Data Archiving and Networked Services
DANS

• Certification of Trusted Digital Repositories and FAIR data

• Heiko Tjalsma, Senior Policy Advisor DANS

The road to FAIR data?
SND Forum, Gothenburg 14 November 2017

FAIR data and certification

Data must be

FAIR - Findable, Accessible, Interoperable, Reusable

And must remain so, and therefore should be preserved in a

CoreTrustSeal Certified Trusted Digital Repository - TDR

FAIR and certification

- FAIR data
- Does a TDR enable FAIR?
- TDR: How do we know? → certification
  - Example: Certification in CESSDA:
- Hurdles for certification
- The relation FAIR – TDR

FAIR Data

PRINCIPLES

A minimal set of community agreed guiding principles to make data more easily discoverable, accessible, appropriately integrated and re-usable, and adequately citable.

In the FAIR Data approach, data should be:

Findable – Easy to find by both humans and computer systems and based on mandatory description of the metadata that allow the discovery of interesting datasets;

Accessible – Stored for long term such that they can be easily accessed and/or downloaded with well-defined license and access conditions (Open Access *when possible*), whether at the level of metadata, or at the level of the actual data content;

Interoperable – Ready to be combined with other datasets by humans as well as computer systems;

Reusable – Ready to be used for future research and to be processed further using computational methods.

Does a TDR enable FAIR?

This is on certifying Trustworthy Digital Repositories:

“Perhaps the biggest challenge in sharing data is trust: how do you create a system robust enough for scientists to trust that, if they share, their data won’t be lost, garbled, stolen or misused?”

(from the Data Harvest report – RDA Europe 2014)

What is a Trusted Digital Repository?

A repository….

- With a mission to provide reliable, long-term access to digital resources, now and in the future
- Understanding threats to and risks to the data within its systems
- Having a regular cycle of audit and/or certification

European framework of certification levels

Basic Certification is granted to repositories which obtain CoreTrustSeal certification

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

Formal Certification is granted to repositories which *in addition t*o Basic Certification obtain full external audit and certification based on ISO 16363

DIN 31644: extended certification

34 criteria written by German NESTOR group and adopted in Germany as DIN 31644

Self-assessment procedure by NESTOR leads to NESTOR seal

Review of the assessment by 2 reviewers, appointed by NESTOR

Self assessment and evidence on website

http://www.langzeitarchivierung.de/Subsites/nestor/EN/nesSiegel/siegel_node.htm l

DIN

ISO 16363: formal certification

Based on Open Archival Information System (OAIS) and Trusted Repository Audit and Certification (TRAC)

Over 100 metrics

Test audits 2011 by PTAB (Primary Trustworthy Digital Repository Authorisation Body)

Full external auditing process

ISO 16919: Requirements for bodies providing audit and certification of candidate trustworthy digital repositories

No ISO certifications yet..

http://www.iso16363.org/

DRAMBORA

Digital Repository Audit Method Based on Risk Assessment

A Toolkit based on Risk Assessment

It assesses how a repository achieves its goals.

The audit produces:

An organisational profile

Clearly identified and documented repository assets, roles and activities

A catalogue of risks and the likelihood that these will occur.

Data Seal of Approval

- Basic, lightweight certification mechanism
- 16 guidelines for Trustworthy Digital Repositories
- Guidelines that relate to Data Producers, Data Repositories, and Data Consumers
- Self-assessment, with no site visit
- Peer review process supervised by DSA Board
- DSA granted for a period of two years
- Online tool for self-assessment and review

ICSU/WDS accreditation

- Basic certification mechanism
- Catalogue of 17 criteria
- The certification criteria apply for regular and network membership of WDS
- Self assessment, reviewed by the WDS Scientific Committee, with possibility of site visit
- Review of accreditation every 3-5 years

RDA Working Group

- DSA and WDS both lightweight mechanisms for repository assessment
- DSA began in social science and humanities, WDS in natural and physical sciences but both expanding in scope
- Developed common catalogue of criteria for basic repository assessment
- Developed common procedures for assessment
- Implementing a shared testbed for assessment

Changes: what is new?

- CoreTrustSeal is valid from September 2016
- Replacing existing version of the DSA and WDS Guidelines,

- **NEW:**
- Two reviewers
- **NOW:**
- Suspension of applications for CoreTrustSeal certification in November and December 2017

- From January 2018 onwards, repositories will be able to apply for certification via the CoreTrustSeal website: https://www.coretrustseal.org/apply/
- But: this involves a fee!

Outline

0. Contextual information

1. Organizational Infrastructure: Requirements 1-6

2. Digital Object Management: Requirements 7-14

3. Technology: Requirements 15-16

0. Context

– ***Repository Type***

– ***Brief Description of the Repository's Designated Community***

– ***Level of Curation Performed***

– ***Outsource Partners*****. If applicable, please list them**

Compliance levels

- ✓ 0 – Not applicable
- ✓ 1 – The repository has not considered this yet
- ✓ 2 – The repository has a theoretical concept
- ✓ 3 – The repository is in the implementation phase
- ✓ 4 – The guideline has been fully implemented in the repository

Organisational Infrastructure

I. Mission / Scope

R1. The repository has an explicit mission to provide access to and preserve data in its domain.

II. Licenses

R2. The repository maintains all applicable licenses covering data access and use and monitors compliance.

III. Continuity of access

R3. The repository has a continuity plan to ensure ongoing access to and preservation of its holding

Organisational Infrastructure

IV. Confidentiality / Ethics

R4. The repository ensures, to the extent possible, that data are created, curated, accessed, and used in compliance with disciplinary and ethical norms.

V. Organizational infrastructure

R5. The repository has adequate funding and sufficient numbers of qualified staff managed through a clear system of governance to effectively carry out the mission.

VI. Expert guidance

R6. The repository adopts mechanism(s) to secure ongoing expert guidance and feedback (either in-house, or external, including scientific guidance, if relevant)

Digital Object Management

VII. Data integrity and authenticity

R7. The repository guarantees the integrity and authenticity of the data.

VIII. Appraisal

R8. The repository accepts data and metadata based on defined criteria to ensure relevance and understandability for data users.

IX. Documented storage procedures

R9. The repository applies documented processes and procedures in managing archival storage of the data.

Digital Object Management

X. Preservation plan

R10. The repository assumes responsibility for longterm preservation and manages this function in a planned and documented way.

XI. Data quality

R11. The repository has appropriate expertise to address technical data and metadata quality and ensures that sufficient information is available for end users to make qualityrelated evaluations.

XII. Workflows

R12. Archiving takes place according to defined workflows from ingest to dissemination.

Digital Object Management

XIII. Data discovery and identification

R13. The repository enables users to discover the data and refer to them in a persistent way through proper citation.

XIV. Data reuse

R14. The repository enables reuse of the data over time, ensuring that appropriate metadata are available to support the understanding and use of the data.

Technology

XV. Technical infrastructure

R15. The repository functions on well-supported operating systems and other core infrastructural software and is using hardware and software technologies appropriate to the services it provides to its Designated Community.

XVI. Security

R16. The technical infrastructure of the repository provides for protection of the facility and its data, products, services, and users.

Additional

XVII. Additional information

R17. Any other relevant information you wish to provide on your repository.

XVIII. Applicant feedback

AND:

Extended Guidance (for reviewers and applicants!)

General Extended Guidance

A. General points

B. Missing evidence

C. Understandability of documentation

D. Documentation in any other language than English

E. Confidentiality of internal documents

CESSDA Work Plan 2014 - 2015

Approved unanimously by the General Assembly in June 2014:

(Priority 1)

Coordinating Trusted Digital Repository status (DSA) for all Service Providers

CESSDA **Trust Working Group**

Support and Training on DSA / CTS Compliance:

DSA Self Assessment and Review Process leading to a **Gap Analysis**

Steps:

- Self Assessment by each SP
- Each Self Assessment reviewed by another SP
- All Self Assessments reviewed by four “experts”

First time in 2013 and repeated in 2016/2017

DSA Compliance in reach? Hurdles?

Three categories Service Provider's:

1. Already certified
2. Close to certification
3. Starting from blank

Not all SP's willl/might reach compliance in the short run:

Often starters: uncertain future / restructuring of organisations as main reasons / no sustainable funding

Hurdles for all: **evidence! In particular on technical infrastructure**

Implementing the FAIR Principles?

Box 2 | The FAIR Guiding Principles

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

15 Criteria

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

See: http://datafairport.org/fair-principles-living-document-menu and https://www.force11.org/group/fairgroup/fairprinciples

Some unresolved FAIR complications:

1. Dependencies among dimensions, difficulty to measure the criteria, no rank order from “low” to “high” FAIRness, grouping of criteria under dimensions is disputable
2. Do we need or want additional dimensions, principles or criteria?
    - Is “openness” a separate dimension, not included in FAIR?
    - Is it desirable/possible to say something about “substantive” data quality, such as the accuracy/precision or correctness of the data?
    - What about the long-term access? For how long does data remain FAIR?
    - Should data security be included?
3. Do we need separate FAIR criteria for different disciplines?
    - e.g. machine actionable data are more important in some fields than in other; note that data accessibility by machines is partly defined by technical specs (A1), partly by licenses (R1.1)

Thank you!

heiko.tjalsma@dans.knaw.nl