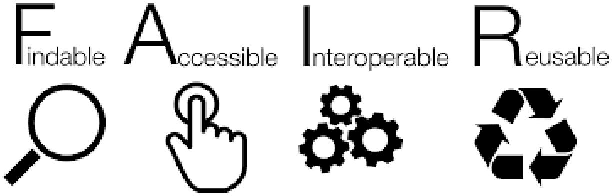
Data Archiving and Networked Services
DANS

DPHEP

Assessing the FAIRness of Datasets in Trustworthy Digital Repositories: a 5 star scale

Peter Doorn, Director DANS
Ingrid Dillo, Deputy Director DANS



2nd DPHEP Collaboration Workshop

CERN, Geneva, 13 March 2017

@pkdoorn @dansknaw

DANS is about keeping data FAIR

EXPERIENCE 50 YEARS EXPERIENCE

Mission: promote and provide permanent access to digital research resources

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

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

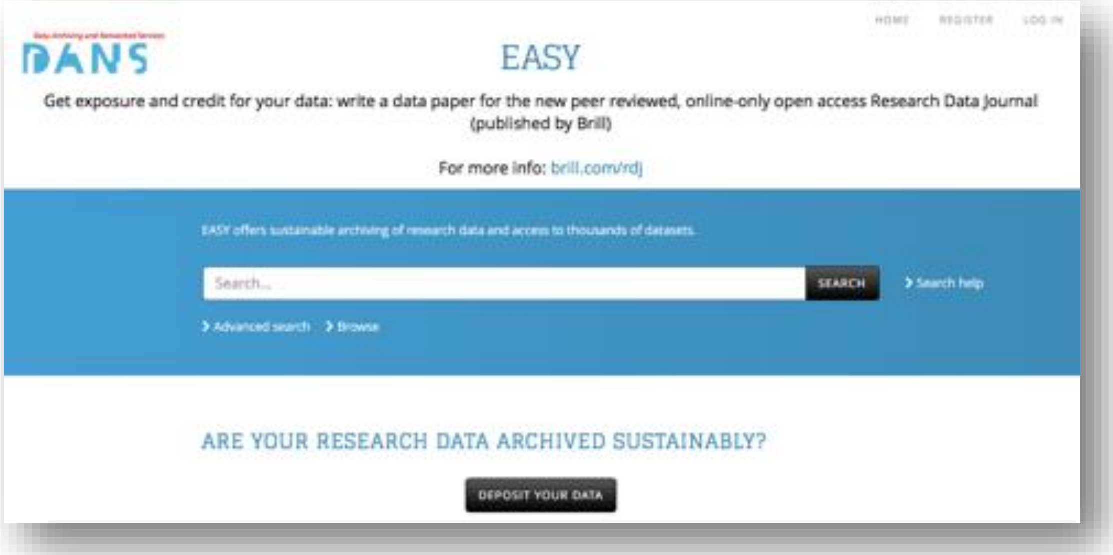
Data Archiving and Networked Services

DANS

Our core services

So, I would really like that everyone made their data available.

Watch our videos on YouTube
https://www.youtube.com/user/DANSDataArchiving



EASY

Get exposure and credit for your data: write a data paper for the new peer reviewed, online-only open access Research Data Journal (published by Brill)

For more info: brill.com/rdj

EASY offers sustainable archiving of research data and access to thousands of datasets.

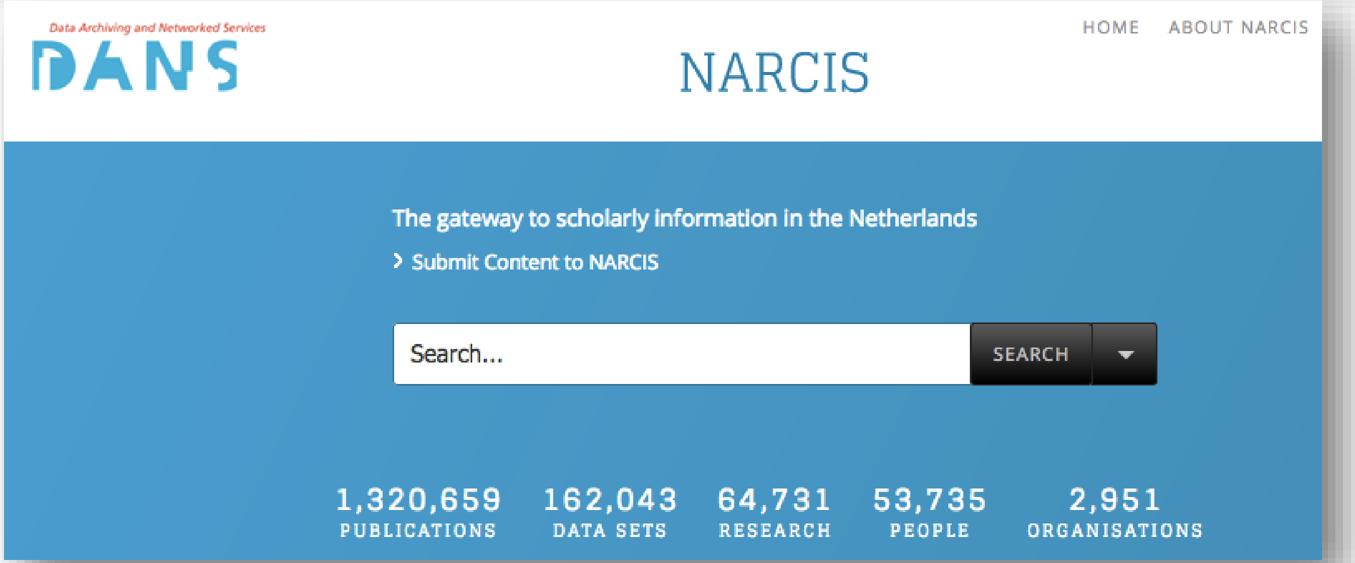
ARE YOUR RESEARCH DATA ARCHIVED SUSTAINABLY?

Data Seal of Approval

nestor Seal 2016

ICSU WORLD DATA SYSTEM

Our core services

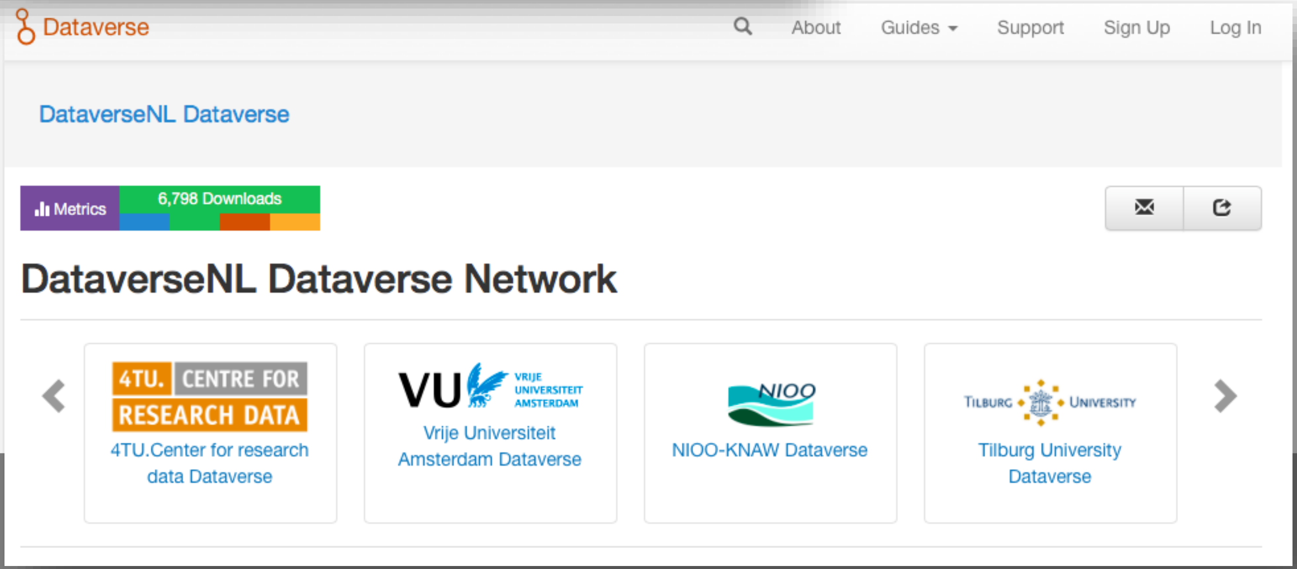


NARCIS

The gateway to scholarly information in the Netherlands

› Submit Content to NARCIS

1,320,659	162,043	64,731	53,735	2,951
PUBLICATIONS	DATA SETS	RESEARCH	PEOPLE	ORGANISATIONS



DataverseNL Dataverse

DataverseNL Dataverse Network

4TU. CENTRE FOR RESEARCH DATA
VU VRIJE UNIVERSITEIT AMSTERDAM
NIOO
TILBURG UNIVERSITY

- 4TU.Center for research data Dataverse
- Vrije Universiteit Amsterdam Dataverse
- NIOO-KNAW Dataverse
- Tilburg University Dataverse

DPHEP
Data Preservation in High Energy Physics
Collaboration for Data Preservation and Long Term Analysis in High Energy Physics

‘Policies and best practices for archival management’

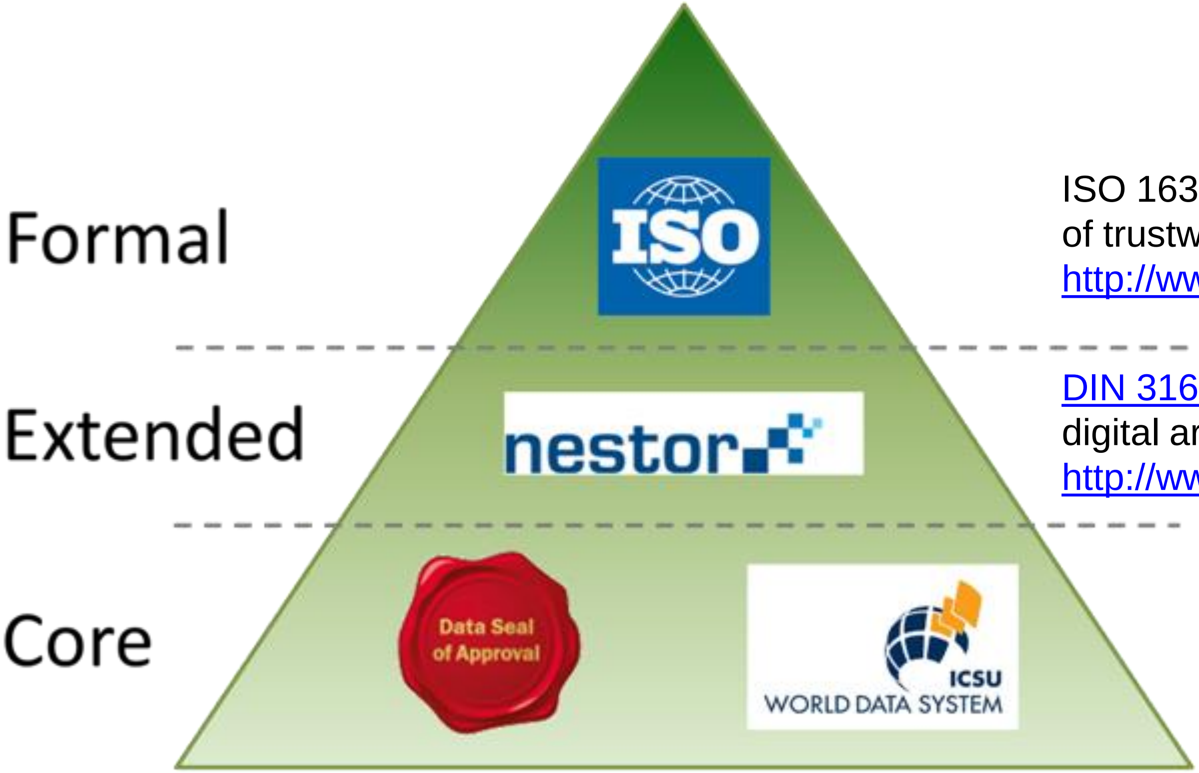
Quality (trustworthiness) of data repositories - DSA principles
Quality (fitness for use) of datasets - FAIR principles

DANS and DSA

Data Seal of Approval

- 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, community review

Complementarity:

- Geographical spread
- Disciplinary spread

Partnership

Goals:

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

Outcomes:

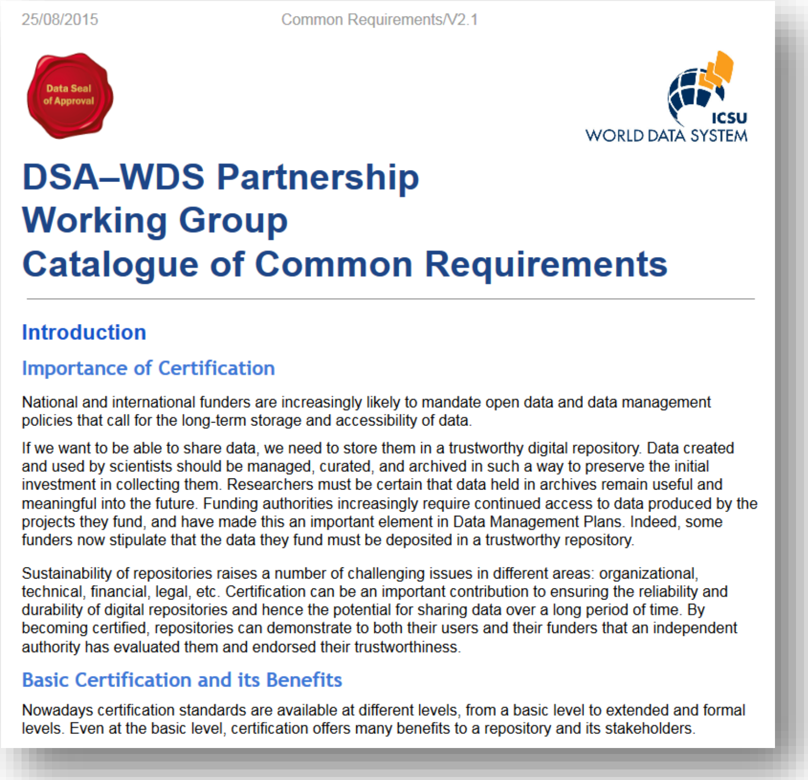
- Common catalogue of requirements for core repository assessment
- Common procedures for assessment
- Shared testbed for assessment

New common requirements

- 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.

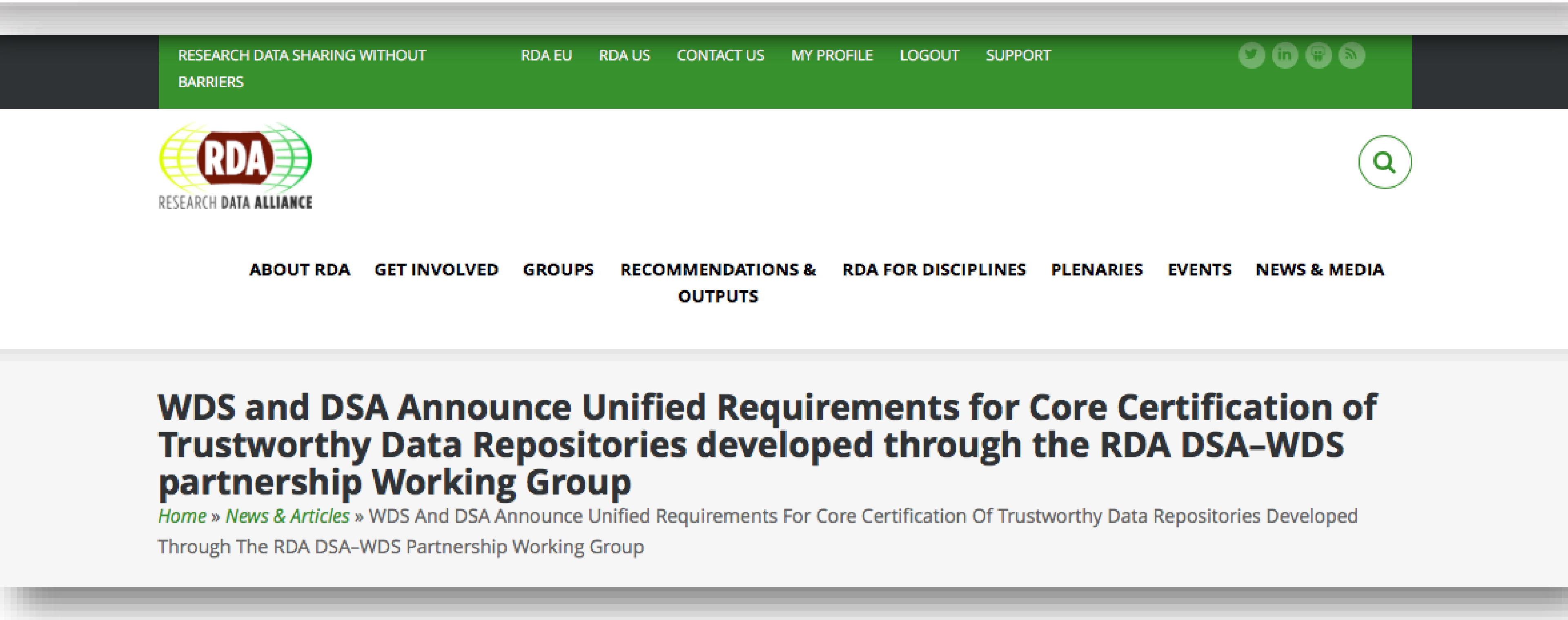
R10. The repository assumes responsibility for long-term preservation and manages this function in a planned and documented way.

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!



RESEARCH DATA SHARING WITHOUT BARRIERS RDA EU RDA US CONTACT US MY PROFILE LOGOUT SUPPORT

RDA RESEARCH DATA ALLIANCE

ABOUT RDA GET INVOLVED GROUPS RECOMMENDATIONS & OUTPUTS RDA FOR DISCIPLINES PLENARIES EVENTS NEWS & MEDIA

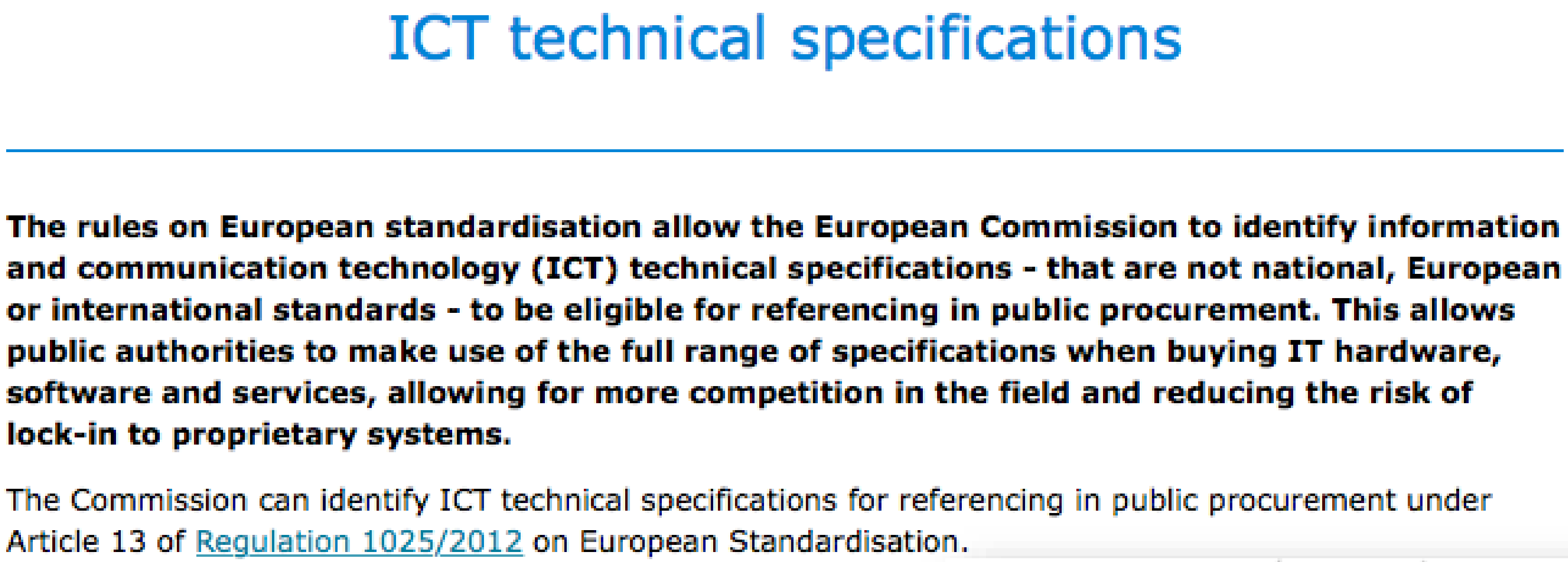
WDS and DSA Announce Unified Requirements for Core Certification of Trustworthy Data Repositories developed through the RDA DSA–WDS partnership Working Group

Home » *News & Articles* » WDS And DSA Announce Unified Requirements For Core Certification Of Trustworthy Data Repositories Developed Through The RDA DSA–WDS Partnership Working Group

http://www.datasealofapproval.org/en/news-and-events/news/2016/11/25/wds-and-dsa-announce-uni-ed-requirements-core-cert/

https://www.icsu-wds.org/news/news-archive/wds-dsa-unified-requirements-for-core-certification-of-trustworthy-data-repositories

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.

European Commission

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.

15 Criteria

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 a way 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 – research data (sets)
- Operationalize the principles as an instrument to assess FAIRness of existing datasets in certified TDRs

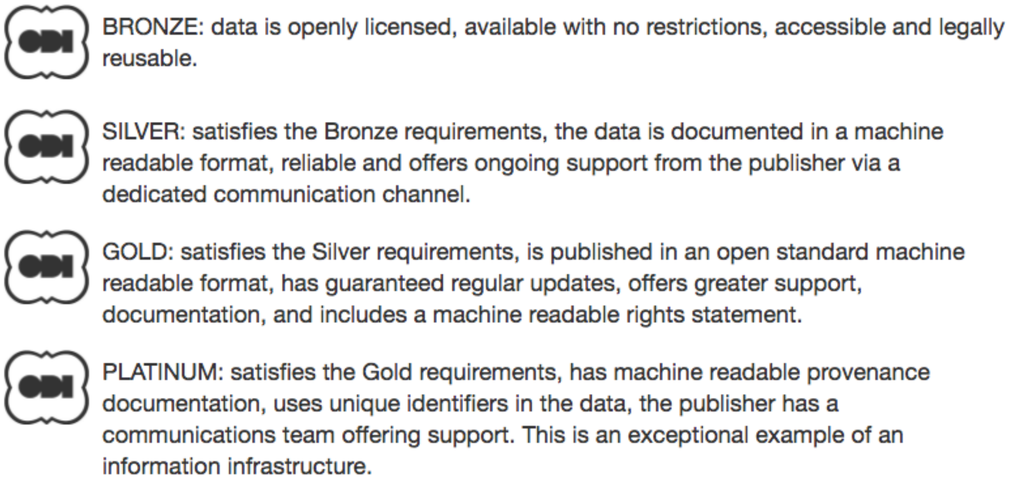
K.I.S.S.

- Keep It Simple, Stupid!
- Keep It Simple and Short!
- Keep It Simple, Short and Specific

Badges for assessing aspects of data quality and “openness”

OPEN DATA
OPEN MATERIALS
PREREGISTERED

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

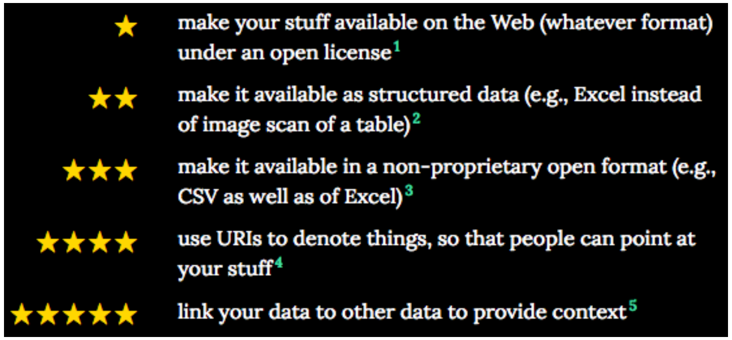


BRONZE: data is openly licensed, available with no restrictions, accessible and legally reusable.

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.

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.

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.



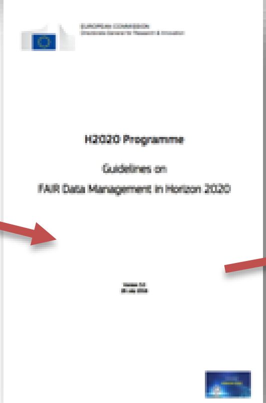
★	make your stuff available on the Web (whatever format) under an open license[1]
★★	make it available as structured data (e.g., Excel instead of image scan of a table)[2]
★★★	make it available in a non-proprietary open format (e.g., CSV as well as of Excel)[3]
★★★★	use URIs to denote things, so that people can point at your stuff[4]
★★★★★	link your data to other data to provide context[5]

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


EUROPEAN COMMISSION
Directorate-General for Research & Innovation

H2020 Programme

Guidelines on

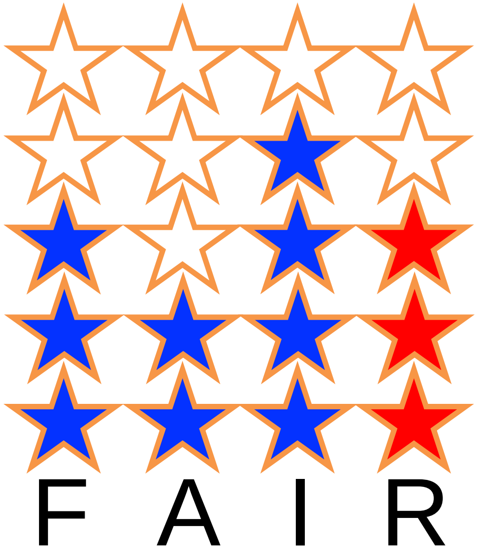
FAIR Data Management in Horizon 2020

[illegible]

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

- First Badge System based on the FAIR principles: proxy for data quality assessment
- Operationalise the original principles to ensure no 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

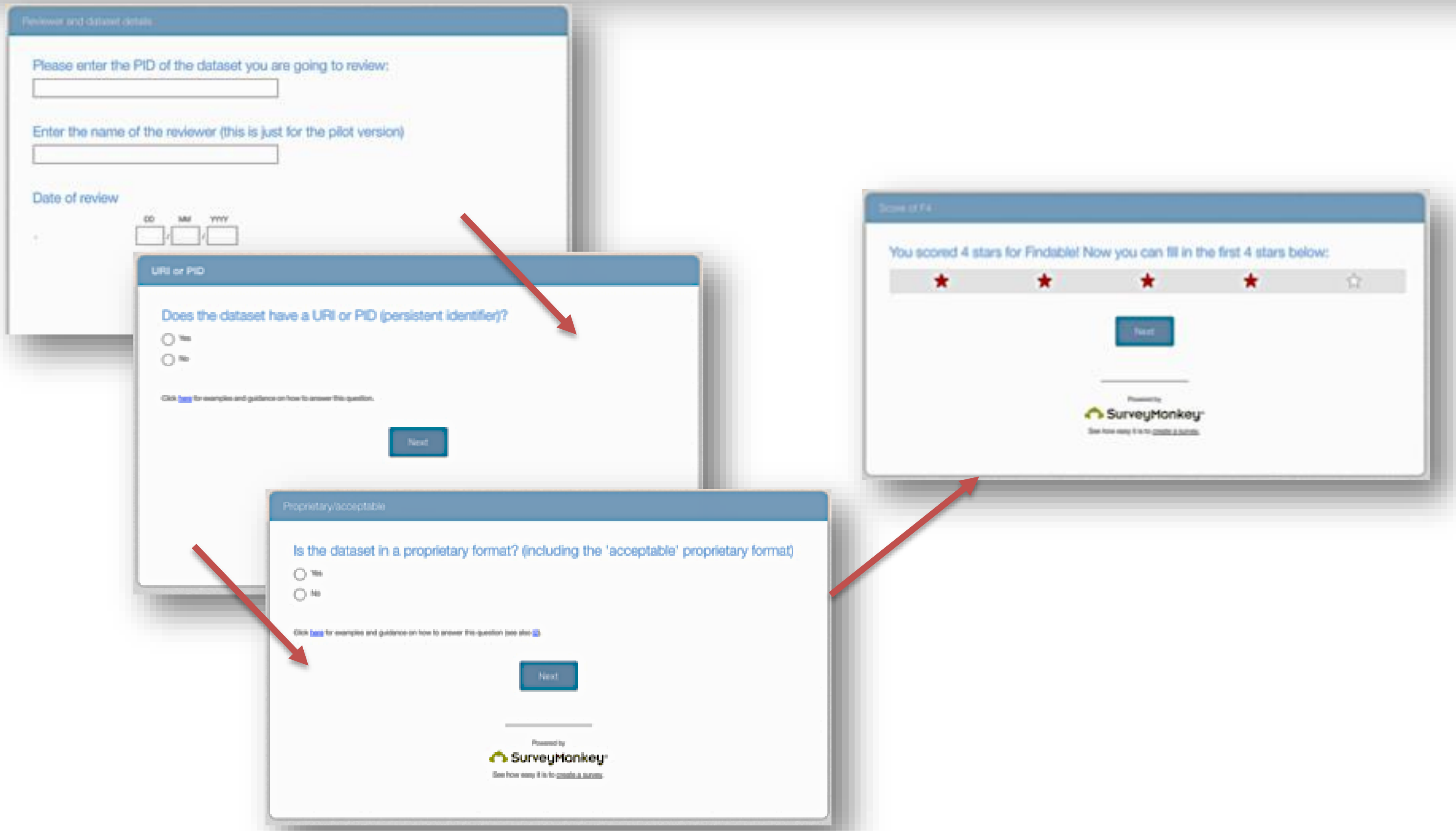
Accessible (defined by presence of user license)

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



Reviewer and dataset details

Please enter the PID of the dataset you are going to review:

Enter the name of the reviewer (this is just for the pilot version)

Date of review

DD / MM / YYYY

URI or PID

Does the dataset have a URI or PID (persistent identifier)?

○ Yes
○ No

Click here for examples and guidance on how to answer this question.

Proprietary/acceptable

Is the dataset in a proprietary format? (including the 'acceptable' proprietary format)

○ Yes
○ No

Click here for examples and guidance on how to answer this question (see also [illegible]).

Powered by SurveyMonkey

See how easy it is to create a survey.

Score of F4

You scored 4 stars for Findable! Now you can fill in the first 4 stars below:

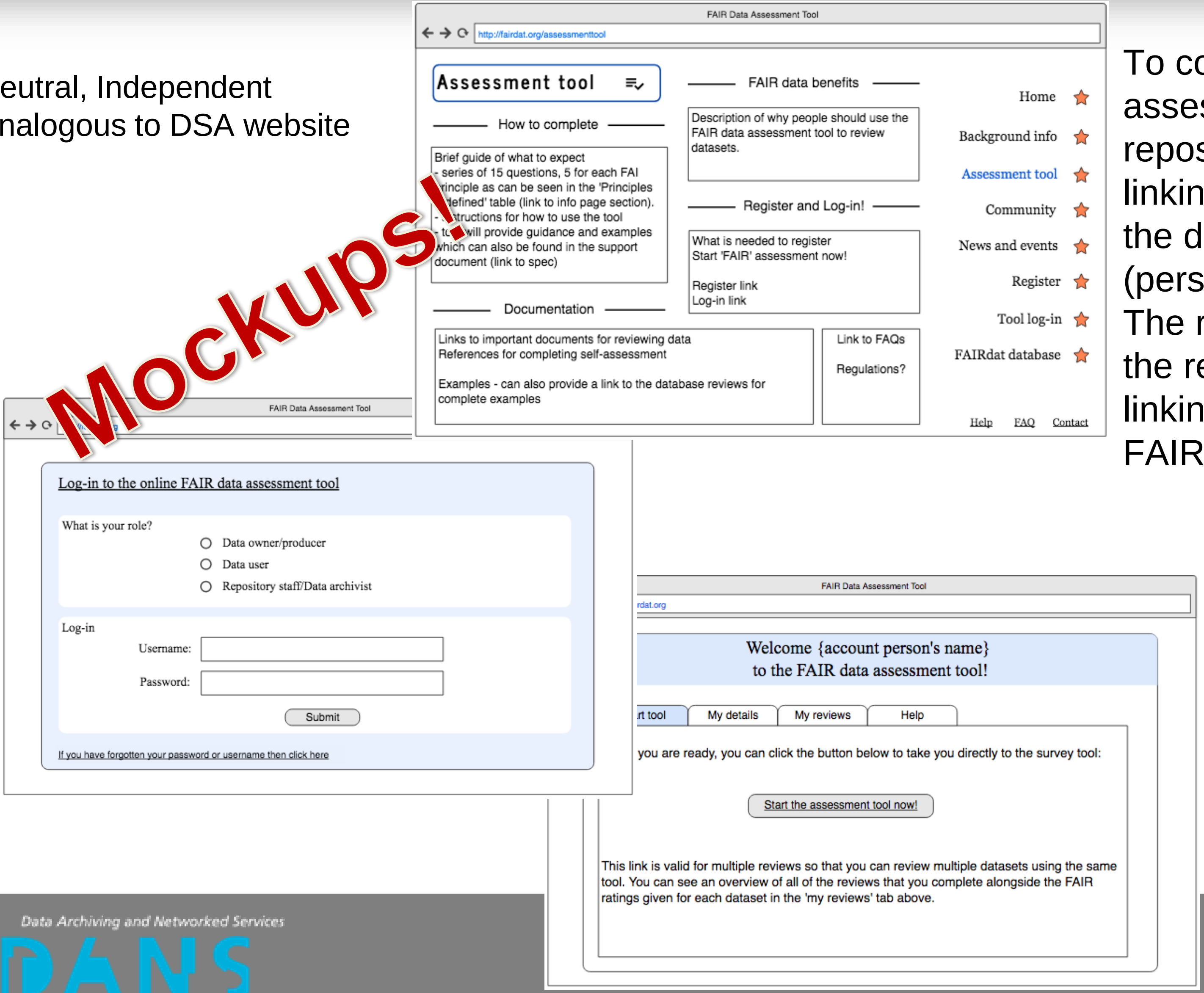
★ ★ ★ ★ ☆

Powered by SurveyMonkey

See how easy it is to create a survey.

Website FAIRDAT

Neutral, Independent
Analogous to DSA website



FAIR Data Assessment Tool

http://fairdat.org/assessmenttool

Assessment tool

How to complete

Brief guide of what to expect
- series of 15 questions, 5 for each FAI rinciple as can be seen in the 'Principles defined' table (link to info page section).
- structions for how to use the tool
- t will provide guidance and examples which can also be found in the support document (link to spec)

FAIR data benefits

Description of why people should use the FAIR data assessment tool to review datasets.

Register and Log-in!

What is needed to register
Start 'FAIR' assessment now!

Register link
Log-in link

Documentation

Links to important documents for reviewing data
References for completing self-assessment

Examples - can also provide a link to the database reviews for complete examples

Link to FAQs

Regulations?

FAIR Data Assessment Tool

Log-in to the online FAIR data assessment tool

What is your role?
○ Data owner/producer
○ Data user
○ Repository staff/Data archivist

Log-in
Username:
Password:

If you have forgotten your password or username then click here

FAIR Data Assessment Tool

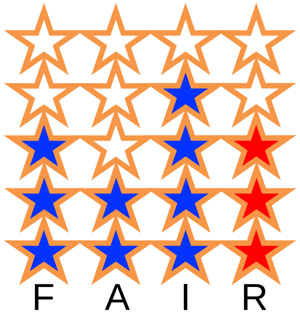
rdat.org

Welcome {account person's name}
to the FAIR data assessment tool!

you are ready, you can click the button below to take you directly to the survey tool:

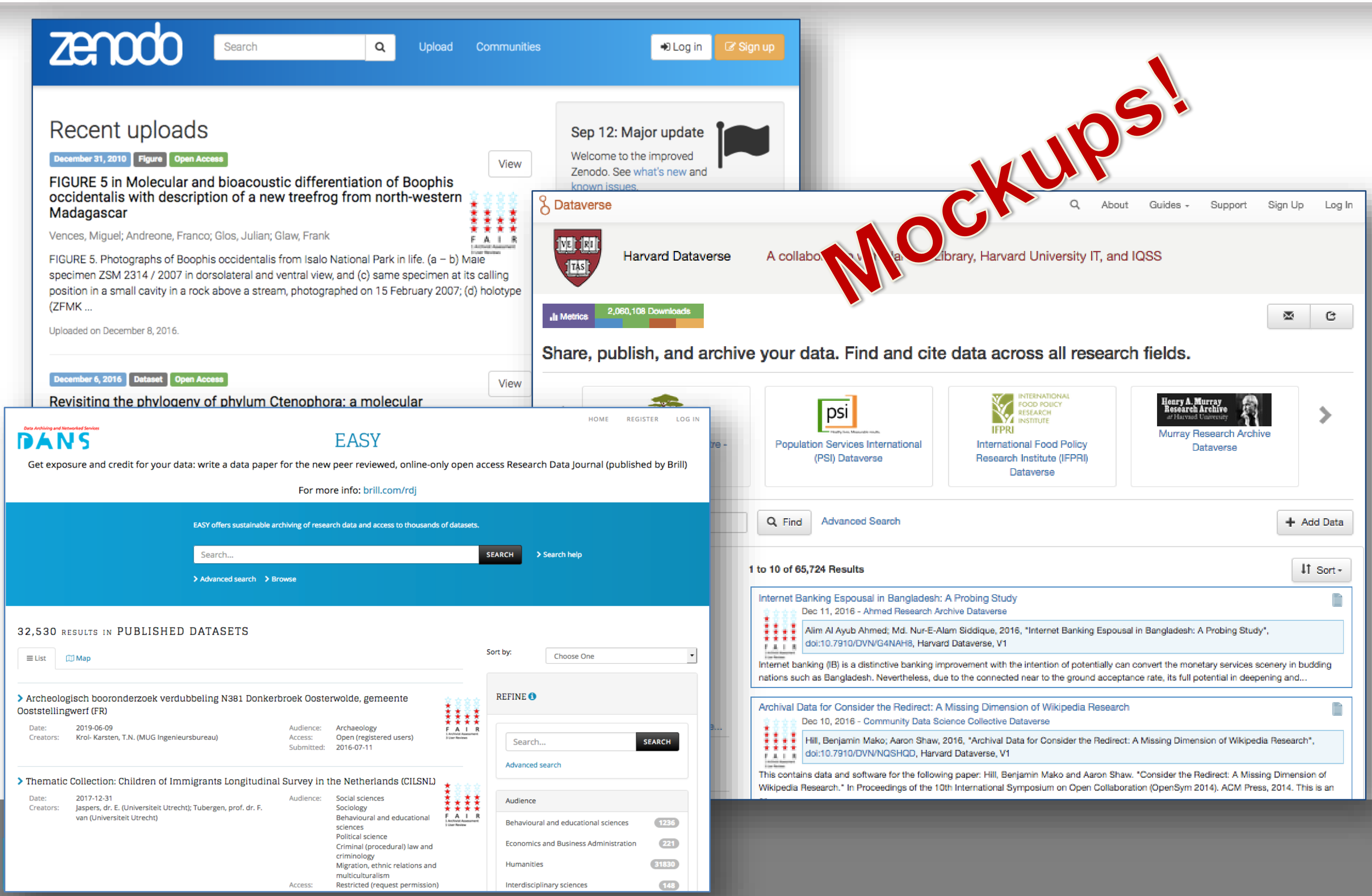
This link is valid for multiple reviews so that you can review multiple datasets using the same tool. You can see an overview of all of the reviews that you complete alongside the FAIR ratings given for each dataset in the 'my reviews' tab above.

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, …)



Recent uploads

December 31, 2010 · Figure · Open Access

FIGURE 5 in Molecular and bioacoustic differentiation of Boophis occidentalis with description of a new treefrog from north-western Madagascar

Vences, Miguel; Andreone, Franco; Glos, Julian; Glaw, Frank

FIGURE 5. Photographs of Boophis occidentalis from Isalo National Park in life. (a – b) Male specimen ZSM 2314 / 2007 in dorsolateral and ventral view, and (c) same specimen at its calling position in a small cavity in a rock above a stream, photographed on 15 February 2007; (d) holotype (ZFMK ...

Uploaded on December 8, 2016.

December 6, 2016 · Dataset · Open Access

Revisiting the phylogeny of phylum Ctenophora: a molecular

Sep 12: Major update

Welcome to the improved Zenodo. See what's new and known issues.

Harvard Dataverse

A collabo… Library, Harvard University IT, and IQSS

Share, publish, and archive your data. Find and cite data across all research fields.

Population Services International (PSI) Dataverse

International Food Policy Research Institute (IFPRI) Dataverse

Murray Research Archive Dataverse

1 to 10 of 65,724 Results

Internet Banking Espousal in Bangladesh: A Probing Study

Dec 11, 2016 - Ahmed Research Archive Dataverse

Alim Al Ayub Ahmed; Md. Nur-E-Alam Siddique, 2016, "Internet Banking Espousal in Bangladesh: A Probing Study", doi:10.7910/DVN/G4NAH8, Harvard Dataverse, V1

Internet banking (IB) is a distinctive banking improvement with the intention of potentially can convert the monetary services scenery in budding nations such as Bangladesh. Nevertheless, due to the connected near to the ground acceptance rate, its full potential in deepening and...

Archival Data for Consider the Redirect: A Missing Dimension of Wikipedia Research

Dec 10, 2016 - Community Data Science Collective Dataverse

Hill, Benjamin Mako; Aaron Shaw, 2016, "Archival Data for Consider the Redirect: A Missing Dimension of Wikipedia Research", doi:10.7910/DVN/NQSHQD, Harvard Dataverse, V1

This contains data and software for the following paper: Hill, Benjamin Mako and Aaron Shaw. "Consider the Redirect: A Missing Dimension of Wikipedia Research." In Proceedings of the 10th International Symposium on Open Collaboration (OpenSym 2014). ACM Press, 2014. This is an

Data Archiving and Networked Services

Get exposure and credit for your data: write a data paper for the new peer reviewed, online-only open access Research Data Journal (published by Brill)

For more info: brill.com/rdj

EASY offers sustainable archiving of research data and access to thousands of datasets.

32,530 RESULTS IN PUBLISHED DATASETS

Archeologisch booronderzoek verdubbeling N381 Donkerbroek Oosterwolde, gemeente Ooststellingwerf (FR)

Date:	2019-06-09	Audience:	Archaeology
Creators:	Krol- Karsten, T.N. (MUG Ingenieursbureau)	Access:	Open (registered users)
		Submitted:	2016-07-11

Thematic Collection: Children of Immigrants Longitudinal Survey in the Netherlands (CILSNL)

Date:	2017-12-31	Audience:	Social sciences Sociology Behavioural and educational sciences Political science Criminal (procedural) law and criminology Migration, ethnic relations and multiculturalism
Creators:	Jaspers, dr. E. (Universiteit Utrecht); Tubergen, prof. dr. F. van (Universiteit Utrecht)		
		Access:	Restricted (request permission)

Behavioural and educational sciences	1236
Economics and Business Administration	221
Humanities	31830
Interdisciplinary sciences	148

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

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