Data Archiving and Networked Services
DANS

FAIR enough:

A researcher friendly checklist and a not so friendly assessment of the FAIRness of repositories

Peter Doorn, Director DANS

CORE TRUST SEAL

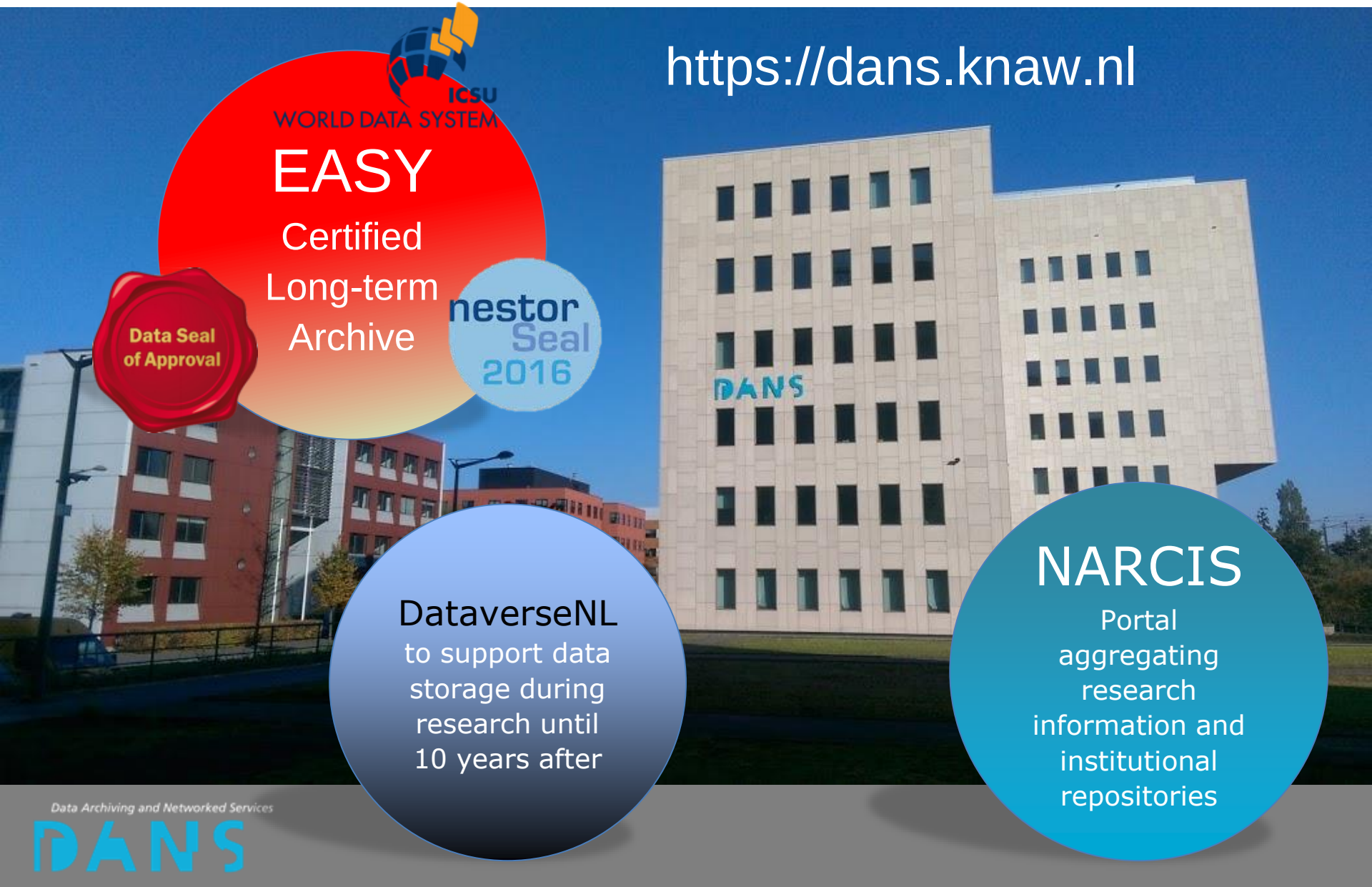
Findable Accessible Interoperable Reusable

eScience - FAIR Science

PLAN-E Workshop @ IEEE eScience Conference
Amsterdam, 29 October 2018

@pkdoorn @dansknaw

DANS is about keeping data FAIR



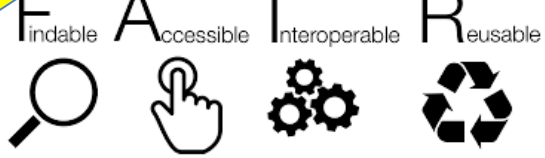
Data Archiving and Networked Services
DANS

One Step Forward, Two Steps Back:

A Design Framework and Exemplar Metrics for Assessing FAIRness in Trustworthy Data Repositories

Peter Doorn DANS

Last spring in Berlin



WG RDA/WDS Assessment of Data Fitness for Use RDA 11th Plenary meeting
Berlin, 22-03-2018

@pkdoorn @dansknaw

Towards a FAIR Data Assessment Tool

DSA Principles (for data repositories)	…ples (for data
data can be fou… internet	…indable
dat…	**A**ccessible
…a **usable format**	**I**nteroperable
data are **reliable**	**R**eusable
data can be **referred** to	(citable)

Previously on RDA: Barcelona 2017

F A I R

…AIR Badging scheme

https://www.surveymonkey.com/r/fairdat

Data Archiving and Networked Services
DANS

https://eudat.eu/events/webinar/fair-data-in-trustworthy-data-repositories-webinar

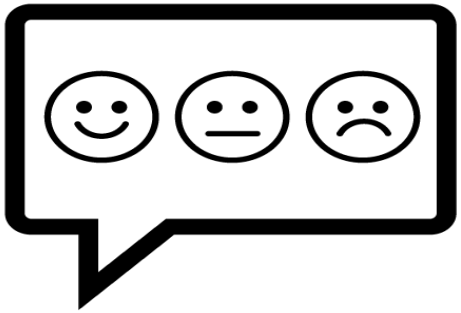
Testing the FAIRdat prototype

Test in 4 repositories, summer 2017

Name of Repository	Number of Datasets	Number of Reviewers	Number of reviews
VirginiaTech	5	1	5
MendeleyData	10	3 (for 8 datasets) 2 (for 2 datasets)	28
Dryad	9	3 (for 2 datasets) 2 (for 3 datasets)	16
CCDC	11	? (no names) 2 (for 1 dataset)	12

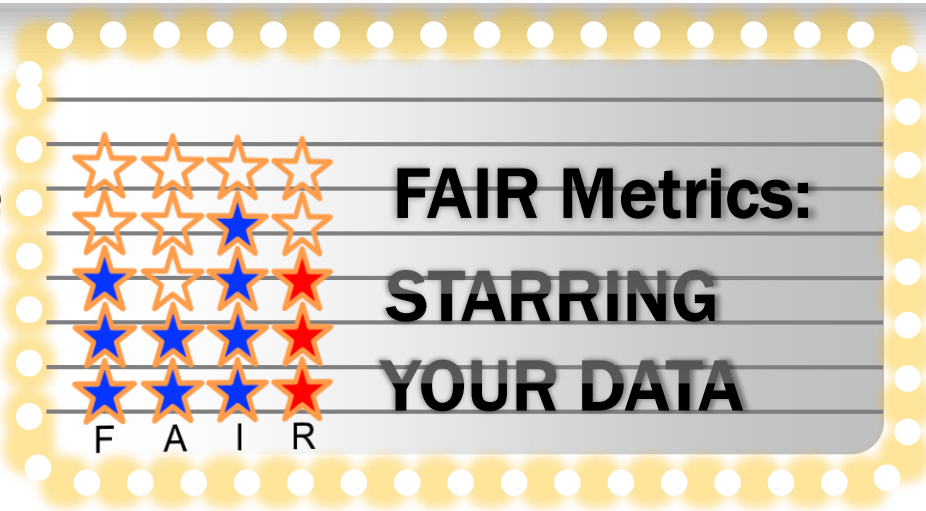
Test at Open science Fair, Athens 2017

17 participants



+ tests within DANS

Pros and Cons of FAIRdat prototype



Pros, positive feedback

- Simple/easy to use questionnaire
- Well-documented
- Useful

Cons, negative feedback

- Questionnaire oversimplified?
- Some requirements of **R**eusability missing/shifted

Other observations

- Variances in FAIR scores across multiple reviewers due to subjectivity
- Some like starring datasets, others not (should open data score higher than closed data?)
- Assessing multi-file data sets with different properties

Other challenges

- **Subjectivity in assessment of principles**
  - F2 “rich metadata”
  - I1 “broadly applicable language for knowledge representation”
  - R1 “plurality of attributes”
  - R1.2 “detailed provenance”
  - R1.3. “domain relevant community standards”
  - Use of standard vocabularies: how to define?
- **Misunderstanding of question/meaning of principle**
- **Several questions can be answered at the level of the repository**

Slide credits: Eleftheria Tsoupra

(Self) assessment of DANS archive on the basis of the FAIR principles (& metrics)

Delft University: DANS EASY complies with 11 out of 15 principles, for 2 DANS does not comply (I2 & R1.2), for 2 more it is unclear (A2 & R1.3)

Self assessment:

- Some metrics: FAIRness of DANS archive could be improved
  - E.g.: Machine accessibility; Interoperability requirements; Use of standard vocabularies; Provenance
- Some metrics: we are not sure how to apply them
  - E.g.: PID resolves to landing page (metadata), not to dataset; Dataset may consist of multiple files without standard PID
- Sometimes the FAIR principle itself is not clear
  - E.g.: Principle applies to both data and metadata; What does interoperability mean for images or PDFs? Are some data types intrinsically UNFAIR? Some terms are inherently subjective (plurality, richly)

FAIR enough checklist

COMING SOON!

Sneak Preview

- by Eliane Fankhauser @ DANS
- Quiz/checklist format
- Easy to understand and fill out by any researcher
- Focus on awareness raising rather than on FAIR orthodoxy

FAIR assessment at the level of the repository

- FAIR principles make no claim to which level of granularity they pertain: repository, collection, data set, file, record, triple, ...
- However, they often mention "(meta)data", which we interpret as pertaining *both* to **data** *and* **metadata**
- For data in trustworthy (certified CTS) repositories, most FAIR principles are taken care of for all data and metadata in the repository
    - Only F2, I2, I3, R1 (R1.2 and R1.3) can vary for metadata in a CTS certified repository
    - Only I1, I2, I3, R1 (R1.3) can vary for data (sets and files) in a CTS certified repository

FAIR principles turned into questions at three levels - Findable

For all Data and Metadata in a Repository	**For Metadata**	**For Data (Sets and Files or other subunits of information)**
F1: Does the repository assign globally unique and persistent identifiers to the **data(sets and files)** in the repository?		
F1: Does the repository assign a globally unique and persistent identifier to the **metadata**?		
F2: Are the **metadata** that the repository offer in a machine-readable form, that is to say, in a format that can be easily processed by machines?	Are the data described with rich metadata?	
F3: Do the **metadata** that the repository offer clearly and explicitly include the identifier of the **data** it describes?		
F4: Are the **data** in the repository registered or indexed in a searchable resource?*		
F4: Are the **metadata** that the repository offers registered or indexed in a searchable resource?		

* This is only possible for openly accessible data

FAIR principles turned into questions at three levels - Accessible

For all Data and Metadata in a Repository	For Metadata	For Data (Sets and Files or other subunits of information)
A1: Are the **metadata** retrievable by their identifier using a standardized communications protocol?		
A1: Are the **data(sets and files)** retrievable by their identifier using a standardized communications protocol?		
A1.1: Is the protocol open, free, and universally implementable?		
A1.2: Does the protocol allow for an authentication and authorization procedure, when required?		
A2: Are the metadata accessible, even when the data are no longer [or not] available		

FAIR principles turned into questions at three levels - Interoperable

For all Data and Metadata in a Repository	For Metadata	For Data (Sets and Files or other subunits of information)
I1: Are the **metadata** in a formal, accessible, shared, and broadly applicable language for knowledge representation?		I1: Are the **data** in a formal, accessible, shared, and broadly applicable language for knowledge representation? [2]
I2. Does the **metadata** schema used by the repository use vocabularies that follow FAIR principles?	I2. Do the **metadata** use vocabularies that follow FAIR principles? [1]	I2. Do the **data** use vocabularies that follow FAIR principles? [3]
I3. Does the **metadata** schema the repository uses include one or more elements to refer to other metadata or data?	I3. Do the **metadata** include qualified references to other metadata?	I3. Do the **data** include qualified references to other data?

[1] in particular F and A
[2] answer may vary for individual files within a dataset
[3] especially F and A; answer may depend on level of granulation

FAIR principles turned into questions at three levels - Reusable

For all Data and Metadata in a Repository	**For Metadata**	**For Data (Sets and Files or other subunits of information)**
R1. Are the attributes of the **metadata schema** relevant and sufficient?	R1. Are the **metadata** sufficient, accurate and relevant?	R1. Are the **data** accurate (or fit for purpose?)
R1.1. Does the repository supply clear and accessible **data** usage licenses?		
R1.1. Does the repository supply clear and accessible licenses for the **metadata**?		
R1.2. Does the **metadata schema** of the repository contain one or more elements to describe the provenance of the data?	R1.2. Is the provenance of the **metadata** accurately described?	R1.2. Is the provenance of the **data** accurately described in the metadata or additional documentation?
R1.3. Does the **metadata schema** of the repository meet domain standards of the community or communities it serves?	R1.3. Do the **metadata** or does the additional documentation meet domain-relevant community standards?	R1.3. Do the **data** meet domain-relevant community standards?

Next Steps

- Paper on FAIR self assessment of DANS EASY
- Establish where and how trustworthy repositories can improve their FAIRness – perhaps propose extensions to Core Trust Seal
- "FAIR enough" checklist for researchers as a tool to raise awareness
- Deal with as many FAIR principles/metrics as possible at the level of the repository
- Focus FAIR assessment on what varies within trustworthy repositories
    - Have separate metrics/questions at level of dataset, metadata and single data file (or other subunits of data collections)
    - Questionnaire approach remains useful as a FAIR data review tool