From repository to eternity: from Delft repository to DARE – the developments of OAI in The Netherlands

To meet the growing demand for accessibility of scientific output, a national-level co-operation has been established in The Netherlands to implement local repositories, known as Digital Academic REpositories (DARE). The repository content will be included in the e-Depot of the National Library of The Netherlands (KB) and therefore in their digital preservation strategies, guaranteeing the accessibility for future generations. This article presents the perspectives of both the Library of the Technical University (TU) of Delft repository and the KB on technical issues concerning harvesting metadata and establishing the infrastructure for a national digital preservation programme supported at the local level.

JUST DE LEEUWE
DARE Project Leader
Delft University of Technology

MIRELLA VAN DER VELDE
DARE Project Manager
KB

DARE to share – OAI in The Netherlands

In 2003 the project DARE (Digital Academic REpositories) was started, to shape the co-ordination between individual scientific data providers. The DARE programme is a joint initiative by all 14 Dutch university libraries and the National Library of The Netherlands (KB: *Koninklijke Bibliotheek,* the Royal Library), along with the Royal Netherlands Academy of Arts and Sciences (KNAW) and the Netherlands Organization for Scientific Research (NWO). The first result was the introduction of a national service provider, named DARENet, which has been operational since early 2004. Through the project, partner agreements have been developed addressing technical issues, metadata collection, acquisition, financing and promotional issues.

The Open Archives Initiative (OAI) is in essence an interoperability framework for freely accessible digital archives and, as such, has been adopted by DARE. The metadata in Dublin Core, included by the data providers (institutional repositories) is harvested according to set technologies. These technologies and the first initiatives stem from the e-print movement, intended to quickly and efficiently distribute scientific results among fellow-scientists, largely within the domain of natural sciences. The OAI technical infrastructure has been elaborated in the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH).[1]

In The Netherlands, the OAI-PMH protocol was first adopted in small-scale initiatives with open access publishers and archives. The processing and acquisition of content is now moving from the pilot phase into a structural phase, where production and management aspects are being developed. Content currently includes materials such as research journal articles and digital versions of theses and dissertations, but it may in the future include other digital assets generated in normal academic life, such as administrative documents, course notes, or learning objects.

The role of the KB

As the National Library of The Netherlands, the KB is responsible for the collection, cataloguing and preservation of all publications appearing in The Netherlands. Motivated by the huge and rapidly growing number of electronic publications, the KB began setting up an electronic depot in 1994. After piloting and prototyping, a dedicated deposit system was developed in collaboration

with IBM and has been operational since 2003.[2] This e-Depot is a fully automated system, devoted to large-scale archiving and long-term storage. It is expected that the e-Depot will store several hundred terabytes in a few years. The e-Depot stores electronic publications on the basis of agreements not only with publishers from The Netherlands but also, since 2002, with international publishers.

Storing the digital data is only the first step towards digital preservation. Long-term preservation and permanent access to the electronic publications stored in the e-Depot have become key tasks for the KB. Considering the pace of technological change in hardware and software, this requires constant attention.

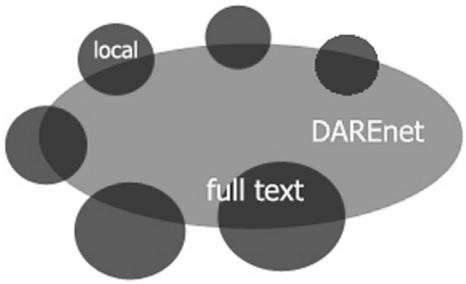
As the initiator of the world's first electronic depot, the KB has a leading role in digital preservation research. A digital preservation department was created in 2003, responsible for the long-term accessibility of the objects in the e-Depot. In order to realize this goal, the department conducts research, participates in a range of (inter)national networks and develops new services for the e-Depot.

The KB carries out extensive research to ensure that the millions of electronic publications stored in its e-Depot remain accessible in the long term. This research is above all aimed at finding ways of ensuring the integrity of the article, but also explores possibilities of making future rendering as user-friendly as possible. All publications are stored in their original file format, which will always be kept. To work on future renderability, both emulation and migration are studied as permanent access strategies.[3]

Research is also focused on the properties of file formats. File formats describe the manner in which information in a computer file is encoded: PDF or MS Word for publications, TIFF or JPEG for images, and so on. In order to determine which strategy for permanent access to follow or which file formats to favour, comprehensive knowledge on file formats is required. For example, consideration has to be given to which features of a PDF are important for the accessibility of the object in the long term; which characteristics should be avoided or encouraged to ensure access into the next decades.

Local versus national

By the end of its first year, DAREnet had accomplished its purpose and demonstrated the usefulness of repositories, permanent storage and open access. Work then started in 2005 to set up a more stable and permanent DAREnet. Currently, DAREnet harvests and makes searchable the metadata of all digital material that is available to everyone from the local repositories, making it searchable. But it limits the harvest to those objects that are available to everyone. Local content, such as material with a non-academic character, is not harvested nationally, nor is material with embargoes or copyright issues. This means that the content of all repositories in The Netherlands adds up to more than the content that can be found in DAREnet. For example, the Delft repository currently contains approximately 3,800 objects, 2,750 of which are harvested. At the moment DAREnet provides access to about 80,000 digital resource objects at 16 institutions.



© SURF-DARE 2004

100,000 new files

Dubbed 'hunDAREd thousand', a new major project was launched in October 2005 in which all participants plan to add a total of 100,000 full-text documents to the DARE ('digital research') archives by the end of 2006. By that time the DARE partners hope to have a total of 150,000 academic publications, dissertations, preprints, datasets and other research material available on DAREnet. The visibility of this metadata means that these documents can also be found through other search engines.

Delft dissertation site to Delft repository

A number of Dutch institutions (such as the University of Utrecht and the University of Amsterdam) have a longer history of open access materials than that of the DARE initiative. At the

library of the Technical University (TU) of Delft, these developments are relatively new. A pilot with e-theses that was set up at the library in 2003 was recently expanded to a broader website that includes more document types. Currently, more than 1,800 Delft doctoral theses are online and by August 2006 almost all theses since 1990 will be online as the result of a large retro-project encompassing around 3,000 doctoral theses.[4]

The introduction of the Delft Repository fits within the theme of ever more important e-services for clients and the demand for fast dissemination of content for all users. In recent years the rapid growth of digital resource management can be observed within the library, of which the growth of e-journals has been the most spectacular: within just a few years, a few hundred online subscriptions has grown to 10,000 titles, of which a large number are e-only. The tools to manage digital information for clients have been set up as Virtual Knowledge Centers.[5]

These digital services have further developed because of the open access movement. Research is now being done on how to integrate open access titles in the catalogue within the profile used by technical services, and how the TU Delft Repository can characterize its position within the university. Starting and managing the repository is a very labour-intensive process: developing the technical infrastructure, retraining personnel, establishing and maintaining consistent metadata, acquiring content and communicating with a proactive attitude towards the scientists.

Library building, TU Delft

KB: repository for eternity?

Special project within DARE: *Promise of Science*

After the successful launch in the spring of 2005 of '*Cream of Science*', which provides access to the publications of over 200 leading Dutch academics, there was an increased focus on new researchers.[6] Every year about 2,500 theses – about five percent of the country's entire academic output – are published. DAREnet is making these theses visible under their own heading, '*Promise of Science*'.[7] Doctoral theses are the showpiece of a university. They usually represent the most recent evolution of ideas by young researchers in their fields, backed up detailed data. The creation of a virtual showcase for universities using doctoral theses not only helps shape their profiles as well as those of the graduates in question, it also enhances access to previously elusive research results. *Promise of Science* is therefore an extension to *Cream of Science*, the partial collection already in place. Alongside 'established' researchers, focused attention is now being given to 'up-and-coming' research talent.

The road to the KB

With the repositories of all 16 universities and scientific institutions basically in place, and the e-Depot at the KB operational, is was necessary to make the transfer from the repository to the 'safe place' at the KB. In order to accomplish that, a number of standards had to be established. As described, the OAI protocol for metadata harvesting had already been set as a standard. However, transferring the object files described by these metadata present more difficulties. Actual harvesting techniques for qualified objects are not widely

implemented in the world but, luckily, important work is being carried out elsewhere to solve the challenges involved.[8] MPEG21-DIDL was adopted as a standard, having evolved from the OAI-PMH framework and offering a workable framework.

The KB and the other participants are now working towards an optimal implementation of data harvesting and all but one of the participating institutions have their DIDL-output in place. All metadata and relevant objects available have been harvested through OAI-PMH/MPEG21-DIDL. Since MPEG21-DIDL offers a framework rather than a rigid set of rules, the DARE community set about agreeing on the specifications of the XML to be rendered, in order to enable automated processing of tens of thousands (and growing) publications within the e-Depot. The agreements were made step by step, starting with 'The Simplest Thing That Could Possibly Work'.[9] Since 80% of the current objects are not complex objects, yet consist of one object file and one metadata file (in XML), this led to a DIDL-structure consisting of three separate parts: the metadata, the object itself (by reference) and the 'splash page', used to render information on the author, arrange copyright issues and such. The use of the splash page is not obligatory. It is expected that in the future, more complex structures will be included in the repositories. The DIDL structure is currently being expanded with rules concerning structural metadata, the metadata describing the order in which items of a complex object should appear.

The current holdings of the e-Depot are voluminous, yet relatively homogenous; large volumes of journals in a strict hierarchy of volume, issue, article, with generally similar formatting and specifications. With the introduction of DARE in 2003, the KB found itself set for a new batch of exciting challenges in the area of technical specifications and digital preservation, most of which would otherwise have been encountered at a later stage, yet due to the DARE programme received a large boost. The material is extremely heterogeneous, in that it stems from 16 institutions, each consisting of many different faculties with their own rules concerning formatting, such as the layout of a PDF. Another example is the interpretation of the DARE use of Dublin Core, detailed in an extensive document, yet prone to variable interpretations and, sometimes, misinterpretations. Since this makes automatic processing impossible, the KB has now built in a quality check, implementing Dublin Core by the rules set within the DARE community.

The way back

DARE objects stored in the e-Depot can be returned to their original institutions at a later date, in case an institution can no longer read the original format or when objects have been lost due to an emergency. In order to deliver objects from the e-Depot to the original owner, a number of functions have been implemented. Regular access to publications in the e-Depot consists of visitors downloading one article per request while they are actually within the KB. In order to enable the extended functional requirements, an authentication and authorization component, as well as a batch delivery component, have been added to the e-Depot system. These components fulfil the functional needs of this new service: a batch of objects can be delivered upon one request, and moreover, the rules of authentication and authorization are elaborated so that only those institutions authorized to do so can retrieve (part of) their collections when necessary. Currently, the KB is discussing with a number of DARE-partners, including Delft, the terms of delivery for the future. Which are the criteria the repositories will want to use to receive their publications back from the e-Depot? Will a formatted request ("We can no longer read our PDF 1.4") and date ("Can we have all publications from 1 January 1992 to 31 December 1999 back?") suffice, or are there other search criteria to be included? Is a month, a week, or a day acceptable for receiving the objects? All these criteria will be specified in a Service Level Agreement that is currently being drafted.

Conclusions

The institutional repositories of Dutch universities are maturing, although there are still serious differences in speed and implementation. This year considerable progress will be accomplished with an increase in content. Local responsibility for the repositories runs parallel to the national control function which is carried out by DAREnet. These dual efforts are producing generic solutions for intense problems such as further defining metadata, establishing service layers, and aspects of

communication. Co-operation between the universities and the KB is essential and allows the content of the repositories to be preserved for future use, establishing a firm foundation for the secure management of Dutch scientific output.

References and notes

1. www.openarchives.org/OAI/openarchives protocol.html
2. Oltmans, E. and Lemmen, A., The e-Depot at the National Library of The Netherlands, *Serials*, 2006, 19(1) 61–67. DOI: http://dx.doi.org/10.1629/1961
3. http://www.kb.nl/hrd/dd/dd_projecten/projecten_intro-en.html
4. http://repository.tudelft.nl
5. http://www.library.tudelft.nl/ws/a/knowledge_ centres
6. In this project, more than 200 prominent Dutch scientists make their scientific publications visible. There are 46,000 documents, of which 60% are full-text available. The remainder either have copyright limitations or could not be retrieved. http://www.darenet.nl/en/page/language.view/keur.page.
7. Promise of Science: http://www.DAREnet.nl/promiseofscience
8. Van de Sompel, H., Nelson, M.L., Lagoze, C. and Warner, S., Resource Harvesting within the OAI-PMH Framework. In *D-Lib Magazine*, December 2004: www.dlib.org/dlib/december04/vandesompel/12vandesompel.html
9. www.artima.com/intv/simplest.htm

■ **Just de Leeuwe**
License Manager
DARE Project Leader
Delft University of Technology
Prometheusplein 1
2628 ZC Delft, The Netherlands
Tel: +31 (0)15 2787813
Fax: +31 (0)15 2572060
E-mail: j.deleeuwe@library.tudelft.nl

■ **Mirella van der Velde**
DARE Project Manager
Research & Development
Koninklijke Bibliotheek
National Libray of
The Netherlands
Prins Willem-Alexanderhof 5
2595 BE Den Haag, The Netherlands
Tel: + 31 70 314 0958
E-mail: mirella.vandervelde@kb.nl
Website: http://www.kb.nl/
DARE: http://www.kb.nl/hrd/dd/dd_projecten/projecten_dare-en.html