Trusted Repository Certification and its Potential to Improve Data Quality

Peter Doorn
Data Archiving and Networked Services (DANS)

8th International Digital Curation Conference Workshop "Data publishing, peer review and repository accreditation: everyone a winner?"
17 January 2013, Amsterdam
Contents

• Relevance of trusted data repositories for journals
• About trust and certification
  – how to build trust
  – pros and cons of certification
• European framework for certification
  – DSA, DIN, ISO
• Ongoing work: the Research Data Alliance Working Group on Certification
• Back to the journals that link data to articles
  – Data Availability Policies
  – Data reviews
Trust

- Trust is at the very heart of storing and sharing data
  - Depositors
  - Repositories
  - Users
  - Funders
- Trust by definition implies uncertainty
  - “You don't have to see the whole staircase, just take the first step.” (Martin Luther King)
- Certification as a means of building trust
What do you rely on?

You can trust us

Can you?

Driven by data
What is a trusted digital repository?

Things are not always what they say they are. Things do not always state what they are.
WHY certification?

- Growing importance of sharing data
  - transparency of research
  - data re-use
- Long term archiving is a condition for sharing
- Trust is at the very heart of storing and sharing data:
  - Funders
  - Depositors
- Certification guarantees trustworthy digital repositories
### Do we need certification?

<table>
<thead>
<tr>
<th>No (devil’s advocate)</th>
<th>Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Trustworthiness of digital repositories is an illusion</strong></td>
<td><strong>Trustworthiness of digital repositories is not an illusion</strong></td>
</tr>
<tr>
<td>- Too complicated to measure</td>
<td>- Three levels are clear and balanced</td>
</tr>
<tr>
<td>- Impossible to maintain over time</td>
<td>- Different levels for different needs</td>
</tr>
<tr>
<td>- Too informal, too expensive and too time consuming</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Objective and consistent auditing is an illusion</strong></th>
<th><strong>Objective and consistent auditing can be done</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>- “If auditing becomes a career, what will happen to objectivity?” (Helen Tibbo)</td>
<td>- Auditing is a career in many other areas</td>
</tr>
<tr>
<td>- Impossible to guarantee consistency around the globe</td>
<td>- Some variation according to local requirements is permitted</td>
</tr>
</tbody>
</table>
HOW to get there?

Awareness raising
• funders, repositories, research communities

Data management and data archiving
• as a condition for funding research proposals

Financial support
• DMPs (including data archiving)
• Certification activities of repositories (including training the auditors)

Driven by data
Certification of digital repositories

- European (→ worldwide) framework
- 3 standards
- 3 levels (basic, extended, formal)

Driven by data
Certification Standards: 
Data Seal of Approval (DSA)

- DANS initiative
- International Board
- 16 guidelines
- Self assessment
- Transparency
- 8 seals rewarded
Certification Standards:
DIN 31644

- Kriterienkatalog vertrauenswürdige digitale Langzeitarchive – NESTOR, Deutsche National Bibliothek
- 34 criteria
- Test audits 2013
Certification Standards:
ISO 16363

- Based on Open Archival Information System (OAIS) and Trusted Repository Audit and Certification (TRAC)
- Over 100 metrics
- Test audits 2011
- Full external auditing process
European Certification Framework

- Three-tiered framework
- Time to implement certification as a new common service!

Standards will be available free from http://www.ccsds.org

There is a hierarchy of ISO standards concerned with good auditing. ISO 16919 is positioned within this hierarchy in order to ensure that these good practices can be applied to the evaluation of the trustworthiness of digital repositories using ISO 16363. It covers principles needed to inspire confidence that third party certification of the management of the digital repository has been performed with impartiality, competence, responsibility, openness, confidentiality, and responsiveness to complaints.

Metrics concerning:
- Organizational Infrastructure
  - e.g. The repository shall have a documented history of the changes to its operations, procedures, software, and hardware.
- Digital Object Management
  - e.g. The repository shall have access to necessary tools and resources to provide authoritative Representation Information for all of the digital objects it contains.
- Infrastructure and Security Risk Management
  - e.g. The repository shall have procedures in place to evaluate when changes are needed to current software.

Basic Certification
Data Seal of Approval
Extended Certification
Audit by external auditors

Standards based Repository Audit and Certification (ISO 16363)

EUROPEAN FRAMEWORK FOR AUDIT AND CERTIFICATION OF DIGITAL REPOSITORIES
to be promoted by the EU

Driven by data
Framework levels

• **Basic Certification** is granted to repositories which obtain DSA certification

• **Extended Certification** is granted to Basic Certification repositories which *in addition* perform a structured, externally reviewed and publicly available self-audit based on ISO 16363 or DIN 31644

• **Formal Certification** is granted to repositories which *in addition to* Basic Certification obtain full external audit and certification based on ISO 16363 or DIN 31644
Allies in implementation

- European Commission
  - recommendation for Horizon 2020

- Research Funders like NWO
  - data access policy

- Research Data Alliance Council
  - working group on certification
On-going work

• Work Package on “Trust” within APARSEN project (special sessions on Ipres and APA conferences)
• Certification Working Group of the Research Data Alliance
• Parties in European Framework for Audit and Certification of Digital Repositories
• Research Infrastructures such as CESSDA, CLARIN, DARIAH started implementing trust requirements
Research Data Alliance Working Group on Certification – case statement

• **Report** on the situation with respect to certification on various continents (Europe, N +S America, Asia, Australia, Africa)

• **Recommendations and Actions** to work towards the **global implementation** of certification

http://rd-alliance.org/
Questions raised for RDA-WG Certification

• Should not repositories develop Service Level Agreements in stead of working towards certification?
• What will you actually certify? The repository or the data? Is not data quality more important than the quality of the repository?
• Who are the stakeholders? What is the role of the researchers?
• How manage a worldwide certification effort?
• Which concrete actions can be taken? Can you go further than a report on the state of affairs and strategic recommendations? How to implement the recommendations?
Data Availability Policies (DAP) by Journals

• Few Journals have a DAP, number is growing
• Variations across disciplines
• DAPs differ in wording and approach:
  – Journal/publisher archives the data
  – Journal outsources archiving to a DataVerse, special/national repository
• Who is responsible: editors, publishers or scholarly societies/communities?
• Whether the data archive is trustworthy is another matter
• Archiving does not guarantee data quality, only that the data can be checked
Two stages of checking

• When a paper is submitted for peer-reviewed publication, data access for reviewers is required
• When the paper is published, readers should be able to access the data
  – Open access and open data are preferred, but not required
  – Data Archives can implement a data review mechanism by their users
Data Reviews at DANS

- Marjan van Egmond and Jeff van Egmond reported on it at IDCC Conference 2011

<table>
<thead>
<tr>
<th>Aspect</th>
<th>Rating</th>
<th>Average rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>data quality</td>
<td>4 2 0 0 0</td>
<td>★★★★☆ (4.67/5)</td>
</tr>
<tr>
<td>quality of the documentation</td>
<td>4 1 0 0 1</td>
<td>★★★★☆ (4.17/5)</td>
</tr>
<tr>
<td>completeness of the data</td>
<td>4 2 0 0 0</td>
<td>★★★★★ (4.67/5)</td>
</tr>
<tr>
<td>consistency of the dataset (if applicable)</td>
<td>4 1 0 0 0 1</td>
<td>★★★★★ (4.8/5)</td>
</tr>
<tr>
<td>structure of the dataset (if applicable)</td>
<td>4 2 0 0 0</td>
<td>★★★★★ (4.67/5)</td>
</tr>
<tr>
<td>usefulness of the file formats</td>
<td>3 2 0 1 0</td>
<td>★★★★☆ (4.17/5)</td>
</tr>
</tbody>
</table>

5 out of 6 reviewers of this dataset recommend the use of it.
1 out of 6 reviewers of this dataset has published using this dataset.
2 out of 6 reviewers of this dataset intend to use this dataset for a publication.
Thank you for your attention

peter.doorn@dans.knaw.nl
www.dans.knaw.nl
www.trusteddigitalrepository.eu
What is DANS?

Mission: promote and provide permanent access to digital research information

Institute of Dutch Academy and Research Funding Organisation (KNAW & NWO) since 2005

First predecessor dates back to 1964 (Steinmetz Foundation), Historical Data Archive 1989

Driven by data
DANS main activities and services

- Encourage researchers to self-archive and reuse data by means of our Electronic Archiving SYstem EASY
- Our largest digital collections are in archaeology, social sciences and history (moving into other domains)
- Provide access, through Narcis.nl, to thousands of scientific datasets, e-publications and other research information in the Netherlands
- Data projects in collaboration with research communities and partner organisations
- Advice, training and support (Data Seal of Approval, Persistent Identifier Infrastructure)
- R&D into archiving of and access to digital information
Thank you for your attention
and visit us at:

www.dans.knaw.nl
www.narcis.nl

peter.doorn@dans.knaw.nl

DANS is an institute of KNAW en NWO