

## FAIR Data Principles and Grey Literature

Peter Doorn (DANS)



19<sup>th</sup> 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

F

- Openness itself is not enough, data must also be:
  - Findable
  - Accessible A
  - Interoperable
  - Reusable
     R



## Do the FAIR principles also apply to Grey Literature?



## **Everybody loves FAIR!** got fairness? Being Fair Findoble • Accessible MORE • loteropersble THAN • Geosable Everybody wants to be FAIR... But what does that 00 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?

# Resemblance Data Seal of Approval – FAIR principles

DSA Principles (for data repositories)	FAIR Principles (for data sets)
data can be <b>found</b> on the internet	Findable
data are <b>accessible</b>	Accessible
data are in a <b>usable format</b>	Interoperable
data are <b>reliable</b>	Reusable
data can be <b>referred</b> 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







# 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

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

Reviewer and datas	et details the PID of the dataset you are going to review: 000/xyz123)	Using an online questionnaire system
URI or Ente Doce	PID s the dataset have a persistent identifier (PID)? s o or 'persistent identifier' provides a permanent citable reference to a certain dataset, this is always a fixed reference to the	Prototype: <u>https://www.surveymonkey.com/r/fairdat</u>
datas Nation Click	User license User license Does the dataset have a user license? O Yes	
	No A user licence is a mechanism that explains the extent to which people and organisations have permission to reuse the dataset and other material which is protected by copyright or database right. For example, CreativeCommons (CC0) license describes that there are no restrictions for access so is completely open. The access rights should therefore be clearly specified or described. Click here for additional guidance on how to answer this question. Any remarks about scoring the dataset at this level:           Prev         Next	Score of F4 You scored 4 stars for Findable. 



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

700000		
ZG IUUU Search Q Upload Cd	Og in     Sign up	
Recent uploads	Sep 12: Major update	
	Welcome to the improved	
FIGURE 5 in Molecular and biogenuatic differentiation of Reaphie	View Zenodo. See what's new and	
occidentalis with description of a new treefrog from north-western	known issues.	nort Sign Lin Log In
Madagascar		port oigin op bog in
Vences, Miguel; Andreone, Franco; Glos, Julian; Glaw, Frank		
FIGURE 5. Photographs of Boophis occidentalis from Isalo National Park in life. (a – b) Male specimen 7SM 2314 / 2007 in dorsolateral and ventral view and (c) same specimen at its c	Harvard Dataverse A collaboration w In National Diversity IT, and IQSS	
position in a small cavity in a rock above a stream, photographed on 15 February 2007; (d) h	olotype	
(ZFMK	Il Metrics 2,060,108 Downloads	2 C
Uploaded on December 8, 2016.		
	Share, publish, and archive your data. Find and cite data across all research fields.	
December 6, 2016 Dataset Open Access	View	
Revisiting the phylogeny of phylum Ctenophora: a molecular	POOD POUCY Hearry A. Murray POOD POUCY Research Archive	
	HOME REGISTER LOG IN FIGURATION CONTINUE IFPRI	th Archive
Get exposure and credit for your data; write a data paper for the new near reviewed online-only open as	re - Population Services International International Food Policy Dataverse Data lournal (oublished by Brill) (PSI) Dataverse Research Institute (IFPRI)	se
Cor more info will som /edi	Dataverse	
For more into: brill.com/raj		
EASY offers sustainable archiving of research data and access to thousands of datasets.	Q Find Advanced Search	+ Add Data
Search	SEARCH > Search help	
> Advanced search > Browse	1 to 10 of 65,724 Results	It Sort-
	Internet Banking Espousal in Bangladesh: A Probing Study	
	Dec 11, 2016 - Ahmed Research Archive Dataverse	-
32,530 RESULTS IN PUBLISHED DATASETS	Alim Al Ayub Anmed, Md. NuF-E-Aliam Siddique, 2016, "Internet Banking Espousal in Bangladesh: A Probled doi:10.7910/DVN/G4NAH8, Harvard Dataverse, V1	ng Study",
≡ List <sup>(1)</sup> Map	Choose One  Choose One	rvices scenery in budding
Archaologisch booronderzeek verdubbeling N291 Donkerbroek Gesterwelde gemeente	nations such as Bangladesh. Nevertheless, due to the connected near to the ground acceptance rate, its full potential	in deepening and
Ooststellingwerf (FR)	Archival Data for Consider the Redirect: A Missing Dimension of Wikipedia Research	
Date: 2019-06-09 Audience: Archaeology Creators: Krol-Karsten, T.N. (MUG Ingenieursbureau) Access: Open (registered users)	Search SEARCH Hill, Benjamin Mako; Aaron Shaw, 2016, "Archival Data for Consider the Redirect: A Missing Dimension of	Wikipedia Research*,
Submitted: 2016-07-11	Advanced search doi:10.7910/DVN/NQSHQD, Harvard Dataverse, V1	
> Thematic Collection: Children of Immigrants Longitudinal Survey in the Netherlands (CILSNL)	This contains data and software for the following paper: Hill, Benjamin Mako and Aaron Shaw. "Consider the Redirect Wikipedia Research." In Proceedings of the 10th International Symposium on Open Collaboration (OpenSym 2014). A	A Missing Dimension of CM Press, 2014. This is an
Date: 2017-12-31 Audience: Social sciences Creators: Jaspers, dr. E. (Universiteit Utrecht); Tubergen, prof. dr. F. Sociology	Audience	
van (Universiteit Utrecht) Benavioural and aducational sciences sciences Political science	Behavioural and educational sciences (123)	
Criminal (procedural) law and criminology	Economics and Business Administration (22)	
Migration, ethnic relations and	numanities (31630)	

# 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?

Don				
k	Extension	File type	N of files	% of total
1	ipg	image/jpeg	2.633.879	64,5%
2	pdf	application/pdf	164.630	4,0%
3	stif	image/tiff	137.408	3,4%
4	csv	data/plain	129.762	3,2%
5	itxt	text/plain	122.136	3,0%
6	mid	gis/mid	57.415	1,4%
7	'mif	gis/mif	57.029	1,4%
8	Scmdi	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 oth	er formats/extensions	331.708	8,1%
	All		4.085.668	100.0%



# Towards a FAIR Framework?





# Thank you for listening!



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

Data Archiving and Networked Services



Thanks to Ingrid Dillo and Emily Thomas for their contributions

