Data seal of approval - assessment and review of the quality of operations for research data repositories

Dr. Henk Harmsen

Data Archiving and Networked Services
The Hague, The Netherlands
henk.harmsen@dans.knaw.nl

Introduction

Data Archiving & Networked Services (DANS) is active in the area of data infrastructure, with two main themes, namely (digital) archiving and making research data available. The field of activity of DANS covers both the social sciences and the humanities. DANS also manages its own data repository of research data.
In 2005, the founders of DANS, the Royal Netherlands Academy of Arts and Sciences (KNAW) and the Netherlands Organization for Scientific Research (NWO), gave DANS the formulation of a data seal of approval as one of its assignments. In February 2008, 17 guidelines were presented under the name Data Seal of Approval, nationally at a KNAW symposium and internationally at the first African Digital Curation Conference. This article will explain more about the backgrounds of the seal of approval: what it is and what it isn't, which international seals of approval exist, how this seal of approval matches them, what its unique selling point is, and what the plans for the future are?

What it is and what it isn't?

The data seal of approval consists of 17 guidelines that may be helpful to an archiving institution striving to become a trusted digital repository (TDR[1]). The guidelines have been formulated in such a way that they are easily understandable and leave sufficient room for a broad interpretation. Standardization was not the objective as the point of departure was that the data seal of approval would remain dynamic during its first years. The seal of approval does not express any views regarding the quality of the data to be archived, but does regarding the provisions an archive has made to guarantee the safety and future usability of the data.
The seal of approval mentions 4 stakeholders: the financial sponsor, the data producer, the data consumer and the data repository, which share an interest and are responsible for a properly functioning data infrastructure. The sponsor is advised to use the guidelines as a condition for financing of research projects. The remaining three stakeholders are addressed in the 17 guidelines. For example, the data producer is expected (three guidelines) to place its data in a TDR and to provide the research data as well as the metadata in the format requested by the data repository. The data consumer must, if it has access to or uses the information in a TDR, respect (inter)national legislation, (scientific) codes of behavior and the applicable licenses (three guidelines). The data repository, in its turn, must ensure that the archive is equipped in such a way that data producer and data consumer are able to meet their obligations. In addition, there are eleven more guidelines for the data repository, regarding organization (mission, dealing with legal regulations, quality management, long-term planning and scenarios), processes (transfer responsibility, data references, integrity and authenticity) and technical infrastructure (OAIS and automated processes).
In other words, the data repository is the stakeholder of which most is expected. Therefore, an assessment document has been formulated for the data repository which, when completed, approved and publicly published, will result in the repository being allowed to use the logo of the data seal of approval. The logo makes the repository recognizable to both data produces and data consumer.
A data repository may be able to delegate some of the guidelines to another archive that bears the logo of the data seal of approval. This way, the concerned repository does not need to execute all the guidelines in order to meet the requirements of the seal of approval.
With regard to auditing the repositories, a minimal system was chosen that is based on trust. The repository publishes its own assessment and then applies for an audit. This audit is carried out by a member of the international *DSA* (data seal of approval) assessment group[2] on the basis of the available assessment document. It determines whether the guidelines have been complied with and whether the logo can be awarded.

[1] The term Trusted Digital Repository (TDR) occurs in almost all seals of approval. However, it is unclear what a TDR is exactly. At the time of writing, Wikipedia does not yet have a description of the concept. Main point of such a repository is ‘trust’. It is the basis of the data seal of approval.

[2] The international DSA assessment group will be launched in the fall of 2008

International initiatives

The text accompanying the seventeen guidelines states[3] that these 'are in accordance with, and match national and international guidelines regarding digital data archiving'.
In this section, I will explore the mentioned initiatives in slightly more detail.

Catalogue of Criteria for Trusted Digital Repositories[4] - NESSTOR

This catalogue has identified criteria that can help in the evaluation of the reliability of digital archives at both the organizational and the technical level. The criteria were defined in close cooperation with a broad range of data institutions and information producers. One of the objectives is to offer a tool enabling archiving institutions to archive and demonstrate reliability. The catalogue is also an opportunity for arriving at the certification of repositories, with a 'standardized national or international process'. Again, 'reliability' or 'trust' plays a role here. The catalogue can be used for conceiving, working out and eventually implementing a 'trusted digital long-term repository' and for working out (in various stages) of a self-assessment.
The criteria catalog employs over fifty criteria organized into fourteen sections that are arranged into three areas of attention namely: Organizational framework, Object management, and Infrastructure and Security.

Digital Repository Audit Method Based on Risk Assessment (DRAMBORA)[5] of the Digital Curation Centre (DCC) and DigitalPreservationEurope (DPE)

The DRAMBORA toolkit is available to support internal audits of archiving institutions. To this end, the party responsible for the archive has the challenge of to tracking down the weaknesses, while at the same time acknowledging the strength of the archive.
DRAMBORA helps track down the many risks any archiving institution runs. This takes place in the form of process description:

- A detailed description of the organization (mission, and activities);
- Formulation of possible risks, organizational as well as technical, that may occur;
- Evaluation of the impact of these risks and making them manageable and controllable.

DRAMBORA gives support by means of templates for the description of risks and codes to assess the severity of the risks. Apart from that, it is an open process which must be shaped by the party responsible for the repository. There is, however, a list of examples of possible risks.

The philosophy of the DRAMBORA authors is clear: by monitoring closely what people are doing and how they are doing it, a repository is capable of keeping the risks involved in archiving of data under control.

Further, the Research Library Group (RLG)[6] developed the *Trustworthy Repositories Audit & Certification (TRAC): Criteria and Checklist.*
This criteria checklist comprises three sections, arranged into a various aspects, in their turn subdivided into more than eighty criteria.
The paper *Foundations of Modern Language Resource Archives* of the Max Planck institution in Nijmegen[7] must not remain unmentioned. The document describes a data seal of approval specifically for language bodies. A language resource archive (LRA) must meet nine principles.
The Research Information Network in the UK[8] developed the *Stewardship of digital research data: a framework of principles and guidelines*. This document is built up of 5 principles, spread across 40 guidelines.
The German Initiative for Network Information (DINI) developed the *Certificate Document and Publication Services of the Deutsche Initiative für Netzwerkinformation*[9], a certificate mainly intended for institutional repositories with their own Document and Publication Services.

Synthesis

The guidelines of the data seal of approval can be seen as a basic set of the above proposals. The data seal of approval wants to facilitate 'awareness' at the archiving institutions. It can serve as a first step toward a 'heavier' assessment and certification. The authors see the data seal of approval as supporting for example TRACK and DRAMBORA. The objective of the data seal of approval was mainly to try and convince archiving institutions to start paying attention to quality management.

Unique selling point

The data seal of approval (DSA) as developed by DANS has a number of unique features: The DSA is oriented toward scientific data, not primarily toward publications. The DSA not only pays attention to the archiving institution, but also to the data producer and the data

[3] Data Seal of Approval, chapter 0.3 Guidelines. See: <http://www.datasealofapproval.org>
[4] See: <http://edoc.hu-berlin.de/docviews/abstract.php?id=27249> [site visited 15 August 2008].
[5] See: <http://www.digitalpreservationeurope.eu/announcements/drambora/> [site visited 15 August 12-2008].
[6] See: <http://www.crl.edu/content.asp?l1=13&l2=58&l3=162&l4=91> [site visited 15 August 2008].
[7] Peter Wittenburg, Daan Broeder, Wolfgang Klein, Stephen Levinson, of the Max-Planck-Institute for Psycholinguistics in Nijmegen, The Netherlands, and Laurent Romary of the Max Planck Digital Library in Munich, Germany. See: <http://www.lat-mpi.eu/papers/papers-2006/general-archive-paper-v4.pdf> [site visited 15 August 12-2008].
[8] See: <http://www.rin.ac.uk> [site visited 15 August 12-2008].
[9] See: <http://edoc.hu-berlin.de/series/dini-schriften/2006-3-en/PDF/3-en.pdf> [site visited 15 August 12-2008].

consumer. This encourages the idea of shared responsibility.
As indicated before, the DSA is not in conflict with for example TRAC, but is rather a step toward it. Where TRAC chooses standardization, the DSA opts for 'trust'. This way of working does on the other hand match the custom of peer review in the scientific world.
The DSA also focuses on smaller organizations. The DSA is relatively light and therefore easy to implement. Openness, dynamics and speed are possible in the actual implementation.
The DSA is formulated as points of attention, not as solutions. Finally, the DSA offers possibilities for subcontracting archiving and still meet the requirements of the DSA. This will be appreciated by research groups with their own data projects.

Future

In 2009, DANS will comply with the data seal of approval and its policy is aimed toward being on the way to meeting the TRAC criteria. Furthermore, DANS uses the code for information security[10].
DANS strives toward internationalization of the data seal of approval. The previously mentioned DSA assessment group will be launched in the fall of 2008, and that same year, four pilots will be planned in The Netherlands as a first step in the area of certification of the DSA.

[10] CVI - The *Code voor Informatiebeveiliging* is the Dutch version of the British Standards 7799, which was later published as ISO/IEC 17799 as international standard for information security in organizations. It is a general code applicable to all institutions that work with information.

Updating DAITSS - Transitioning to a web service architecture

Randall Fischer, Carol Chou, Franco Lazzarino

Florida Center for Library Automation
5830 NW 39th Avenue
Gainesville, FL 32605, USA
rf@ufl.edu, cchou@ufl.edu, flaz@ufl.edu

Abstract

The Florida Digital Archive (FDA) is a long-term preservation repository for the use of the libraries of the public universities of Florida. The FDA uses locally-developed software called DAITSS, which was designed to perform the major functions of Ingest, Archival Storage, Data Management and Dissemination in the OAIS reference model. A DAITSS 2 project is in process to re-write the application based on a distributed, Web services model. This paper describes the major changes in store for DAITSS 2.0, the rationale behind them, and the issues involved in their design and implementation. These changes include: moving from a monolithic to distributed processing environment; implementation of modular RESTful services; incorporation of existing tools, services, and registries; and revising the internal data model to be more conformant with the PREMIS data.
.

Introduction

The Florida Digital Archive (FDA) is a long-term preservation repository for the use of the libraries of the public universities of Florida. It has been in operation since late 2005, and as of July 1, 2008 has archived 52,000 information packages comprising 3.6 million files (10.4TB). Nine universities have agreements with the FDA to archive their submissions, which are being ingested at an average rate of 30-60 GB per day.

The FDA uses locally-developed software called DAITSS, which was designed to perform the major functions of Ingest, Archival Storage, Data Management and Dissemination in the OAIS reference model. DAITSS implements format-specific preservation strategies including normalization, migration and localization. ([Caplan 2007])

DAITSS was a pioneering digital preservation system. When it was designed and developed, there were few models of true preservation repositories and few external tools available for performing specific functions such as format validation and metadata extraction. It is somewhat remarkable that in three years of FDA operations, no major functional flaws have been discovered and few enhancements to functionality are pressing. The architecture of the application, however, requires major redesign. DAITSS was coded as a monolithic, self-contained system. A DAITSS 2 project is in process to re-write the entire system based on a distributed, Web services model.

The fundamental principles governing the original design of DAITSS have not changed. These include:

- strict conformance to the OAIS functional model;
- a requirement that the archived data store be self-defining, so that if the DAITSS system were lost, all known information about archived objects could be recovered from the data store itself;
- data once written to archival storage cannot be altered; modified objects are in effect new objects;
- original versions of archived files must be retained unaltered.

In conformance with these principles, files are modified only during the Ingest process as the SIP is transformed into the AIP. DAITSS relies upon format normalization and migration as preservation strategies, and these are implemented as part of Ingest. All files in the SIP as originally submitted are retained unaltered in perpetuity, but other versions may be derived and added to the AIP.

The basic unit of storage and processing is an Information Package. Each Information Package consists of an XML descriptor and all of the content files required to assemble one (and possibly more) representations of an information object. The Information Package is the only unit of input and output; that is, even if only a single file in an AIP is needed, the entire IP must be disseminated.

Because many years may pass between the time a file is ingested and when it requires some preservation treatment, dissemination requests are filled by a three-step process. In the first step, the AIP is exported from the repository and placed in the Ingest queue as a SIP. In the second step, the AIP-cum-SIP is re-ingested, and undergoes file identification, validation, and transformation processing according to the current version of the software. In the final step, the resulting AIP is reformatted into a DIP and delivered to the requestor.

This model will be retained in DAITSS 2. It has worked well in practice and in fact has beneficial side-effects. For example, the ingest model makes updates extremely simple, and the dissemination model allows the FDA to implement migration on request or mass migration depending on the circumstances.

Another governing principle was to use standard formats and metadata schemes whenever possible. However, at the time DAITSS was initially developed, there were few