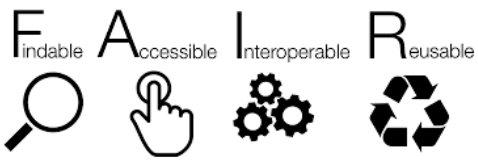
Data Archiving and Networked Services
DANS

@pkdoorn @dansknaw

FAIR Data in Trustworthy Repositories:
Everybody wants to play FAIR, but how do we put the principles into practice?



Peter Doorn, Director DANS
Ingrid Dillo, Deputy Director DANS

IASSIST 2017
May 23–26
Lawrence, Kansas
Data in the Middle:
The common language of research

Session B2: Repository Strategies across Communities
Wednesday, 24 May 2017, 10:30 – 12:00, Room Big 12

DANS is about keeping data FAIR

Data Archiving and Networked Services
DANS

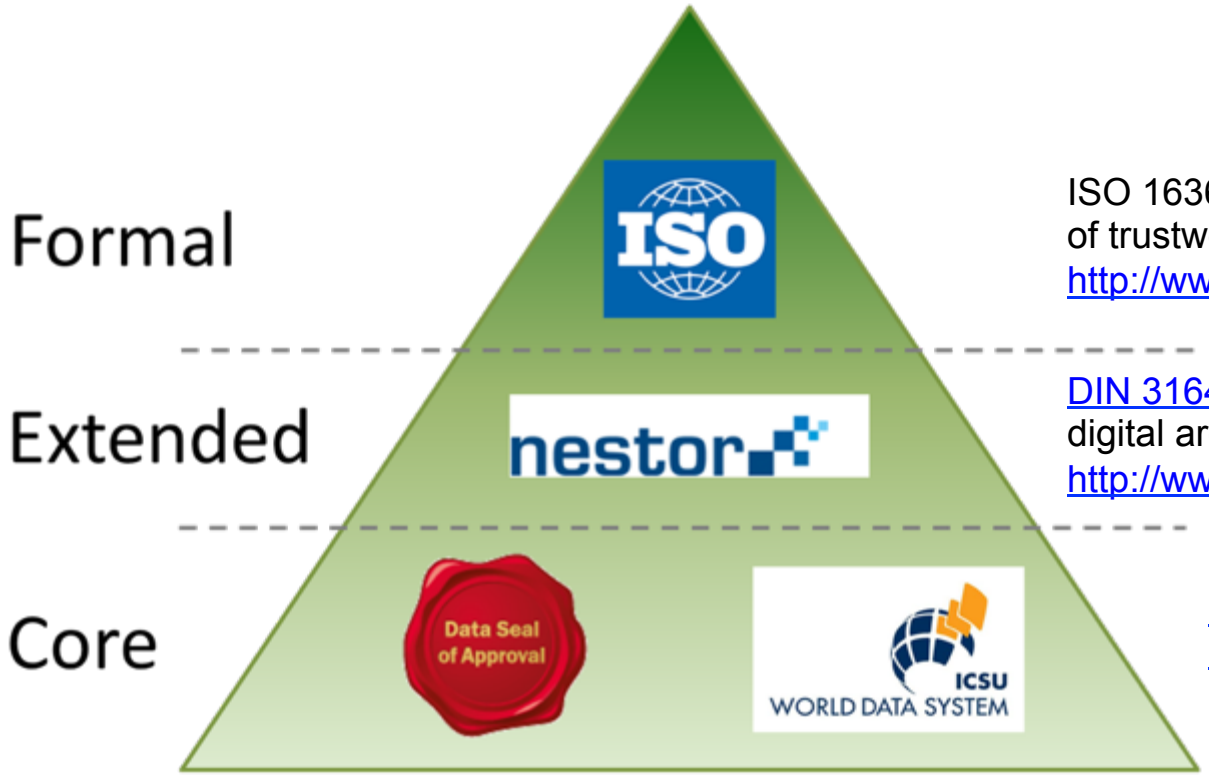
What will we present?

- Quality (trustworthiness) of data repositories
- Quality (fitness for use) of datasets

DANS and DSA

- 2005: DANS to promote and provide permanent access to digital research resources
- Formulate quality guidelines for digital repositories including DANS
- 2006: **5 basic principles** as basis for 16 DSA guidelines
- 2009: international DSA Board
- Almost 70 seals acquired around the globe, but with a focus on Europe

The Certification Pyramid



ISO 16363:2012 - Audit and certification of trustworthy digital repositories
http://www.iso16363.org/

DIN 31644 standard “Criteria for trustworthy digital archives”
http://www.langzeitarchivierung.de

http://www.datasealofapproval.org/
https://www.icsu-wds.org/

DSA and WDS: look-a-likes

Communalities:

- Lightweight, self assessment, community review

Complementarity:

- Geographical spread
- Disciplinary spread

Partnership

Goals:

- Realizing efficiencies
- Simplifying assessment options
- Stimulating more certifications
- Increasing impact on the community

Outcomes:

- Common catalogue of requirements for core repository assessment
- Common procedures for assessment
- Shared testbed for assessment

New common requirements: CoreTrustSeal

- Context (1)

- Organizational infrastructure (6)
- Digital object management (8)
- Technology (2)

- Additional information and applicant feedback (1)


25/08/2015 Common Requirements/V2.1

Data Seal of Approval

ICSU WORLD DATA SYSTEM

DSA–WDS Partnership Working Group
Catalogue of Common Requirements

Introduction

Importance of Certification

National and international funders are increasingly likely to mandate open data and data management policies that call for the long-term storage and accessibility of data.

If we want to be able to share data, we need to store them in a trustworthy digital repository. Data created and used by scientists should be managed, curated, and archived in such a way to preserve the initial investment in collecting them. Researchers must be certain that data held in archives remain useful and meaningful into the future. Funding authorities increasingly require continued access to data produced by the projects they fund, and have made this an important element in Data Management Plans. Indeed, some funders now stipulate that the data they fund must be deposited in a trustworthy repository.

Sustainability of repositories raises a number of challenging issues in different areas: organizational, technical, financial, legal, etc. Certification can be an important contribution to ensuring the reliability and durability of digital repositories and hence the potential for sharing data over a long period of time. By becoming certified, repositories can demonstrate to both their users and their funders that an independent authority has evaluated them and endorsed their trustworthiness.

Basic Certification and its Benefits

Nowadays certification standards are available at different levels, from a basic level to extended and formal levels. Even at the basic level, certification offers many benefits to a repository and its stakeholders.

Requirements (indirectly) dealing with data quality

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

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

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

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

Requirements (indirectly) dealing with data quality

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

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

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 quality-related evaluations.

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

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.

New requirements are out now!

https://www.datasealofapproval.org/en/news-and-events/news/2017/3/30/new-standards-and-certification-entity-horizon/

RDA endorsed recommendation and European recognition

ICT technical specifications

The rules on European standardisation allow the European Commission to identify information and communication technology (ICT) technical specifications - that are not national, European or international standards - to be eligible for referencing in public procurement. This allows public authorities to make use of the full range of specifications when buying IT hardware, software and services, allowing for more competition in the field and reducing the risk of lock-in to proprietary systems.

The Commission can identify ICT technical specifications for referencing in public procurement under Article 13 of Regulation 1025/2012 on European Standardisation.

Resemblance DSA – FAIR principles

DSA Principles (for data repositories)	FAIR Principles (for data sets)
data can be **found** on the internet	**F**indable
data are **accessible**	**A**ccessible
data are in a **usable format**	**I**nteroperable
data are **reliable**	**R**eusable
data can be **referred** to	(citable)

The resemblance is not perfect:

- usable format (DSA) is an aspect of interoperability (FAIR)
- FAIR explicitly addresses machine readability
- etc.

A certified TDR already offers a baseline data quality level

Implementing the FAIR Principles

To be Findable:

F1. (meta)data are assigned a globally unique and eternally persistent identifier.
F2. data are described with rich metadata.
F3. (meta)data are registered or indexed in a searchable resource.
F4. metadata specify the data identifier.

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.

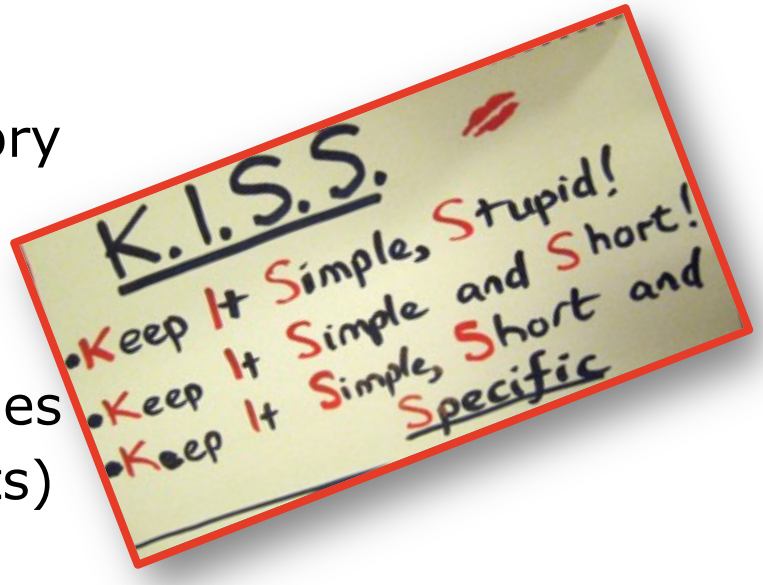
To be Re-usable:

R1. meta(data) have 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 their 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

Combine and operationalize: DSA & FAIR

- Growing demand for quality criteria for research datasets and ways to assess their fitness for use
- Combine the principles of core repository certification and FAIR
- Use the principles as quality criteria:
  - Core certification – digital repositories
  - FAIR principles – research data (sets)
- Operationalize the principles as an instrument to assess FAIRness of existing datasets in certified TDRs



Badges for assessing aspects of data quality and “openness”

These badges do not define good practice, they certify that a particular practice was followed.

ODI

BRONZE: data is openly licensed, available with no restrictions, accessible and legally reusable.

ODI

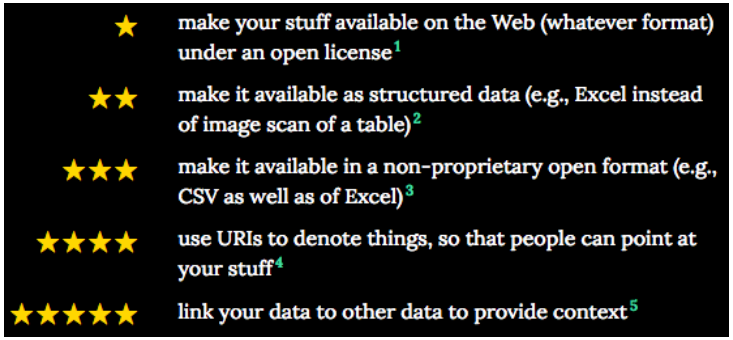
SILVER: satisfies the Bronze requirements, the data is documented in a machine readable format, reliable and offers ongoing support from the publisher via a dedicated communication channel.

ODI

GOLD: satisfies the Silver requirements, is published in an open standard machine readable format, has guaranteed regular updates, offers greater support, documentation, and includes a machine readable rights statement.

ODI

PLATINUM: satisfies the Gold requirements, has machine readable provenance documentation, uses unique identifiers in the data, the publisher has a communications team offering support. This is an exceptional example of an information infrastructure.

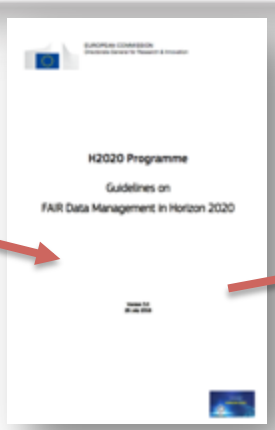


5-star deployment scheme for Open Data

Sources: Open data institute (UK), Centre for open science (US), Tim-Berners Lee

Different implementations of FAIR

Requirements for new data creation



Establishing the profile for existing data

Transformation tools to make data FAIR (Go-FAIR initiative)

FAIR badge scheme

2 User Reviews
1 Archivist Assessment
24 Downloads

- Proxy for data "quality" or "fitness for (re-)use"
- Prevent interactions among dimensions to ease scoring
- Consider Reusability as the resultant of the other three:
  - the average FAIRness as an indicator of data quality
  - (F+A+I)/3=R
- Manual and automatic scoring

First we attempted to operationalise R – Reusable as well… but we changed our mind

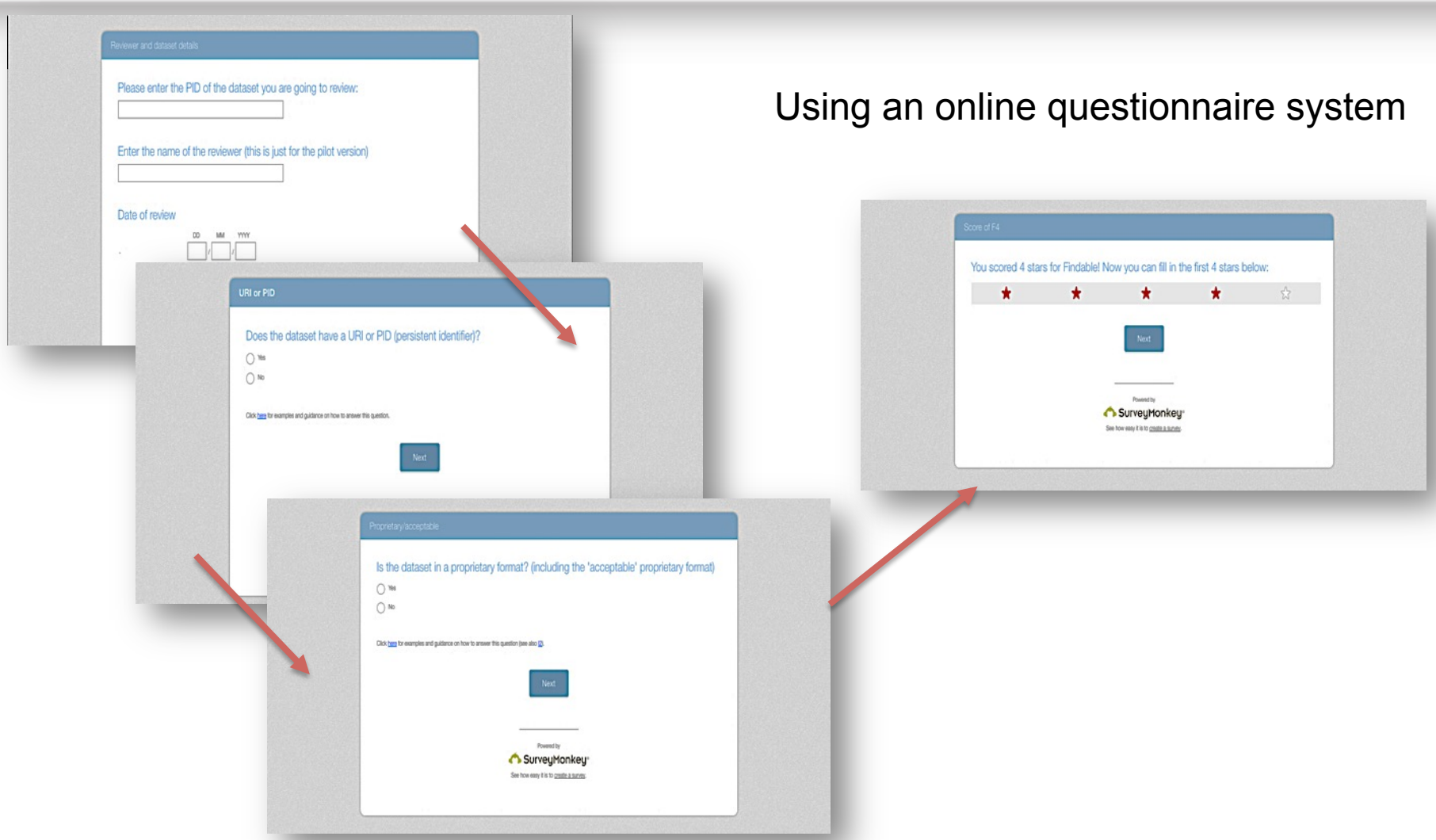
Reusable – is it a separate dimension? Partly subjective: it depends on what you want to use the data for!

Idea for operationalization	Solution
R1. plurality of accurate and relevant attributes	≈ **F2**: "data are described with rich metadata" → **F**
R1.1. clear and accessible data usage license	→ **A**
R1.2. provenance (for replication and reuse)	→ **F**
R1.3. meet domain-relevant community standards	→ **I**
Data is in a TDR – unsustained data will not remain usable	Aspect of Repository → Data Seal of Approval
Explication on how data was or can be used is available	→ **F**
Data is automatically usable by machines	→ **I**

Findable (defined by metadata (PID included) and documentation)
1. No PID nor metadata/documentation 2. PID without or with insufficient metadata 3. Sufficient/limited metadata without PID 4. PID with sufficient metadata 5. Extensive metadata and rich additional documentation available
Accessible (defined by presence of user license)
1. Metadata nor data are accessible 2. Metadata are accessible but data is not accessible (no clear terms of reuse in license) 3. User restrictions apply (i.e. privacy, commercial interests, embargo period) 4. Public access (after registration) 5. Open access unrestricted
Interoperable (defined by data format)
1. Proprietary (privately owned), non-open format data 2. Proprietary format, accepted by Certified Trustworthy Data Repository 3. Non-proprietary, open format = ‘preferred format’ 4. As well as in the preferred format, data is standardised using a standard vocabulary format (for the research field to which the data pertain) 5. Data additionally linked to other data to provide context

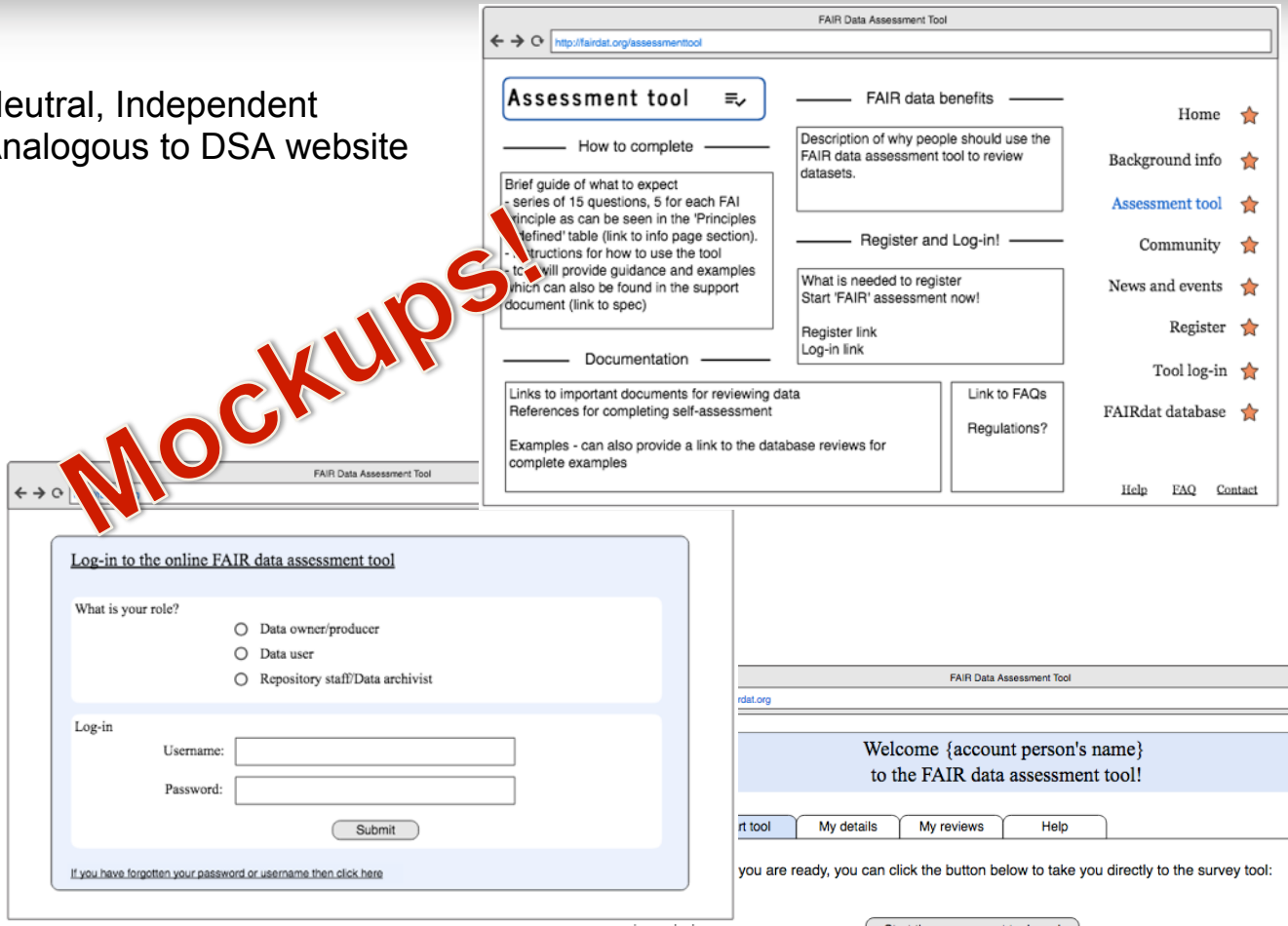
Creating a FAIR data assessment tool

Using an online questionnaire system



Website FAIRDAT

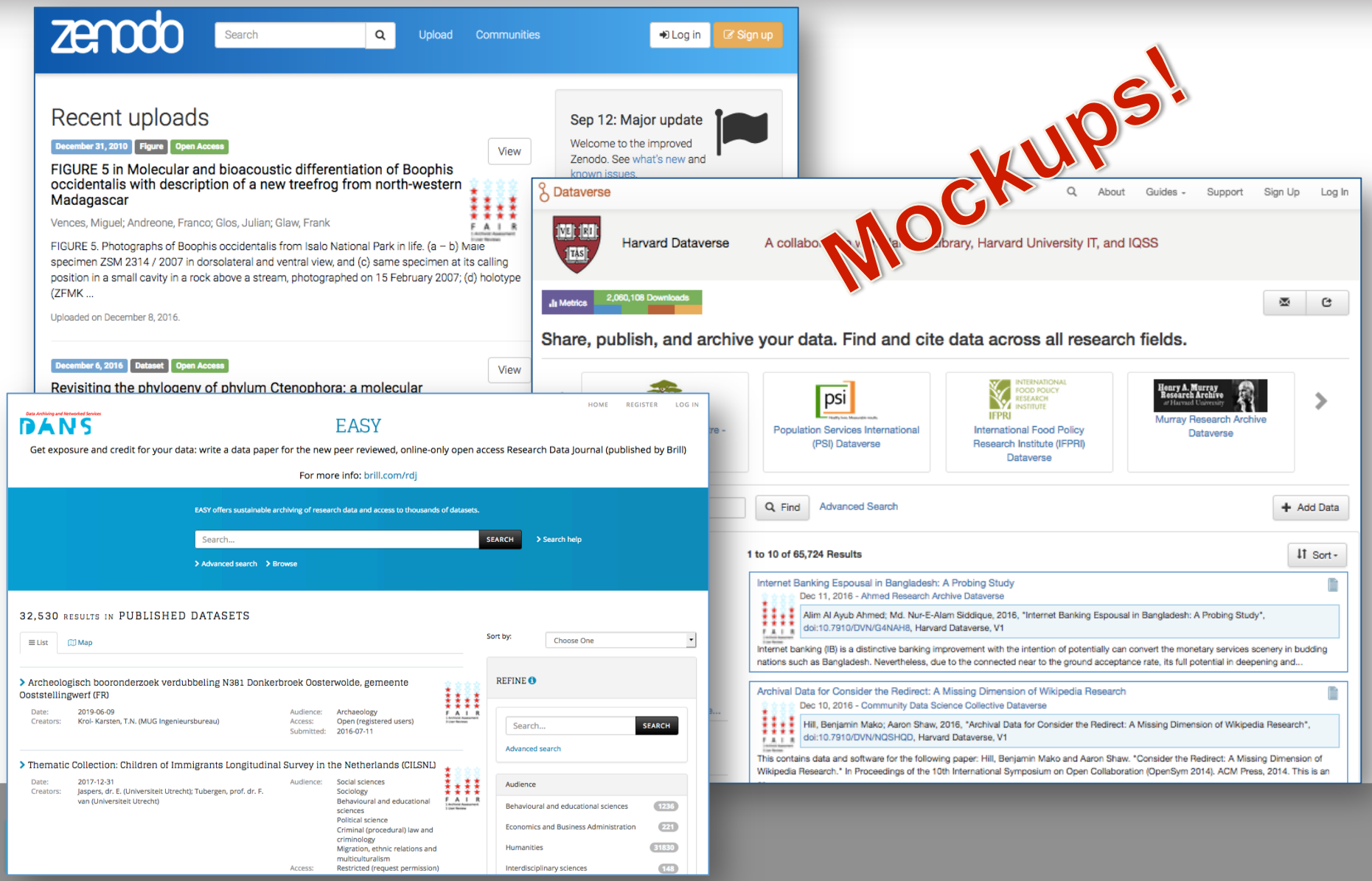
Neutral, Independent
Analogous to DSA website



To contain FAIR data assessments from any repository or website, linking to the location of the data set via (persistent) identifier
The repository can show the resultant badge, linking back to the FAIRDAT website

2 User Reviews
1 Archivist Assessment
24 Downloads

Display FAIR badges in any repository (Zenodo, Dataverse, Mendeley Data, figshare, B2SAFE, …)



Can FAIR Data Assessment be automatic?

	Criterion	Automatic? Y/N/Semi	Subjective? Y/N/Semi	Comments
F1	No PID / No Metadata	Y	N	
F2	PID / Insuff. Metadata	S	S	Insufficient metadata is subjective
F3	No PID / Suff. Metadata	S	S	Sufficient metadata is subjective
F4	PID / Sufficient Metadata	S	S	Sufficient metadata is subjective
F5	PID / Rich Metadata	S	S	Rich metadata is subjective
A1	No License / No Access	Y	N	
A2	Metadata Accessible	Y	N	
A3	User Restrictions	Y	N	
A4	Public Access	Y	N	
A5	Open Access	Y	N	
I1	Proprietary Format	S	N	Depends on list of proprietary formats
I2	Accepted Format	S	S	Depends on list of accepted formats
I3	Archival Format	S	S	Depends on list of archival formats
I4	+ Harmonized	N	S	Depends on domain vocabularies
I5	+ Linked	S	N	Depends on semantic methods used

Optional: qualitative assessment / data review

Some measuring problems encountered in tests of FAIR data assessment tool

- Assessing multi-file data sets:
  - Which are in different formats, some open, some proprietary
  - Some files well-documented, others less so
  - Some are openly accessible, others are protected
- Quality of metadata: when is metadata minimal / insufficient / sufficient / extensive / rich ?
- Use of standard vocabularies: how to define?
  - Often these apply only to a subset of the data, e.g. specific variables

Thank you for listening!

peter.doorn@dans.knaw.nl
ingrid.dillo@dans.knaw.nl
www.dans.knaw.nl
http://www.dtls.nl/go-fair/
https://eudat.eu/events/webinar/fair-data-in-trustworthy-data-repositories-webinar