Moving theory into practice. How can research data archives benefit from digital preservation research?

René van Horik
DANS – Data Archiving & Networked Services
Anna van Saksenlaan 10
2593 HT Den Haag
+31 70 3446484
rene.van.horik@dans.knaw.nl

Peter Doorn
DANS – Data Archiving & Networked Services
Anna van Saksenlaan 10
2593 HT Den Haag
+31 70 3446484
peter.doorn@dans.knaw.nl

ABSTRACT

It is a challenge for research data archives to establish and maintain robust, long-term "trusted digital repositories" while keeping track of new insights from digital preservation research. This paper discusses the way the Dutch research data archive DANS intends to conduct research into digital preservation in order to improve its data management services. Moving from theory into practice concerns the setting up of a research program that has the ambition to stay informed on the latest insights into digital preservation and pays attention on the way these insights can improve the quality of data archiving services.

Keywords

Data archive, Data archiving, Digital preservation, Trusted Digital Repository, Research agenda

1. INTRODUCTION

An important activity for research data archives is to establish "trusted digital repositories" that facilitate the durable storage of and access to research data sets. This digital preservation infrastructure has to meet quality standards as set by the digital preservation community and facilitate the daily data archiving routines. As digital preservation research covers a wide range of topics and often provides alternatives for a given problem it is a challenge for data archives to select relevant research topics, to allocate resources to carry out the research and ways to translate the newly found insights into practical solutions.

In this paper, we discuss the eResearch program of DANS[1] with an emphasis on the activities related to digital preservation. The eResearch program, entitled "Exploring the Long Term Availability of Research Data", spans the period 2012 – 2015 and is thus at its beginning.

2. Research themes

A number of models exist that illustrate the cyclic character of research data management and the data activities that are related to it[2]. These models can provide a framework for research activities for research data archives.

The themes of the DANS research program are structured around the idea of the life cycle of research data. We see the cycle of data production, data analysis and data archiving as integrated parts of the cycle of scientific knowledge production.

We identify three research themes: (1) data archiving, (2) data cultures, and (3) data processes – from creation, enrichment and exploitation to analysis. All research themes contribute to one or both of the two main focus points of research at DANS as research data archive: digital preservation and enhancing access to digital data.

2.1 Drivers for research on data archiving

Data archiving concerns the durable storage and long-term access of research data. Sharing and reuse over the long-term of data is a theme mentioned in several research agenda's, see e.g. [1].

Three drivers guide the research activities of DANS on data archiving. First of all the research builds upon the existing knowledge base of procedures and guidelines as used to archive research data in the day to day practice. An example is the assessment of the preferred data formats used by the data archive and that are considered as durable.

Secondly, data archivists are consulted to define research topics that will contribute to the solutions for specific data archiving problems. An example is the archiving of software for which currently no solution exists.

In the third place research in the framework of existing (international) projects is driving the research activities on data archiving. DANS participates in the EU project APARSEN[3] that provides resources to stay informed on research in the field of

[1] See: http://www.dans.knaw.nl/en. DANS - *Data Archiving & Networked Services* - promotes sustained access to digital research data in the Netherlands. For this purpose, DANS encourages scientific researchers to archive and reuse data in a sustainable manner, by storing them in a trustworthy repository such as the online archiving system EASY (see: http://easy.dans.knaw.nl. DANS also provides access, through the NARCIS system (see: http://www.narcis.nl) to thousands of scientific datasets, electronic publications and other research information in the Netherlands. Moreover, DANS provides training and advice and performs research into sustained access to digital information. DANS provides information services to researchers from all disciplines, although its expertise is primarily in the (digital) humanities and social sciences.

[2] E.g. the DCC life cycle model, see: http://www.dcc.ac.uk/curation/curation-lifecycle-model

[3] See: http://www.aparsen.eu

digital preservation and data archiving. An example is the activities that are carried out in the field of the auditing and certification of trusted digital repositories.

3. eResearch for research data archives

eScience or “enhanced science”, is commonly understood as the “new science” of the digital age. In [3] the data-intensive character of eScience is motivated. In line with these changes in the scientific landscape, research activities related to research data management can be characterized as eResearch.

The formulation of a research agenda seldom starts from scratch. Also the formulation of the eResearch program at DANS is based on activities related to data archiving that currently are carried out by DANS or that were finalized in the recent past.

Two types of activities are distinguished. In the first place the exploration of the research community. By consulting the research community DANS tries to gain insight in the demands and wishes concerning the role of research data. Long-term access and usability of research data is an important aspect of this. Secondly, the involvement of DANS in the development and implementation of a number of digital preservation tools and services is described. The overview of activities related to data archiving puts them in the broader scope of the lifecycle of research data.

3.1 Exploration of the user community

The main research communities served by DANS are scholars active in the humanities and social sciences. This type of cooperation goes back to the 1960’s when the first social science data sets were created and archived. These data sets are part of the data archive collection curated by DANS.

Consultation activities are carried out to determine the directions to enhance existing services or to develop new services, also in the field of digital preservation. Examples of consultation activities are the interviewing of key persons active in a scientific discipline and the execution of a survey on the role of data in the scientific discourse The results of the exploration on the role of data in the psychological discipline in the Netherlands are published in [6].

A recent exploration of the role of data in the exact sciences (see [7]) made clear that despite the fact that large-scale data processing and data storage facilities do exist, the durability of the data is suffering from fragmentation in tasks and responsibilities. In many cases, awareness of the importance of data management is still in its infancy. The development of data management services for new scientific audiences is an important driver for the research activities of DANS. The involvement in the design and execution of a course on data management is an example of this. Also the long-term storage of data sets that increasingly grow in size is a recurring issue in the explorations. The durability of grid and cloud storage of big data sets is an emerging research topic.

3.2 Tools and services for data archiving

This section contains five examples of tools and services that contribute to the durability of research data and that are the result of research activities carried out by DANS or by DANS in cooperation with other parties. The tools and services form the basis for further research on digital archiving as part of the DANS research program.

3.2.1 File format migration

Obsolescence of file format specifications is one of the main factors that threaten the long-term access to digital data. The aim of the MIXED[4] project is to develop a sound theoretical framework and tool to convert binary file formats, which might get obsolete, upon ingest in a data archive in a durable XML representation. Upon dissemination this XML file is converted from this generic format into a vendor format of choice. More information on the MIXED tool and project can be found in [8].

3.2.2 Audit and Certification of repositories

The auditing and certification of data repositories contributes to the long-term usability of data objects stored in these repositories. DANS took the initiative to develop a “low-threshold” audit and certification framework that consists of a number of quality guidelines to support the digital durability of repositories. This framework is called the DataSeal of Approval (DSA)[5]. Research is carried out to formulate the quality guidelines of the DSA and to create an international community that supports the implementation and governance of the DSA.

Within the APARSEN project DANS and other research data repositories underwent an audit based on the ISO16363 standard (Audit and Certification of trustworthy digital repositories). A report of the audit process can be found in [9].

3.2.3 Persistent identifiers

Persistent identifiers enable the creation of trusted and sustainable references to scientific resources in repositories. DANS is involved in initiatives to create an infrastructure for the assignment and resolving of persistent identifiers based on two components: the Uniform Resource Name (URN), a unique and permanent identifier for electronic resources on the internet, and the National Bibliographic Number (NBN), assigned by national libraries.

3.2.4 Peer review of research data

Peer review of publications is seen as an instrument for assessing the quality of the research results. DANS carried out a pilot study to extend the peer review to research data. This will make the research data much more visible and it provides a way to evaluate the degree to which the data are fit for re-use. A report of the study on the peer review of research data can be found in [10]. The pilot results have convinced DANS that peer review of open data in an archival context is feasible and yields valuable information for a large audience.

3.2.5 Cost models for digital archiving

Financial sustainability is an important attribute of a trusted, reliable digital repository. Research was carried out to gain insight in the cost issues relating to running the DANS research data archive. The outcomes of the research can be found in [11]. The “activity based costing (ABC) model” was used for estimating the costs of preserving digital research data. This model has been tested on empirical cost data from activities performed by all DANS employees. This resulted in the “euros per dataset” unit of cost measurement. The outputs of this model were connected to the strategic goals of the organization. This resulted in a detailed insight in the labor intensity of activities in DANS. It became clear that only a small portion of the activities and costs could be assigned to preservation activities. Project management and

[4] See: http://mixed.dans.knaw.nl

[5] See: http://www.datasealofapproval.org

administrative support e.g. are also labor and cost intensive activities.

3.3 Moving theory into practice

The research themes, the drivers for the execution of research activities and the examples of tools and services as described above are illustrative for research data archives. They form the theoretical basis for further explorations into the long-term usability of research data. Currently the DANS research group consists of about 10 members, active in a number of projects in the field of eResearch. To a considerable degree these projects provide the means to carry out the research activities. Phd candidates are active in the research group related to projects that are carried out in close cooperation with Dutch universities[6]. Next to the resources provided by projects, the research group has resources provided by general funding means.

4. Conclusion

The requirements made by researchers with regard to data services are continuously changing at a rapid pace. The DANS eResearch group is formed within the institute to maintain and improve the level of services. The research activities always have an applied component. Digital archiving is one of the research themes for the coming years.

Explorations of research practices in specific scholarly communities and consultation of professionals active in research data archives provide guidance with respect to the formulation of research questions. A number of research issues as presented in this paper will be further explored.

Despite the fact that a research on digital preservation of research data is work in progress we can distinguish a number of general issues that will guide the next steps. They regards to tools and methodologies that are needed to enable the long-term storage of research data as well as the evaluation of them. Also insights from computer science and information science will be transferred into digital data archiving practices. A third focus point of research on data archiving concern the further professionalization of data archivists.

5. REFERENCES

[1] Faniel, I. M. and Zimmerman, A. 2011. Beyond the Data Deluge: A research agenda for large-scale data sharing and reuse. In *The International Journal of Digital Curation*. Issue 1, Volume 6 (2011). http://www.ijdc.net/index.php/ijdc/article/view/163

[2] Hey, T. Tansley, S. and Tolle, K. (2009) *The Fourth paradigm. Data-intensive scientific discovery*. Microsoft Research, Mountain View

[3] Voorbrood, C. 2010 *Data – Voer voor Psychologen. Archivering, beschikbaarstelling en hergebruik van onderzoeksdata in de psychologie*. DANS Studies in Digital Archiving 4. http://www.dans.knaw.nl/content/categorieen/publicaties/dans-studies-digital-archving-4

[4] Dillo, I, and Doorn, P. 2011. *The Dutch data landscape in 32 interviews and a survey*, DANS (2011). http://www.dans.knaw.nl/en/content/categorieen/publicaties/dutch-data-landscape-32-interviews-and-survey

[5] Horik, R. van and Roorda, D. 2011. Migration to Intermediate XML for Electronic Data (MIXED): Repository of Durable File Format Conversions. In *The International Journal of Digital Curation*. Issue 2, Volume 6 (2011). http://www.ijdc.net/index.php/ijdc/article/view/195

[6] *Report on peer review of digital repositories*. Report of the APARSEN project (2012). http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2012/04/APARSEN-REP-D33_1B-01-1_0.pdf

[7] Grootveld, M. and van Egmond, J. 2011. Peer reviewed open research data: results of a pilot. *Proceedings of the 7th Digital Curation Conference* (2011) http://eprints.rclis.org/bitstream/10760/16668/1/Grootveld-vEgmond.pdf

[8] Palaiologk, A. Economides, A., Tjalsma, H. and Sesink, L. 2012. An activity-based costing model for long-term preservation and dissemination of digital research: the case of DANS. In *International Journal on Digital Libraries* 2012. DOI = http://dx.doi.org/10.1007/s00799-012-0092-1

[6] An example is the computational humanities research project “CEDAR - Census Data Open Linked” aimed to realise semantic access of data from the Dutch census of the last two centuries. See: http://cedar-project.nl/