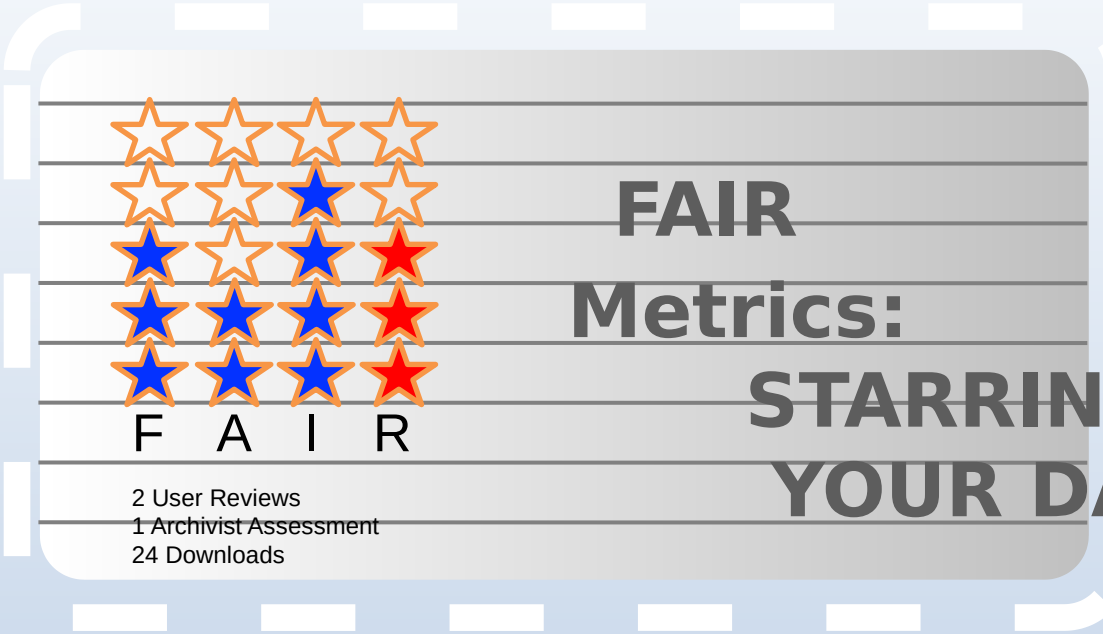
Data Archiving and Networked Services

DANS

FAIR Data Principles and Grey Literature



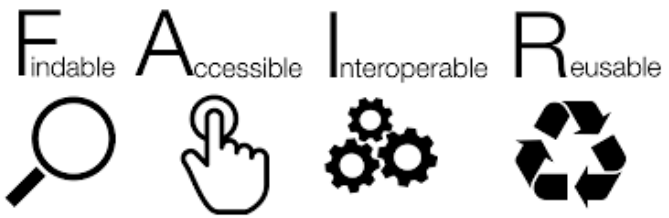
Peter Doorn (DANS)

19th International Conference on Grey Literature

Public Awareness and Access to Grey Literature
Rome, October 23-24, 2017

What is FAIR data and why is it important?

- Open Science has become a policy priority: not only publications need to be openly accessible, but also other research outputs, especially research data
- Principle: open if possible, protected if necessary
- Legitimate restrictions to openness:
  - to protect privacy
  - data creator must be able to publish first
- Openness itself is not enough, data must also be:
  - Findable F
  - Accessible A
  - Interoperable I
  - Reusable R



Do the FAIR principles also apply to Grey Literature?

Everybody loves FAIR!

Everybody wants to be FAIR… But what does that mean? How to put the principles into practice?

FAIR Metrics Group

GOALS:

- Develop a broadly useable framework and tool for FAIR assessment
- Harmonize current ongoing efforts

Expected outcomes:

- Checklist of clearly defined metrics
- Indication of how these metrics can be readily implemented (e.g. self-reporting, automated)
- Proposal(s) for undertaking a FAIR assessment of a digital resource
- One or more implementations for performing an assessment

FAIR Metrics Form

Metric Identifier
Metric Name
To which principle does it apply?
What is being measured?
Why should we measure it?
What must be provided?
How do we measure it?
What is a valid result?
For which digital resource(s) is this relevant?
Examples of their application across types of digital resource
Comment

Can we do it?

The European Open Science Cloud for Research pilot project is funded by the European Commission, DG Research & Innovation under contract no. 739563
www.eoscpilot.eu

Resemblance Data Seal of Approval – 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


All data sets in a Trusted Repository are FAIR, but some are more FAIR than others
Data Seal of Approval
CORE TRUST SEAL

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

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

DANS FAIR Core

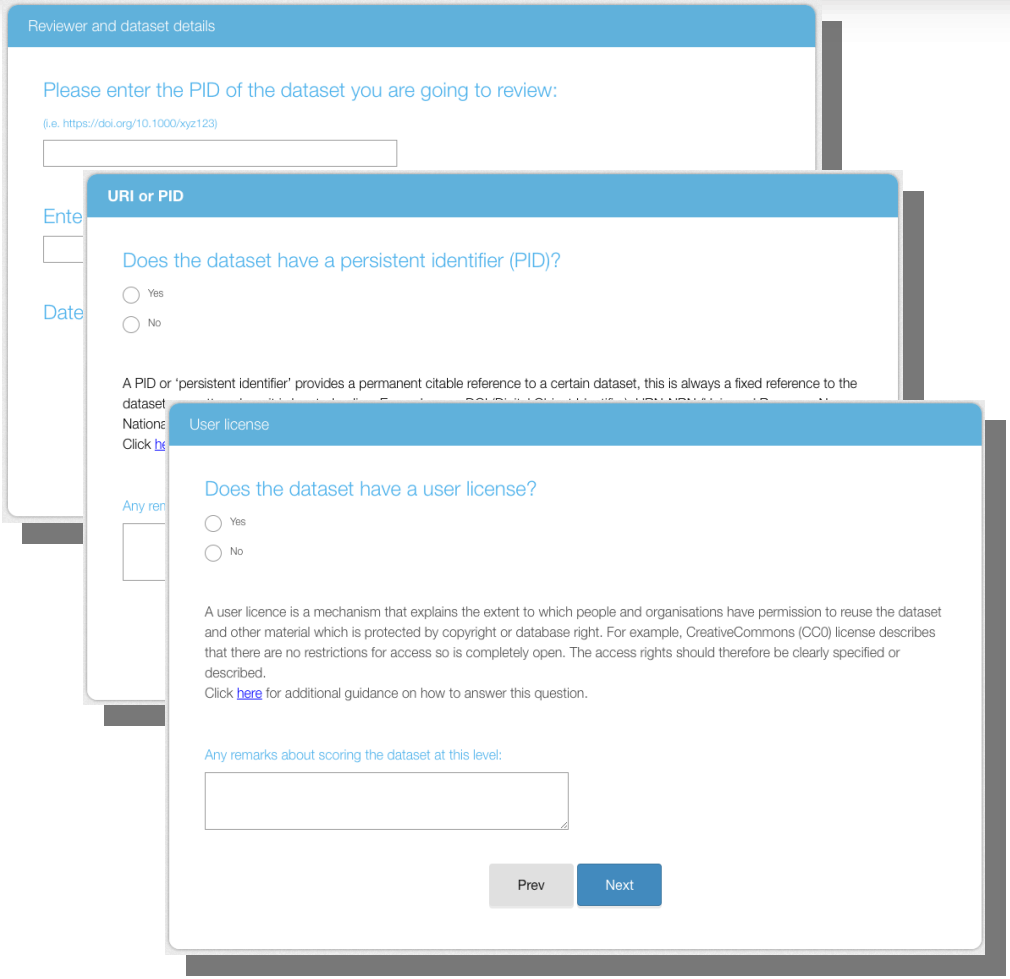
Accessible (defined by presence of user license)

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)

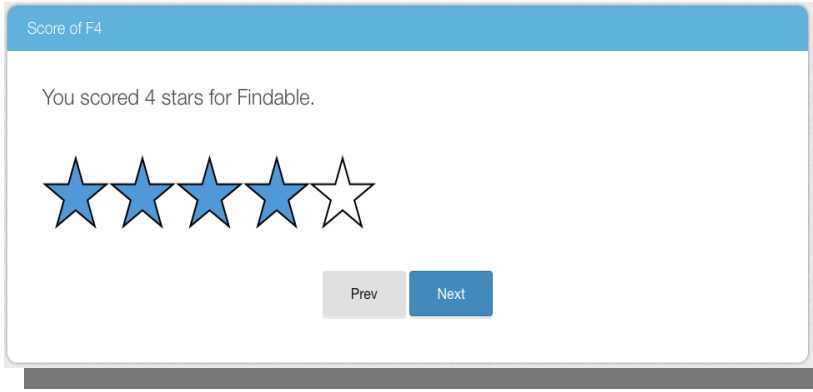
1. Proprietary (privately owned), non-open format data
2. Proprietary format, accepted by Certified Trustworthy Data Repository
3. Non-proprietary, open format = 'preferred format'
4. As well as in the preferred format, data is standardised using a standard vocabulary format (for the research field to which the data pertain)
5. Data additionally linked to other data to provide context

Creating the FAIR Data Assessment Tool

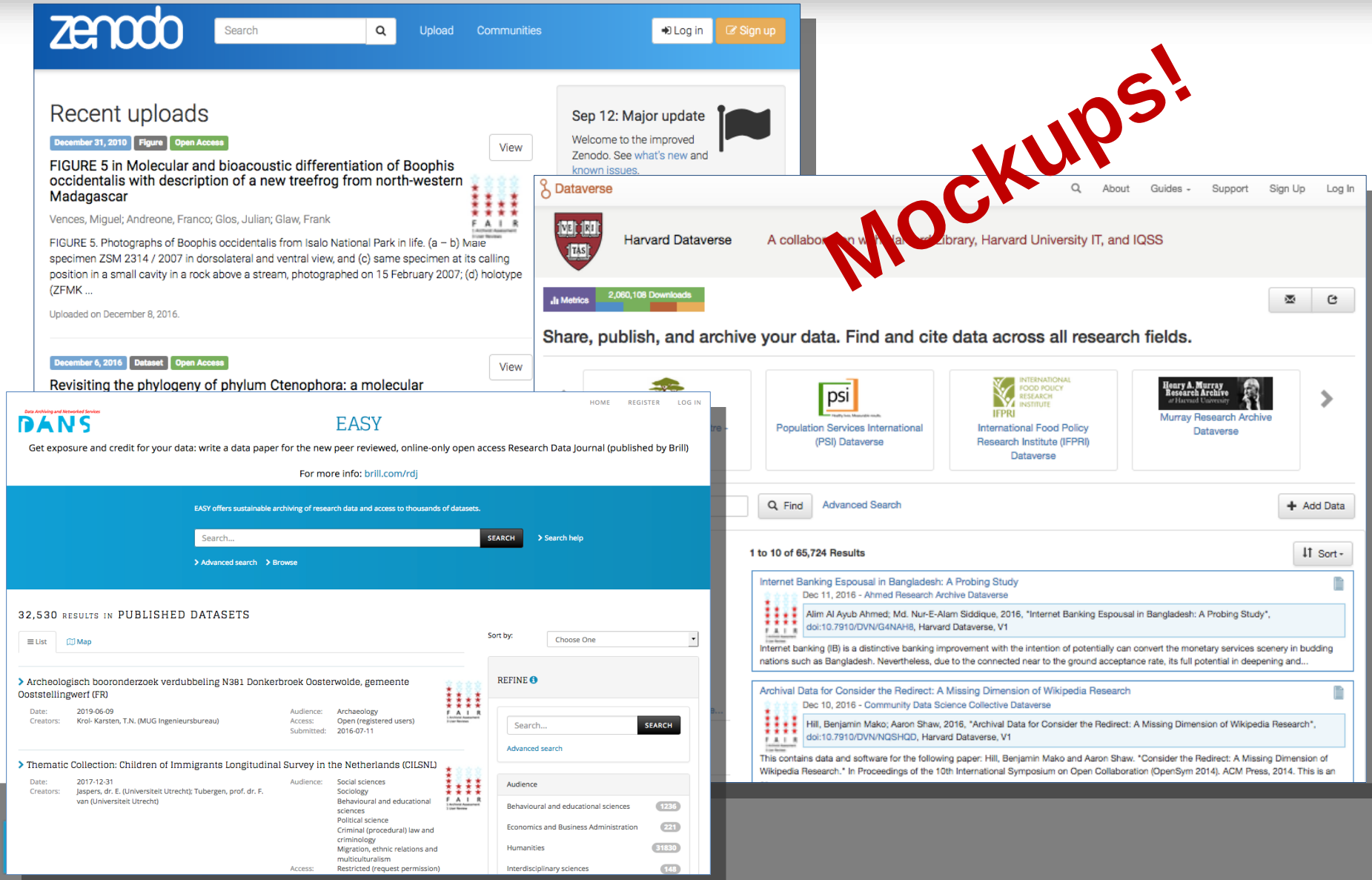


Using an online questionnaire system

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



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



Testing the prototype

- How to assess multi-file datasets?
  - e.g. DANS archives: 35,000 data sets, 4 million files (on average > 100 files per dataset)
- Do the same principles apply to all data types?
  - eg. images, PDFs?
- Operationalizing Reusability?
- Max score for Accessibility of sensitive data?
- Multiple reviewer variance?
- Scores are additive
- Do the metrics reflect the FAIR principles well?
  - Replace DANS metrics by those of Go FAIR metrics group?

Distribution of top 20 of file types/formats in DANS EASY archive

How to interpret FAIRness of:

- images
- pdf
- gis-files
- ms-word documents

Does that make sense at all?
Are these also “research data”?
How to assess multi-file datasets?

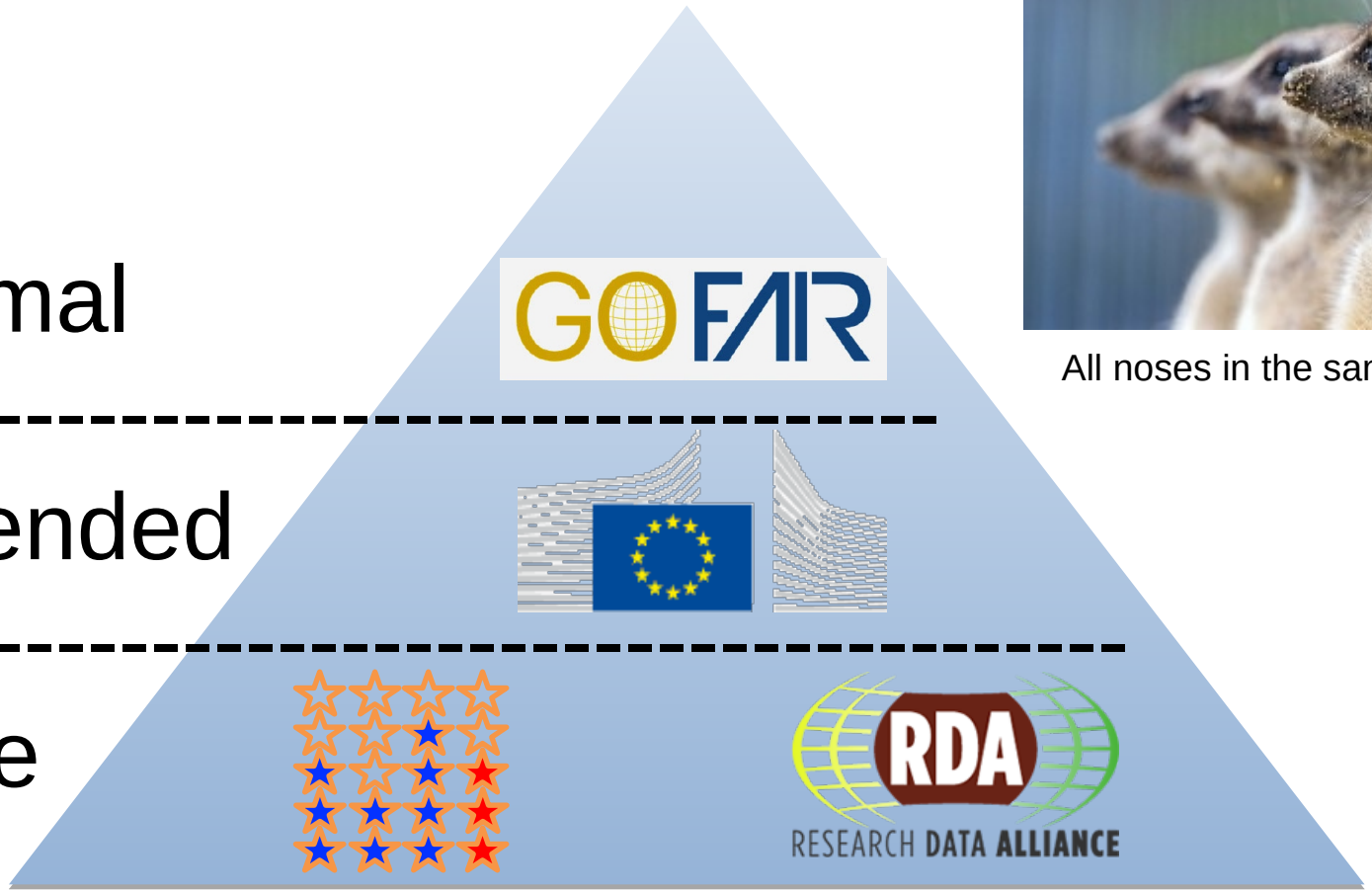
What does it mean for Grey Literature to be FAIR?

Rank	Extension	File type	N of files	% of total
1	jpg	image/jpeg	2.633.879	64,5%
2	pdf	application/pdf	164.630	4,0%
3	tif	image/tiff	137.408	3,4%
4	csv	data/plain	129.762	3,2%
5	txt	text/plain	122.136	3,0%
6	mid	gis/mid	57.415	1,4%
7	mif	gis/mif	57.029	1,4%
8	cmdi	text/xml	53.054	1,3%
9	tab	text/plain	46.431	1,1%
10	map	gis/map	43.297	1,1%
11	id	application/octet-stream	42.759	1,0%
12	dat	data/	41.671	1,0%
13	dbf	data/dbase	41.358	1,0%
14	xml	text/xml	35.681	0,9%
15	gz	application/gzip	30.327	0,7%
16	doc	text/ms-word	30.281	0,7%
17	shx	???/octet-stream	21.947	0,5%
18	shp	gis/shape-file	21.946	0,5%
19	gif	image/gif	21.687	0,5%
20	dxf	image/x-dxf	21.262	0,5%
	Subtotal	top 20 formats	3.753.960	91,9%
	> 1200 other formats/extensions		331.708	8,1%
	All		4.085.668	100,0%

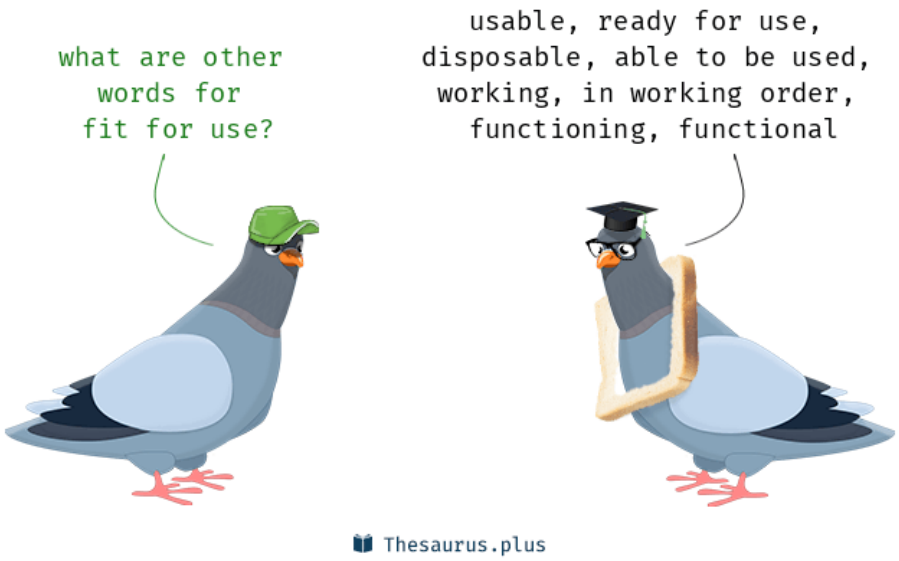
Towards a FAIR Framework?

Analogous to Certification Framework?

All noses in the same direction?



Thank you for listening!



peter.doorn@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

Data Archiving and Networked Services
DANS

Thanks to Ingrid Dillo and Emily Thomas for their contributions