Institutional Repositories, Long Term Preservation and the changing nature of Scholarly Publications

by

Paul Doorenbosch and **Barbara Sierman**
(Koninklijke Bibliotheek, the Netherlands)

Introduction

In Europe over 2.5 million publications of universities and research institutions are stored in institutional repositories. Although institutional repositories make these publications accessible over time, a repository does not have the task to preserve the content for the long term. Some countries have developed an infrastructure dedicated to sustainability. The Netherlands is one of those countries. The Dutch situation could be regarded as a successful example of how long term preservation of scholarly publications is organised through an open access environment. In this contribution to the Open Repository Conference 2010 it will be explained how this infrastructure is structured, and some preservation issues related to it will be discussed.

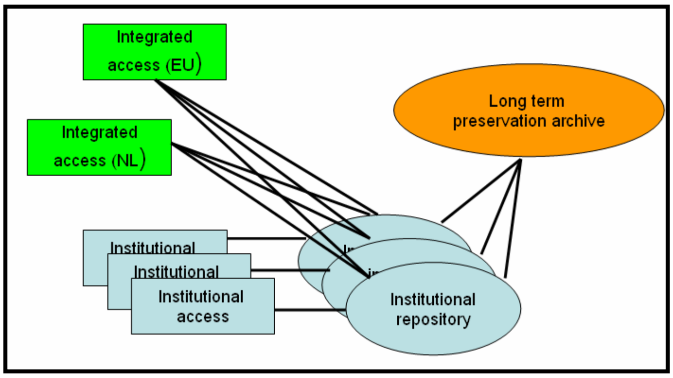
This contribution is based on the long term preservation studies into Enhanced Publications, performed in the FP7 project DRIVER II[1] (2007-2009). The overall conclusion of the DRIVER studies about long term preservation is that the issues are rather of an organisational nature than of a technical one.

The nature of publications in scholarly communication is changing. Enhanced Publications and Collaborative Research Environments are new phenomena in scholarly communication using the wide range of possibilities of the digital environment in which researchers and their audience act. This rapidly changing digital environment also affects long term preservation archives. Raising awareness of long term preservation in the research community is important because researchers are responsible for public dissemination of their research output and need to understand their role in the life cycle of the digital object. At the moment of the creation of the digital object choices are made that will influence the long term preservation changes of the objects. Researchers should be aware that constant curation and preservation actions must be undertaken to keep the research results fit for verification, reuse, learning and history over time.

[1] DRIVER II (Digital Repository Infrastructure Vision for European Research II, WP 4 Technology Watch Report, part 2, Long-term Preservation Technologies (Deliverable 4.3/Milestone 4.2). http://www.driver-repository.eu/
The official report is downloadable at: http://research.kb.nl/DRIVERII/resources/DRIVER_II_D4_3-M2_demonstrator_LTP__final_1_0_.pdf ; the public version is part of *Enhanced Publications : Linking Publications and Research Data in Digital Repositories*, by Saskia Woutersen-Windhouwer et al. Amsterdam, AUP, 2009, p. 157-209; downloadable as: http://dare.uva.nl/aup/nl/record/316849

Infrastructure for Institutional Repositories in the Netherlands

Research universities, universities for applied studies, research institutions etc. in the Netherlands, coordinated by SURFfoundation[2] (the innovation platform for scholarly information and network), have developed open access repositories to make the output of their research community available. In most cases the institutional or university library is in charge of coordination and maintenance. Although every organisation provides access to its own repository, there is also an integrated single access point for this open access material: NARCIS[3]. NARCIS harvests the metadata from the repositories and builds services on it. On a European level the Dutch repositories are harvested by DRIVER. The DRIVER website provides integrated access to the metadata of open access research material in European repositories.



By agreement – the Netherlands has no deposit legislation - between the repositories and *Koninklijke Bibliotheek* (KB, national library of the Netherlands), the National Library is harvesting the publications from all Dutch repositories together with the accompanying metadata, and stores them in the e-Depot,[4] where they are safeguarded for long term preservation and access.

Dutch Organisational Approach to Long Term Preservation

The KB is not the only party involved in long term preservation in the Netherlands. In winter 2009-10 four Dutch organisations, collaborating in the *National Coalition for Digital Preservation* (NCDD)[5], offered a proposal to the Dutch Government on how long term preservation of digital material in the Netherlands could be organised. This proposal intends to make formal what is currently more or less reality (both in the analogue and in the digital world). A division of responsibilities over four organizations is proposed: the *National Library* will take care of textual materials, *Digital Archive and Networked Services* (DANS) of the scientific data, the *National Archive* of national governmental information, and *The Netherlands Institution for Sound and Vision* will take care of the audiovisual material. Making these responsibilities formal is a big step forwards in the organisation of preserving the digital heritage.

Enhanced Publication

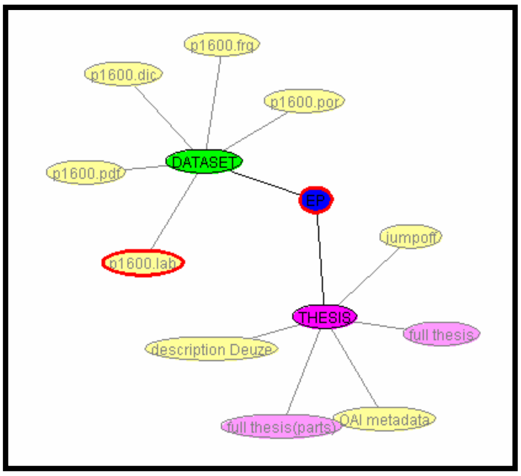
For 'traditional' ways of publication this works well, where research output is most of the time a document of a single nature: text or film or dataset, but that might no longer always be the case. A current development in scholarly publications is the Enhanced Publication (EP) or Compound Publication. In DRIVER II the definition of an Enhanced Publication is: *"Enhanced publications are envisioned as compound digital objects which can combine various heterogeneous but related web resources. The basis of this compound object is the traditional academic publication. This latter term refers to a textual resource with original work which is intended to be read by human beings, and which puts forward certain academic claims. […]*

[2] SURFfoundation: http://www.surffoundation.nl/en/

[3] NARCIS: http://www.narcis.nl/

[4] The e-Depot is the long term preservation environment for publications and other digital material in the KB (http://www.kb.nl/hrd/dd/index-en.html)

[5] NCDD: http://www.ncdd.nl/en/index.php

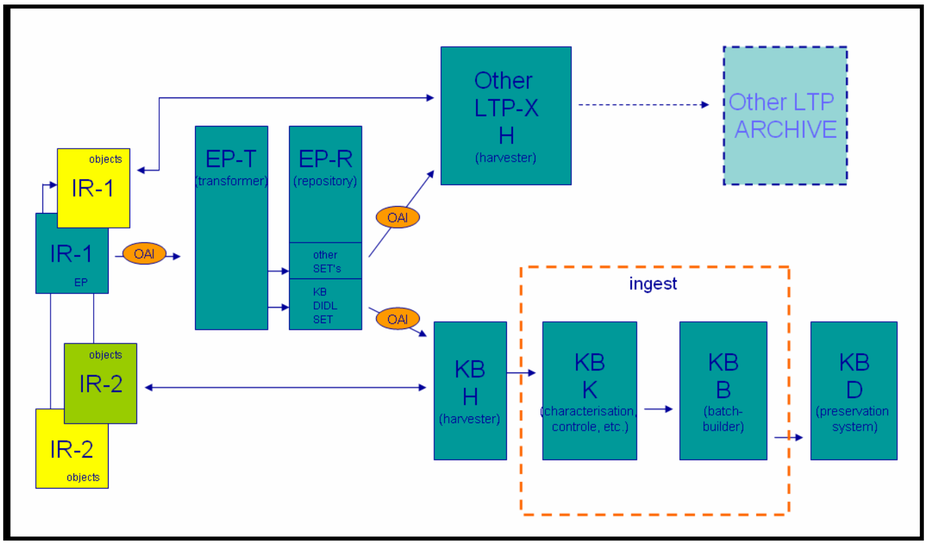


Enhancing a publication involves adding one or more resources to this ePrint. These can be the resources that have been produced or consulted during the creation of the text. […]" [6]

Even during the DRIVER project this definition turned out to be too traditional. An EP can be any combination of files (video and annotation, datasets and documentation, etc.). The traditional publication in digital form is not always the main file of an enhanced object. For practical reasons we kept to our original definition in DRIVER II, but we are aware of its limitations. Nevertheless it served the research goal in the Driver II project: how could we archive such an EP in the existing Dutch infrastructure and what issues to be resolved came out of this research? These issues turned out to be rather generic, independent of the definition.

Demonstrator

In DRIVER II the partners built a harvesting and transformation tool that verifies how we could store an EP in the existing Dutch infra-structure where the responsibility for research data is lies with another organisation than the one which stores the textual material. In this tool an EP is retrieved from a repository and divided into two packages (one with the data, one with the text part) whereby each contains a file with the complete EP structure, including the identifiers of all the consisting parts. One package goes to the "data" archive and one to the "text" archive. Because the OAI-ORE file (the resource map with the structural information and the identifiers) is stored with every part, the EP can be restored for access purposes.



The demonstrator to explain and show the preservation process for simple EPs is available at http://research.kb.nl/DRIVERII/EP-LTP_demonstrator.html. Issues we detected during this research are rather generic for long term archiving, but have some special features because of the nature of an EP.

Issues with the Preservation of Enhanced Publications

The very nature of EPs raised some issues that might influence the activity of preserving these publications for the long term, like rights management, ownership, persistency etc. Some of these

[6] Driver II, D4.2: *Report on Object Models and Functionalities*, 2008, by Verhaar, Peter et al. p. 11-12. http://wiki.surffoundation.nl/display/standards/Objectmodel+Enhanced+Publications

will be described briefly, but the list can be extended without any doubt. Further research on these aspects is needed.

- Ownership. As research is often done with partners from various (international) organisations, an EP can have several owners, sometimes geographically distributed. So first of all creators have to be identified because they own the intellectual property rights. In the Netherlands we assign a unique Digital Author Identification (DAI) to every creator. These numbers are planned to become connected to the VIAF[7] to have a global unique identification for creators.

- Rights. It has to be clear what an archive is permitted to do with the EP on the short term and on the long term, and with its consisting parts, which could all be subject to different legislation. This relates not only to access rights but also to the rights related to preservation actions such as: under what agreement is the copy archived; what actions are agreed to preserve the content and/or the form, etc. Copyright could be held by the author, the organisation the author is working for, or any institution or person the rights holder has transferred his rights to. Currently the KB's solution is to have the institutional repositories declare in the archival agreement that it will only store open access material in its repository, so there is no need for an advanced access rights system.

This approach will only work for a fairly simple and transparent EP. In complicated situations where the different types of data are more intertwined, it is harder to record exactly which part of the publication is under a free access licence and which part is not. Take for example the situation where an EP has a textual part that is already available as open access, but where the related (commercially interesting) data underneath are still not publicly available. In that case it is necessary to record for every part of the publication the exact copyright holder, the national law under which the licence is given, the form of licensing, the neighbouring rights, the clearance by the owner of the publication or by the deposit holder, the exceptions owner and deposit have agreed on, the period for which agreed propositions are valid, etc. This all will bring a lot of work, but it is preferably to do this at the deposit moment and not in a later stage.

- Nature of the consisting files. Is it possible to divide files according to their textual nature or data nature, and could features like location and subject be detected automatically when the EP needs to be reconfigured for access? In case there are separate archives for different types of material, subject or country, a distributing mechanism should understand automatically where to deliver the parts of an EP. Although "type" of object might seem the easiest part to detect automatically, current characterisation tools still have lots of problems with determining the nature of files.

- Versioning/update policies. Data tend to be part of sets that could be subject to change. A 'traditional' text is a unity in itself and it is relatively easy to determine a new version. Modern texts become more and more a patchwork of information units (like a hypertext), that could be subject to updates on its own. A dataset could be changed as a whole, but also the items within a dataset could be changed. An update policy can be based on a schedule, but this will not always be the right method, depending on the nature of the publication. More advanced and differentiated solutions have to be developed

- Authenticity. How can the future user trust that the Enhanced Publication as a whole still is the same as intended when created? A variety of measures need to be taken, like integrity checks, adding metadata about the origin of the parts, etc.

- Persistency. Not only the files themselves need to be kept integer, but also the linking between the files and the identification of the files stored in different places should be persistent [persistent identifiers]). A basic requirement of a long term preservation archive is that it can guarantee the persistency of the ingested files. In the Dutch academic world we use the URN:NBN[8] as the system for persistency of identifiers. A persistent identifier is essential for identification, retrieval, referring and linking. For datasets HANDLE[9] is an upcoming standard in the Netherlands. There is no problem in mixing several systems for persistent identifiers, as long as it is clearly stated which system is used.

[7] VIAF (Virtual International Authority File) http://www.viaf.org/

[8] URN:NBN : http://wiki.surffoundation.nl/display/standards/URN-NBN

[9] HANDLE : http://www.handle.net/

- Future use. A clear vision of the expected use of the archived Enhanced Publication will support shaping the preservation policies, but it is not an easy task to describe the future users and expected use. Preservation is only useful if the goal of preservation is access on the long term. But how will systems and formats develop; what do users in the future expects from information created in our time? What aspects do we need to record to provide the right context for a future user to understand and reuse information from long time ago?

The Consequences of these Issues

Some of the above mentioned issues became very clear during the building of the prototype. For this prototype software we built, the starting point was a rather simple situation: a text and a dataset that could easily be separated and distributed over two separate archives as long as the references between the parts were guaranteed. But soon we needed to conclude that this model was too theoretical model and not be feasible in the long run. By separating the different parts and distributing them over several archives, the Enhanced publication is seriously violated and it will become very difficult to reengineer the distributed files into its original form, despite the persistent identifiers. The only way to solve this issue will be to archive an Enhanced Publication completely in one single archive.

To be able to do that our archival systems need much more flexibility to adept changes, new formats, structural information etc. Apart from the technical challenges, we will need to nuance and adapt the above mentioned national agreements between the four archives to prevent redundancy.

New Developments: Collaborative Workspaces

(Enhanced) publications are the concrete output of a research process. The researcher or the research group determines when the publication will be delivered to a repository for access. Collaborative virtual research environments are considered to be the new workspaces for researchers. Future scholarly communication will take place in this environment and the environment itself could be part of intermediate and end results of scientific research. As a consequence these environments could become the new way of publishing scientific output. Repositories should make connections to these environments to harvest and distribute the scientific outcomes. Long term preservation archives should archive and preserve relevant aspects of collaborative virtual research environments for future use.

But also in that environment it is the researcher who has to decide in the end what is ready for publication and thus for archiving. It is too early to describe how this could be done but archives have to make it easy for a researcher or research group to transfer the scientific data and publications to the long term archive. The easiest way is to archive everything, but with it then comes the obligation to preserve (not only store) and give access to it all. However, it will be unavoidable that the costs of digital preservation will force organisations to select what to preserve. Every selection is a decision about what is now considered valuable for the future without knowledge of those future users. Nevertheless, even if we come to a selection of what is really valuable and relevant for future users, history has taught us that history of science and history of man could only be studied well if also the artefacts that were judged non-valuable in those days could be taken into account. Currently the Dutch SURFfoundation is performing pilots with virtual research environments and is discussing to make the issue of preservation in this environment and the selection of data for future use part of their investigation. First reports are expected later this and next year.

(2010-06-03)

[Acknowledgments/ The authors: page 6]

Acknowledgment
The partners in WP 4 (D4.3/M4.2) of the DRIVER II project were:
Paul Doorenbosch (WP-leader), Barbara Sierman (Koninklijke Bibliotheek, The Netherlands)
Eugène Durr (3.TU, The Netherlands)
Jens Ludwig, Birgit Schmidt (Göttingen State and University Library, Germany)
Maarten Hoogerwerf (DANS, The Netherlands)
The visualisation of an EP on page 2 comes from the EP-demonstrator by DANS et. al
(http://driver2.dans.knaw.nl/)

The authors

Paul Doorenbosch MA, is head Research & Development at the Koninklijke Bibliotheek, national library of the Netherlands. He studied Dutch literature of the 19th and 20th century at the University of Amsterdam. in 2001 he joined the KB as programme manager for the development and realisation of the Dutch national digitisation programme 'Memory of the Netherlands' (http://www.geheugenvannederland.nl/?/en/homepage). In 2005 he was appointed head of KB's Product and Services Development Department and from 2008 until 2009 also interim head National and International programmes. Before 2001 he was employed by the *Royal Dutch Academy of Arts and Sciences as* manager and scientific editor. He is a member of the board of the multidisciplinary programme CATCH (Continuous Access to Cultural Heritage, www.nwo.nl/catch), a joint research programme of computer science, cultural heritage and humanities, and member of the advisory boards of CLARIN (The Netherlands) and D-SPIN (Germany), both large scale infrastructural programmes for the Humanities.
E: paul.doorenbosch@kb.nl

Barbara Sierman MA, is Digital Preservation Manager at the Koninklijke Bibliotheek. She studied Dutch literature of the Enlightenment at the University of Amsterdam. She then joined Pica (now OCLC) as a library consultant. She had various jobs at IT companies as a consultant, last at Cap Gemini. In 2005 she started at the KB at the Research and Development Department. She is engaged in the EU projects PLANETS and DRIVER and participates in international working groups on digital preservation, like TRAC, GDFR, IIPC and JHOVE2. She gave presentations on digital preservation, preservation metadata and organising digital preservation and published several articles on these topics.
E: barbara.sierman@kb.nl

Koninklijke Bibliotheek: **www.kb.nl**