or more archives. The OAI protocol itself is a simple and easy to implement protocol that lacks sophisticated search facilities.

ARNO focuses on university wide archive servers. In principle there is just one archive server per university that gives access to the

publications of the university. However, an archive server can also be used in other contexts, and a university can implement more than one server.

The ARNO archive server consists of two main components: a database storing metadata plus an application for maintaining these data; and a document store for the full text documents. An institution that uses an ARNO server can decide that placing the publications in the document store is optional. This implies that a metadata record can refer to a document that is physically outside the archive. The ARNO server can mirror such external documents in case, for one reason or another, a remotely stored document disappears from the Web. However, if an institution thinks that document integrity is important, then it is advised to store the documents on the ARNO server. It is also possible to store different versions of a document. In this way the history of a document, from a preprint to the peer-reviewed version published in a scholarly journal can be tracked.

ARNO servers are aware of three groups:

1. Contributors of documents and the corresponding metadata. The authors themselves can be contributors, but intermediaries like library staff or administrators of a department or faculty can also function as contributors to the system.
2. Administrators who oversee the process of entering data in the metadata database and the document store. The role of an administrator can be very marginal, with no influence on the quality of the publications, but an administrator might also function as an editor who can block the admission of a publication to the archive.
3. Service providers that collect the metadata of the archive.

Both contributors and administrators use the same Web interface, but with different privileges. It is possible to use external authentication mechanisms like LDAP and Athens.

Central in the ARNO data model is the notion of an organisation. Users, i.e. contributors and administrators, and documents belong to an organisation. An organisation refers to administrative units of a university, like a research institute, a department or a faculty. The user privileges, and policies with respect to quality control and the electronic

formats of the documents that are accepted, are defined per organisation or unit. An administrator can set the privileges and the policies that apply to his or her organisation through the Web interface of the ARNO server.

The ARNO server is written as a Perl cgi-script. An ARNO server must have access to a relational database management system that supports the Perl DBI interface. The metadata are maintained with a feature list. This is a table in which each row represents a metadata element of a document. To describe a document several rows are used: for each element that has a value for the document there is a row in the table. When new metadata elements are defined, for example elements describing the outcome of peer review, these can easily be added to the existing descriptions: updating the table with new rows that refer to the new elements. The database itself does not need to be reorganised. Per document type the relevant metadata elements can be defined. Both document types and corresponding metadata elements are extensible without database reorganisation. With this flexible data model, the system can support existing and future metadata standards.

The support of the OAI protocol by ARNO proved to be a relatively easy job. It passed successfully the validation tests that are available via the OAI website (2). This implies that an ARNO archive can be harvested by OAI compliant harvesters; that Dublin Core, the default metadata standard of the OAI protocol, is supported; and that ARNO can supply the metadata in valid XML documents. In a later phase of the project the ARNO archives will be harvested to demonstrate how, using a search engine, a sophisticated service can be implemented that gives access to the academic output stored in the archive servers of the participating universities.

When the software is finalised, it will be made available as Open Source.

PEER REVIEW

This section examines the possibilities to link peer review and other quality-control and certification mechanisms to university archive servers. The findings are based on experiences with existing e-print archives (CERN Document Server (CDS) (4), Los Alamos e-print arXiv (5), University of Amsterdam Beta preprint and Publication Server (6)),

a literature study and attending the 'Workshop on the Open Archives Initiative (OAI) and Peer Review journals in Europe', hereafter referred to as the Workshop (7).

The question underlying this topic is how the research communications infrastructure can be reconfigured to take maximum advantage of newly evolving electronic resources. Rather than 'electronic publishing' which seems to imply a rather straightforward cloning of the paper methodology to the electronic medium, many researchers involved in the innovation of the information chain would prefer to see new technology leading to some form of global 'knowledge network', and for this to happen sooner rather than later.

As argued by Odlyzko (8), the current methodology of disseminating research results and their validation is based on a paper medium that is difficult to produce, difficult to distribute, difficult to archive, and difficult to duplicate. The paper medium requires numerous local distribution points in the form of research libraries. The electronic medium is opposite in each of the aspects mentioned above. Therefore, if the research community were to start from scratch today to design a quality-controlled distribution system for research findings, it would likely be shaped differently. It would be different from the current paper system and different from the electronic clone this paper system would spawn without more constructive input from the research community.

An important incentive to change the research communications infrastructure is the possible large reduction in the costs of this process. By a redesign of the whole process of production, validation and distribution of scholarly information it can be organised in a more efficient way.

Ginsparg (9, 10) has developed a vision of a new infrastructure in which the research communications process can be organised. In his model there are tree levels. At the data level, the model suggests data providers (as defined by the OAI), including, for example, the Los Alamos e-print arXiv (and, by extension, its international mirror network), a university library system (for example the California Digital Library), and a typical foreign funding agency (for example the French Centre National de Recherche Scientifique). These examples are intended to convey the importance of library and international components. Where 'information' is usually comprised of data and

metadata, in the Ginsparg model the second, information level, shows a generic public search engine (like Google), a generic commercial indexer (for example the Institute for Scientific Information), and a generic government resource (for example the US PubScience), suggesting a mixture of free, commercial, and publicly funded resources at this level. At the knowledge layer, the model shows currently a tiny set of examples of physics publishers (the American Physical Society, the Journal of High Energy Physics, and Advances in Theoretical Mathematical Physics) which are more or less overlay journals for the Los Alamos e-print arXiv. In the model, these are the third parties that can overlay additional synthesizing information on top of the information and data levels, and partition the information in sectors according to subject area, overall importance, quality of research, or degree of pedagogy.

The three levels are interconnected in more ways than one. The knowledge level provides in principle much more information than contained in just the author-provided data. The link between the knowledge level - critical here - and the other levels represents how journals of the future can exist in an overlay form, in other words as a set of pointers to selected entries at the data level.

In the Workshop there was general agreement about the way quality control and certification has to be reorganised in the new infrastructure. This corresponds with the suggested strategies of William Arms (11). The first strategy separates peer review from publication. In this model authors publish articles in an e-print archive. Publishers can then provide services, like organising peer review, add an indicator of the quality, and indexing, while the publication itself remains in the e-print archive. Journals than exist as overlay journals by selecting papers on quality in an e-print archive followed by the peer review process. The overlay journal can be represented by a set of links to the e-print archives. For instance: a physicist deposits a paper in the Los Alamos arXiv and notifies the XYZ society. XYZ arranges reviewers who suggest changes. The physicist revises the paper and deposits the revised version in arXiv, noting that the paper has been reviewed by XYZ. XYZ in turn links to the paper through its overlay service.

The second strategy is based on the exchange of quality / certification metadata. The question here is: given a digital object, how can a reader

discover if there is a review or other metadata about its quality? This can be done by creating quality / certification metadata and make the exchange of this metadata possible. In the Workshop, suggestions were made to make quality / certification metadata available for harvesting by means of an extension of the Open Archives protocol. There was strong support for this extension. Given the focus of the workshop on peer-review, concrete actions were suggested to address the exchange of quality / certification metadata using the OAI protocol in a trusted environment. Representatives from the American Physical Society and the Los Alamos arXiv volunteered to participate in a prototype. Actions will be taken by the OAI to facilitate such a prototype and to involve technical experts from the US and Europe. Also quality / certification schemes will be created, building on existing efforts (Dublin Core, W3C), where possible.

However, there are already forms of quality control of the papers to be stored in archive servers. A first mechanism is that the name and reputation of a research institute or department of a university is linked to the submitted paper. Another consideration is that not all submitted documents are suitable for an official publication process (work reports, essays, etc.). A certain indication of the quality will be useful for the reader/user of the archive server. Suggestions for ways to implement validation processes are based on the existing processes by the CERN Documents Server and the Beta Preprint and Publication Server of the University of Amsterdam. The principle suggestions are moderation and refereeing.

With moderation, the author submits a paper electronically to the sever. A moderator (librarian or representative of the research institute) quickly decides whether the report fits in the archive server or not. If the paper is accepted, it will be public available on the server. With refereeing, the document is submitted electronically to the server. It is then kept in a restricted area as long as the referee (research leader) does not approve it. The referee reads the paper extensively and makes suggestions for changes in the text. In the end, if the paper is approved by the referee, it will be publicly available in the server.

The university archive server in ARNO is technically equipped to support both processes. For the future it is recommended that quality metadata is created in both processes and stored as metadata on the

university archive server. The quality metadata on the university archive server must be exchangeable following the future extensions of the OAMH protocol. University archive servers must also be suitable to act as an e-print archive for publishers to produce overlay journals. This calls for a good version management system of the papers in the archive server.

REFERENCES

1. IWI is the Dutch acronym for Innovation of Scholarly Information, http://www.surf.nl/iwi/home.html (in Dutch)
2. http://www.openarchives.org/
3. http://www.uba.uva.nl/en/projects/arno/
4. http://cds.cern.ch/
5. http://arxiv.org/
6. http://preprint.beta.uva.nl/
7. Workshop on the Open Archives Initiative (OAI) and Peer Review journals in Europe, http://documents.cern.ch/AGE/v2_0/fullAgenda.php?ida=a01193
8. ODLYZKO, A. Tragic loss or good riddance? The impending demise of traditional scholarly journals, *Intern. J. Human-Computer Studies*, 42 (1995), pp. 71-122, and ODLYZKO A. Competition and cooperation: Libraries and publishers in the transition to electronic scholarly journals, *Journal of Electronic Publishing*, 4(4) (June 1999) http://www.press.umich.edu/jep/04-04/odlyzko0404.html, and in J. *Scholarly Publishing* 30(4) (July 1999), pp. 163-185.
9. GINSPARG, P. Electronic Clones vs. the Global Research Archive. http://arXiv.org/blurb/pg00bmc.
10. GINSPARG, P. Winners and losers in the global research village, *Conference on Electronic Publishing in Science*, UNESCO: Session: Scientists' View of Electronic Publishing and Issues Raised, Paris, 19-23 February 1996) http://arXiv.org/blurb/pg96unesco.html
11. ARMS, W.Y. Quality control in scholarly publishing. What are the alternatives to peer review?, *Workshop on the Open Archives Initiative (OAI) and Peer Review journals in Europe.* http://documents.cern.ch/archive/electronic/other/agenda/a01193/a01193s4t3/transparencies/Arms.ppt