Purdue Libraries

International Association of Scientific and Technological University Libraries, 31st Annual Conference

Purdue Libraries *Year* 2010

Building a 'data repository' for heterogenous technical research communities through collaborations

Jeroen Rombouts* Alenka Princic†

*Delft University of Technology, j.p.rombouts@tudelft.nl

†Delft University of Technology, a.princic@tudelft.nl

BUILDING A ‘DATA REPOSITORY’ FOR HETEROGENEOUS TECHNICAL RESEARCH COMMUNITIES THROUGH COLLABORATIONS

Jeroen Rombouts and Alenka Princič

TU Delft Library, Delft University of Technology, Prometheusplein 1, 2628 ZC Delft, Netherlands.
T: +31 15–2787816, F: +31 15-25 72060, E: j.p.rombouts@tudelft.nl, W: www.library.tudelft.nl

Abstract
The paper describes the project ‘3TU.Datacentrum’, an initiative of the libraries of the three Dutch Technical Universities. Its aim is to build a data curation facility for the improvement of data management, providing data curation services for data producers of the Technical Universities and enabling data reuse.

The libraries initiated this project in their function as information- and knowledge brokers in emerging e-science. Playing a role in the management of research data offers an opportunity to get more thoroughly involved in the scientific process and to interconnect research data with all other sources managed by the libraries.

The project builds on the experience from previous TU Delft research (E-Archive and Darelux). Initial interviews with managers and researchers in leading research areas of the Technical Universities were followed by in-depth investigation of the benefits and barriers for data producers. Additional work with research groups in technology- and engineering science confirmed the need for collaboration in data management. Data producers from these heterogeneous research communities identified benefits for data producers in three clusters: quality improvement, increase in research impact and efficiency (saving time on administration).

Building the data curation infrastructure and setting up the data librarianship were the primary challenges for the library staff. In collaboration with national and international ‘colleagues’ of the data center the project is currently expanding the data set collection and selecting and developing formal license agreements, guidelines and tools, data citability, as well as selection and usage criteria for long-term access to and preservation of research data.

Preliminary conclusions are that while the data curation principle is easily adopted, the data producers will not as easily invest their time in data archiving. Building a data curation facility to meet the diverse needs of heterogeneous research communities requires considerable efforts that can only be realized by (inter)national collaboration between data centers and data users.

Keywords: data repository, data center, data curation, data management, ‘3TU.Datacentrum’, data librarianship, heterogeneous research communities.

Introduction
Sharing scientific data through publication is the dominant building stone of the discovery process and the dominant means by which scientists can earn credits for their discoveries. Recently, science has changed and is migrating slowly but inevitably to e-science. New technologies generate very large data sets that can no longer be adequately represented in an article. Interdisciplinary collaborations and research communities are being formed addressing sharing data and resources. Along with these changes, the terms 'digital curation', 'digital data management' and 'data repository' are finding their place in daily use. More and more often these days, research success is measured not only by the publications produced but also by the data it generates (Nature, 2009). More and more funding agencies now require proof of good data management practices along with the grant proposals. These data must be made available to the large scientific community.

Digital curation! A general definition of *digital curation* is the selection, preservation, maintenance, and collection and archiving of digital assets ("Digital curation", 2010).
Research data, however, are viewed as a distinct sector with specific characteristics. One important difference with, for example government data, are the reasons for preservation (Tjalsma, 2010). Basically we distinguish three reasons to preserve research data for the long term:

1. Re-use within or outside the research discipline in which the data were created. Also often described as secondary use.
2. Verification of data on which publications are based. Existing codes of conduct for research often prescribe keeping available the data for verification for a, mostly limited, period, like The Netherlands Code of Conduct for Scientific Practice.
3. Historical research, in particular for the history of science, or cultural heritage.

Because the business case for investing in digital curation of research data is mainly based on re-use the project targets are wider then just digital curation according to the definition above. Furthermore acquisition of data is crucial for the early stages of developing a data curation facility as is the definition of services. Therefore we use the 'ANDS Data Sharing Verbs' (Burton and Treloar, 2009) to illustrate the experiences of the 3TU.Datacentrum combined with more familiar data curation terms from the OAIS model.
The ANDS verbs are: Create, Store, Describe, Identify, Register, Discover, Access and (Re)use.

What exactly research data means is not yet very well defined; the authors view at the moment is that research data means any output from research projects that is not a publication. A very clear property of research data as used in this paper is that we mean only digital data; any non-digital data are not taken into consideration.

Whereas it is known that some research communities have been quite open to sharing for a long time, such as GeneBank, Ribosomal Database Project or Protein Data Bank in molecular biology; arXiv.org by Cornell University in Ithaca, New York in mathematical and computer science; or the International Virtual Observatory Alliance in astronomy community; to mention only a few, there are no examples of large-scale institutional data repositories for versatile science disciplines consisting of small research groups. One of the reasons might be, as stated in the scarp synthesis report (Key Perspectives Ltd, 2010), institutional data repositories will face a considerable challenge to serve their communities due to the different data-related needs and expectations of researchers working different disciplines.

Organizations in the UK, for example, have made a good start in digital data management. The Joint Information Systems Committee (JISC), established by seven research councils already back in 1993, has made data sharing a priority (www.jisc.ac.uk). JISC also helped establishing a Digital Curation Center at University of Edinburg, as a national focus for research and development into data issues.

The Dutch landscape
Other European agencies have also pursued initiatives in digital data curation. So how is this issue being addressed in The Netherlands?

The Netherlands Coalition for Digital Preservation (NCDD), was established by a number of public sector organizations actively involved in digital preservation across the sectors government, scholarly communication and culture & heritage. This Coalition will act as a catalyst and joint platform for sharing expertise and advocacy issues (Angevaare, 2009).

Another Dutch initiative is Data Archiving and Networked Services (DANS) (www.dans.knaw.nl), an institute under the auspices of Royal Netherlands Academy of Arts and Sciences (KNAW), which is also supported by the Netherlands Organization for Scientific Research (NWO). DANS is not related to any university and has a focus on storing and making research data permanently accessible in the arts and humanities and social sciences. In the Netherlands, it is (also) still believed that natural sciences and engineering are taking good care of their own research data.

The Dutch SURF Foundation (www.surf.nl) is an agency comparable to JISC and aims to create a common infrastructure to facilitate access to research information. Their current program finances small projects to survey the needs and requirements concerning research data. It also provides a forum for knowledge exchange between people involved with research data.

A variety of institutions can be involved in data curation if research is to flourish due to readily available data collections. Ideally, data management should make part of every course in science (Joy Davidson, personal communication). But who should host these data? Certainly, these should be institutions that can take responsibility for preserving digital assets and making them accessible over the long term. The university libraries that can provide for robust long-term funding are obvious candidates to take on this role. University libraries are at the intersection between research producers and consumers, between scientists and public. The mission of TU Delft Library is to function as the hub of knowledge exchange for technical and scientific information in the Netherlands. This library in particular supports not only research and education within the university faculties but also at the national level, as appointed by the national government for natural sciences and engineering.

Playing a role in the management of research data offers an opportunity to get more thoroughly involved in the scientific process and offer even better support to the researchers. It connects research data with all other sources managed by the libraries, such as for example, the libraries e-journal collection or publications in an Institutional Repository, which renders this institution a modern, innovative and prestigious digital library (‘e-library’) for the world of science and technology, both in the Netherlands and abroad.

There are three universities of technology in the Netherlands: Delft University of Technology, Eindhoven University of Technology and the University of Twente. In 2007 they joined forces in the formation called 3TU.Federation. Temporary funding from the national government facilitates the forming of the 3TU.Federation.The federation aims to maximize innovation by combining and concentrating the strengths of all three universities in research, education and knowledge transfer. For efficient operational processes, projects are being executed within the 3TU.Federation focusing on two main aims: Facilitation, providing the best possible support for joint activities, and standardization and shared services to increase efficiency in operational management.

The ‘3TU.Datacentrum’ (for the purpose of this paper we use a term ‘Data center’) started in 2008 as a three-year project under auspices of 3TU.Federation. The project is a collaboration of the libraries of the three universities who view their role in data management as a natural addition to their activities as knowledge-brokers for their universities.

The 'Data center'
Researchers at the three Dutch universities of technology expressed the need for sustainable data management. This was shown by a questionnaire conducted in 2006. The registering of data sets was seen as a weak point as it strongly depends on personal input. Concerns were raised that '*one size does not fit all'*. However, it is expected that advantages will be obtained from a structured method of storage and with regard to the collaborative services (the setting up of a data lab) a number of respondents saw opportunities particularly for EU-research projects.

The 'Data center' is both a project and a strategic focus for the libraries of the three Universities of Technology in the Netherlands. The mission of the 'Data center' is to become the foremost facility for the permanent accessibility of technical-scientific research data in the Netherlands. The focus is on (national and international) programmes and projects in which Dutch research groups are involved. The 'Data center' also provides access to technical-scientific data stored elsewhere in the world.

The project's aim is to build a data curation facility for the improvement of data management, enabling science data reuse and preparing data for preservation. By 2011 the libraries want to have gained experience with:

- A data lab, where data producers can store, process and share data for current research. Data sets are stored in formats defined by data-producers. Sharing, versioning and storing will be facilitated to support scientists. Added services, like advice on data management, metadata (standards), licenses are provided by the 3TU.Datacentrum staff to ease possible transition to the data archive at a later stage;
- A data archive, where data producers can deposit and data consumers can retrieve data sets.
  At the moment data is accepted, original files will be kept in the archive with additional preservation description information, in time data sets too big to store multiple versions are expected and probably formats which are difficult to preserve or convert. By then recommendations on preferred formats and different levels of preservation need to be defined;
- Data services, like training, data management support and assistance in locating to data stored elsewhere in the world.

The project 'Data center' builds on previous research at TU Delft Library as well as initial interviews with managers and researchers in leading research areas of the Technical Universities and in-depth investigation of the benefits and barriers for data producers. This research confirmed the need for collaboration in data management in technology- and engineering science. Data producers from these heterogeneous research communities identified benefits for data producers in three clusters: quality improvement, increase in research impact and efficiency (saving time on administration). The main statements collected among the producers and users were 'the existence of a data center is justified by the frequency of reuse of archived data', and 'the reuse is dependent on accessibility, on the support offered to data producers and data users and on the *value* (quality) of the data collection. The researchers were of the opinion that the quality of the data should be a responsibility of the researcher while the data center should provide tools for automated quality control. Altogether, public access to the data sets should lead to a better communication among researchers and efficient research methods without unnecessary duplication of experiments. Hence, incentives for the data producers to deposit data, were seen as an essential parameter for building significant data collections.

Naturally, not all the data produced can be and should be archived in a data repository. But who can or should set criteria and implement quality control for this purpose? Preliminary research in these matters confirmed that currently only for a very small number of disciplines selection criteria are being applied. Permanent, and open, access to research data is certainly not yet achieved in large areas of the Dutch research world. Selection of data for long term preservation does not seem to be the most urgent issue among the researchers most academic disciplines. Even more, making selection decisions does not seem to be the major concern even among the existing data archives

and repositories. Inevitably, this will change, as some managers of repositories have indicated: 'selection decisions might become an increasingly important topic when collections are growing, regardless whether these decisions are to be taken at the moment of creation of the data, at the ingest (transfer into a repository) or years later'.

Experiences

The 'Data center' project encountered a range of the challenges and diverse needs of the data producers within the first two years of setting up a data curation facility for the three Dutch universities of technology. This was acquired by processing a number of cases and other significant questions. Each case represented particular challenge, needed detailed investment form the project's side and brought to light specific benefits. Some of these are briefly presented below and in Table 1.

Table 1: Overview of cases in the first two years of setting up a data curation facility, their versatile needs, investements, challenges and benefits.

	Needs	*Investments*	*Challenges*	*Extra benefits*
Combustion experiments	Distribution support Pilot data set	Submission (acquisition)	Store (archival format)	Improved accessibility High quality set from leading research group
Hydrology observations	Interoperability and long term access	Quality assurance Archiving (data model, conversion)	Create (quality assurance)	Improved quality and discovery Work with data producers
Stevin lab	Ease documentation and collection building	Archiving (conversion)	Describe (automate generation of descriptive information)	Potential demonstrator for efficiency benefits to data producers
Drizzle radar measurements	Sharing large data sets and preservation	Archiving (size testing and storage capacity)	Store (size of data)	Experience with large data sets
Transmission Electron Microscope images	Sharing data with publication and easy citation	Description and submission (model)	Identify (link to publication)	Increased visibility of data set and repository
Marine & coastal collaboratory	Preservation (mirror data) and register data	Submission (setting up sync between collaborator and repository)	Register ('releasing' data to repository)	Share knowledge of science tools and practices. Co-developing integrated services.
Jet-ski data	Registration of pre-publication data	Description (meta data and registration)	Discover (reuse and intellectual property claims)	Potential for high reuse and demonstrator pre-publication benefits of sharing data
Anthropometric data/Benchmark log files	Sharing and preservation	Submission (support creation and description)	Create (support documentation/ interactive elements)	Potential increase of data repository visibility

The cases spanned from simple, small experimental data sets to large and complex ones. The combustion lab experiments case was about published data from a finished research project that was shared with any data consumer upon request. Descriptive information and quality assurance were on a high level. The main challenge was selecting a model and format for storing the numerous small files of this one-time data set. The main investment of the data center for this set was receiving the submission; only after several discussions explaining the data center's goals and services, the data set was offered for submission. Nevertheless, the combustion experiments were an interesting case revealing some hesitation, concerns and perhaps skepticism or maybe just unawareness of the data producers with regard to the services available or the benefits of enabling open access to the research data.

On the other hand, floods of bites such as in the 'Drizzle radar measurements' represented yet another challenge. This ongoing research with radar measurements of the rain clouds produces large data sets with high demand on long term preservation for climate studies. The data sets

consist of continuous processed data with samples of raw data; the research group wanted first of all a reliable backup for their data and also wanted to share, or show, their data but did not have an infrastructure available to do so. Apart from a copy of the data the centralized storage and OPeNDAP server of the Data center enabled sharing, discovery and querying the data sets. The data are characterized by high-level descriptive information and quality assurance. The main investment for this set from the data center side was tests with file size and acquiring storage capacity while the challenge still is developing an easy way for delivering the data from the measurement site to the data center.

Several requests related to the publishing of data with a publication. Perhaps the most interesting case was the data set from the 'Transmission Electron Microscopy'. The images represent underlying data set of completed and published experiments. Still, the publisher could not facilitate hosting medium-sized files in other cases publishers did not facilitate publication of data sets in usable formats. This data set consisted of several experiments with each several phases of the data - from raw to processed - and visualizations. All these needed to be linked to the publication. The main investment for this set was extracting descriptive information from the publication and supplementary material. The main challenge was modeling the data set to ease linking and citation. This is also the first data set that received a digital object identifier (DOI), rendering it a unique citable unit.

In addition to these versatile cases we also encountered some other challenges, such as very confidential data sets amongst others; at the moment possibilities for restricted access are being developed, but long-term preservation of very confidential data and access will (at the moment) not be facilitated by the 'Data center'. Once the strategy goals and business model of the data center elaborated, preserving such confidential data sets might be (re)considered as a (paid) service in the future. Similarly to the acquisition for publications there was occasionally resistance to investment of time. An interesting observation is that we experienced acquisition for a data repository to be easier than for a publication repository.
Also data sets difficult to reuse, mostly caused by insufficient descriptive information were submitted, so far we have assisted in improving the documentation as much as possible. Barriers for reuse can also be caused by a very high complexity of data sets or slightly different, by a very low reuse value. For example because the data generation process is very easy to repeat (like simulations),investment in long term data preservation does not make sense in these cases. In the near future we will have to look into preservation of models and code instead (preserve VirtualBox Disk Images, vdi's) for meeting the needs of the last mentioned example.

Conclusions
During the first two years of the project 3TU.Datacentrum we encountered several interesting cases and high-value data of leading research groups. From over 3.000 data sets ingested so far the vast majority are repetitive sets within two longitudinal collections. Looking at the remaining cases we see that for about half of the data sets offered significant investment in modeling, description and ingest are not justified. These are mostly relatively small, one-time datasets. These sets are stored in BagIt format containing the original formats with checksums and a file list. For a similar number of cases quite a bit of labor has gone into modeling the data, definition and collection of metadata, and subsequent ingest actions. Every time the 'costs' were carefully balanced against 'profit', especially regarding reuse potential. We expect this distribution to remain the stable for the next few years with a slow decrease in cases for which significant investment is required.
We strongly suspect that we covered just a small tip of the iceberg.

Following the 'iceberg-rule' that only a small part (one third) is visible and a large part hidden, it seems too soon to tell what the most important data producing communities, standards and file formats will be for natural science and engineering in the Netherlands.

Collaboration among the three university libraries was the princile of starting the data curation studies providing strong funding base for data management by 3TU.Federation. This also enabled

better insight into the research at different universities, different organization models and setup of libraries and their repositories. In the course of the project, other collaborations were set up to combine the experience and knowledge, as well as capacity (Figure 1).

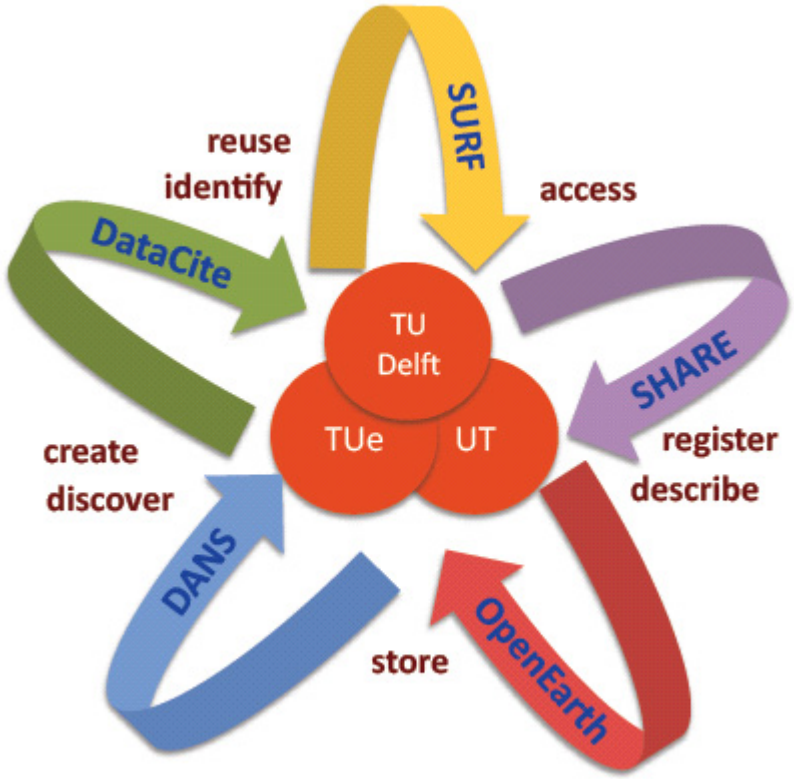

Figure 1: Collaborations of the project '3TU.Datacentre' for building a data repository and their services: Create, Store, Describe, Identify, Register, Discover, Access and (Re)use. The three universities UT, University of Twente; TU/e, Technical University of Eindhoven; TU Delft, Delft University of Technology operate together under the 3TU.Federation umbrella.

Collaboration between data centers and data users is the key solution to building a data curation facility as they provide benefits for both the data producers and consumers as well as the libraries and institutions building and maintaining the data centers. Collaboration with OpenEarth provided us with knowledge on process and tools used by a community of data producers. Also acquisition efforts are shared with the community, and we recently started developing plans for integration of data management into the data creation process of several lab facilities. Collaboration with other research partners, such as SHARE (Sharing Hosted Autonomous Research Environments) enabled data producers to share tools without distributing the tools.

We also conclude that no institution can do this alone on a national scale, collaboration will be needed on digital curation.
DataCite, a recently founded consortium, enabled assignment of digital object identifiers to datasets. This pioneering development contributed significantly to demonstrating the secure preservation of the digital treasure of the researchers. It is these persistent identifiers (DOIs) that function as the convincing incentive of the data center.
Joining forces with DANS on research data curation hopefully leads to a single (virtual) back office organization and distributed front offices for handling all research data independent of discipline.

Last but not least we make a plea to act now! Through collaboration we were able to apply the guideline ‘Do not redo work already done elsewhere with regard to data curation’. Too much is still uncharted territory to waste resources on competing. For institutional data repositories this means a careful selection of gaps to focus on and setting up close relations with relevant ‘colleagues’ on aspects important for the institutes’ community. Although important, just creating an overview of the important data, needs, available infrastructure, possible partners etc. itself will require a considerable investment of time and resources while existing communities already need support.

References

Angevaare, Inge. (2009, September). A future for our digital memory: permanent access to information in the Netherlands, interim report – summary in English. NCDD – Netherlands Coalition for Digital Preservation. Retrieved February, 2010, from http://www.ncdd.nl/en/publicaties.php

Burton, Adrian, & Treloar, Andrew. (2009, December). Designing for Discovery and Re-Use: the 'ANDS Data Sharing Verbs' Approach to Service Decomposition. 5th International Digital Curation Conference.

Digital curation (2010). Retrieved March, 2010, from http://en.wikipedia.org/wiki/Digital_curation

Key Perspectives Ltd (2010, January).
Data dimensions: disciplinary differences in research data sharing, reuse and long term viability. SCARP Synthesis Study. (ISSN 1759-586X).
Digital Curation Centre, University of Edinburgh.

Nature. Data's shameful neglect. Vol. 461, issue 7261, p 145.

Tjalsma, Heiko. (2010, April). Forthcoming report by DANS and 3TU Data Centre on Selection of Research Data.