"Eerlijk is vals"

FAIR Principles in de praktijk

Dr. Patrick J.C. Aerts

Publieksversie

Driven by data

Accelerating discovery

Achtergrond van de FAIR Principles

Prof. dr. Barend Mons

Heeft al sinds de 90-er jaren wetenschappelijke belangstelling voor het onderwerp Malaria. Publiceerde daar veel over, maar had ook al vroeg zorg over de toegankelijkheid van de moderne literatuur voor de mensen die het het meest nodig hadden: onderzoekers, artsen en anderen in Afrika en andere landen waar Malaria heerst. Mijn persoonlijke mening: dit zal een belangrijke drijfveer geweest zijn om aandacht te vragen voor en te handelen ter bevordering van de toegankelijkheid van wetenschappelijke literatuur en data: open access, open data, etc.
De *principles* voor de omgang met data vormen dan ook vast niet per ongeluk het woord FAIR!

FAIR Principles geboren

- The Value of Data
  - *Nature Genetics* **volume 43**, pages 281–283 (2011)
  - http://dx.doi.org/10.1038/ng0411-281
  - 16 auteurs, Barend Mons (contactpunt)

- The FAIR Guiding Principles for scientific data management and stewardship
  - Scientific Data **volume 3,** 160018 (2016) (A Nature publication)
  - http://dx.doi.org/10.1038/sdata.2016.18
  - 53 auteurs, Barend Mons (contactpunt)

- Merk op, hoeveel auteurs en hoe lang geleden!

Meanwhile in Europe

- In September 2015 Barend Mons was appointed chair of the High Level Expert Group on the European Open Science Cloud
- In August 2016 Horizon 2020 Commission expert group on Turning FAIR data into reality (E03464)
  1) Map best practices to turn FAIR components into reality
  2) Propose indicators to measure progress on each of the FAIR components
  3) Give input to a proposed action plan on how to make data FAIR
  4) Contribute to the evaluation of the Horizon 2020 DMP template and future revisions of the template, in light of harmonisation with funders across the EU, including the development of additional sector/ discipline specific guidance (if desired)
  5) Provide input into the issue of costing and financing of data management activities (on EU/ Member State/ international level).
- June 2018

- July 2 2018: Metadata en interoperabiliteit: Hoe dan?

netherlands eScience center

DANS

De FAIR Data Principles

- De FAIR Data Principles zijn een *guidence*, geen voorschrift, laat staan een blauwdruk voor hoe ze geïmplementeerd moeten worden
- Over de FAIR Data Principles is goed nagedacht. Het feit dat het *principes* zijn en geen voorschriften is niet zonder opzet.
- Dus daarom goed voor ogen houden *waarom* er FAIR Data Principes zijn:

Primair

- Versnelling wetenschappelijk vooruitgang
- Geen dubbel werk verrichten
- Communicatie optimaliseren over de wetenschappelijke stand van zaken
- Toegankelijkheid met minimale barrières
- “Level playing field”
- Niet alleen de netto publicatie, maar ook de data zelf delen
- Als het kan, meegaan in de trend van automatische (gerobotiseerde) zoek en analyse-processen (AI)

Secundair

- Bijdragen aan accountabililty
- Tegengaan van malversaties
- Stroomlijnen eigen onderzoeksproces (workflow)
- Bijdragen aan een vernieuwde zienswijze van onderzoekers op hun bijdrage aan de wetenschap: “delen helpt”
- Bijdragen aan het verkleinen van de “digital divide” (ontwikkelingslanden)

Top Down

- Top Down requirements komen van
  - De Europese commissie
  - H2020 en vergelijkbaar
  - Research Councils/Funding Agencies
  - Universitaire/Institutes besturen/directies
- Hun belang bestaat uit:
  - Verantwoording kunnen afleggen over (aan projecten) besteed geld
  - Een show case hebben van mooie –financieel ondersteunde- projecten
  - Tonen van het belang van hun functie: bevorderen wetenschap en economie
  - Eenvoudige administratieve afdoening
  - Dus stroomlijning van processen

Bottom up

- Bottom up requirements komen van
  - De research communities
  - Vanuit verschillende niveaus:
    - zonodig per subdiscipline
    - Zonodig per data type
  - Deskundige wetenschappers, die invloed hebben binnen hun domain
  - Onderzoekers die een voorbeeld stellen
- Hun belang ligt in:
  - Versnelling van het onderzoeksproces
  - Meer gewaardeerde output (dan alleen p.r. publikaties)
  - Meer erkenning
  - Intensievere samenwerking
  - En meer…

In between

- En daartussen zitten de “executive organisations”
  - De partijen die diensten aanbieden op het gebied van data-opslag, data services
  - De partijen die (helpen) data van meerwaarde (te) voorzien (zoals DANS)
  - Partijen die expertise bezitten
    - Op het gebied van metadata, thesauri, indelingssystematieken, “knowledge eperts”
    - Bibliotheken van instellingen
    - Reken of data centra van de instellingen
    - Data stewards, zittende en nieuwe
  - Wij, hier dus

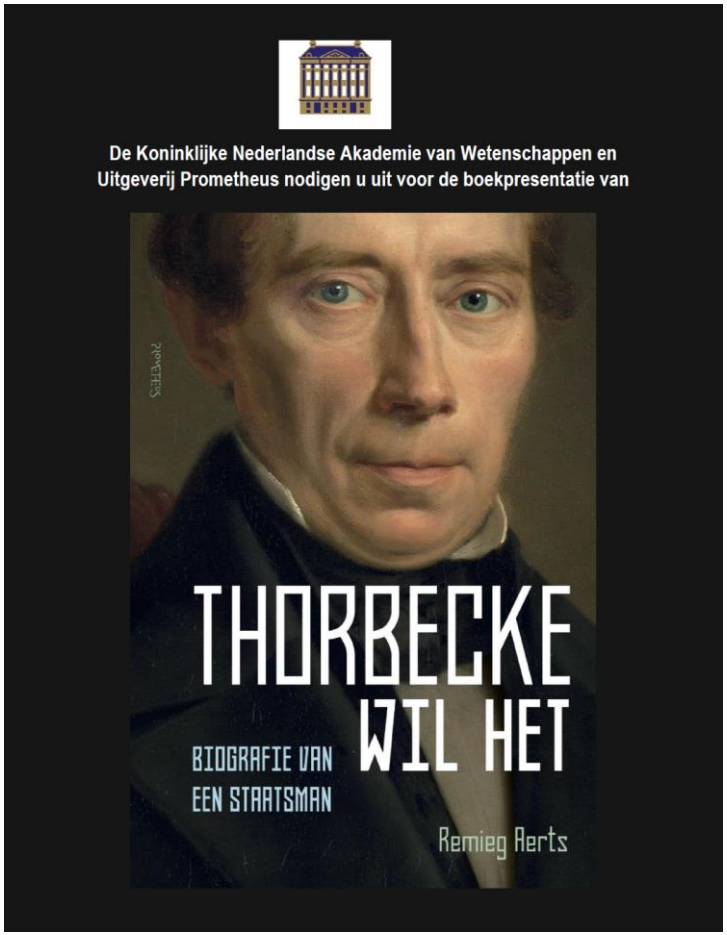
Experts (zijn wij dat?)

- Eerlijk is vals duidt vooral daarop, dat FAIR-ness een prachtig streven is (“eerlijk zullen we alles delen”), maar de materie evenzo weerbarstig:
  - De F is wel te doen: goede afspraken rond metadata, een goede infrastructuur, eventueel zoekmachines,…
  - De A is al wat lastiger: kun je bij de data komen, mag je ze wel gebruiken, is het gratis of betaald, is het private of openbaar, of iets er tussen, …
  - De I is het beest dat getemd moet worden. Wat is “interoperability” van iets dat zelf niet “opereert”. Het is (doorgaans) geen software, het “werkt” niet ergens op, het “is” wat het is.
  - Maar zonder I zijn data FAR en je wilt ze NEAR, dus verzin een list of wat.
  - De R is ook nog volop in discussie, maar als de F,A en I gedefinieerd zijn, zou de R dan nog een bijzonder probleem zijn?

De I van FAIR

- De “officiële” onderverdeling leest als volgt:
  - I1. Proprietary, non-open format data
  - I2. Proprietary format, accepted by Certified Trustworthy Data Repository
  - I3. Non-proprietary, open format (= “archival format”)
  - I4. Data is additionally harmonized/standardized, using a standard vocabulary (for the research field to which the data pertain)
  - I5. Data is additionally linked to other data to provide context
- En hier blijkt ook al de weerbarstigheid van FAIR:
  - Als hieraan voldaan is, waarom zijn data dan ineens interoperable?
  - PDF has become basically open, (ISO standardized), but not all aspects are equally open, data in PDF cannot always be copy-pasted, etc. This shows that even for PDF, I2 and I3 may not be sufficient
  - I5 is related to provenance, that could as well have been taken care of under the F or the A
  - Some communities consider the I as to guarantee data to be machine readable ánd machine-interpretable (in the sense of usability for one’s purpose)
  - What is the interoperability of a book?

Intermezzo



- (Zou NWO zoiets als projectvoorstel gehonoreerd hebben?)
- 7-9 jaar intensieve research
- Veel onderzoek in bibliotheken, archieven, sommige zaken digitaal, veldonderzoek
- De uitkomst is een belangwekkende publicatie (1e druk in twee weken uitverkocht)
- De uitkomst is toegankelijk (want een boek)
- Wat als je een research data managementplan had moeten schrijven, wat betekent FAIR in zo'n geval?

Ergo: blijf denken over functionaliteit van voorschriften en beleid en het hoe en waarom van (FAIR) principes

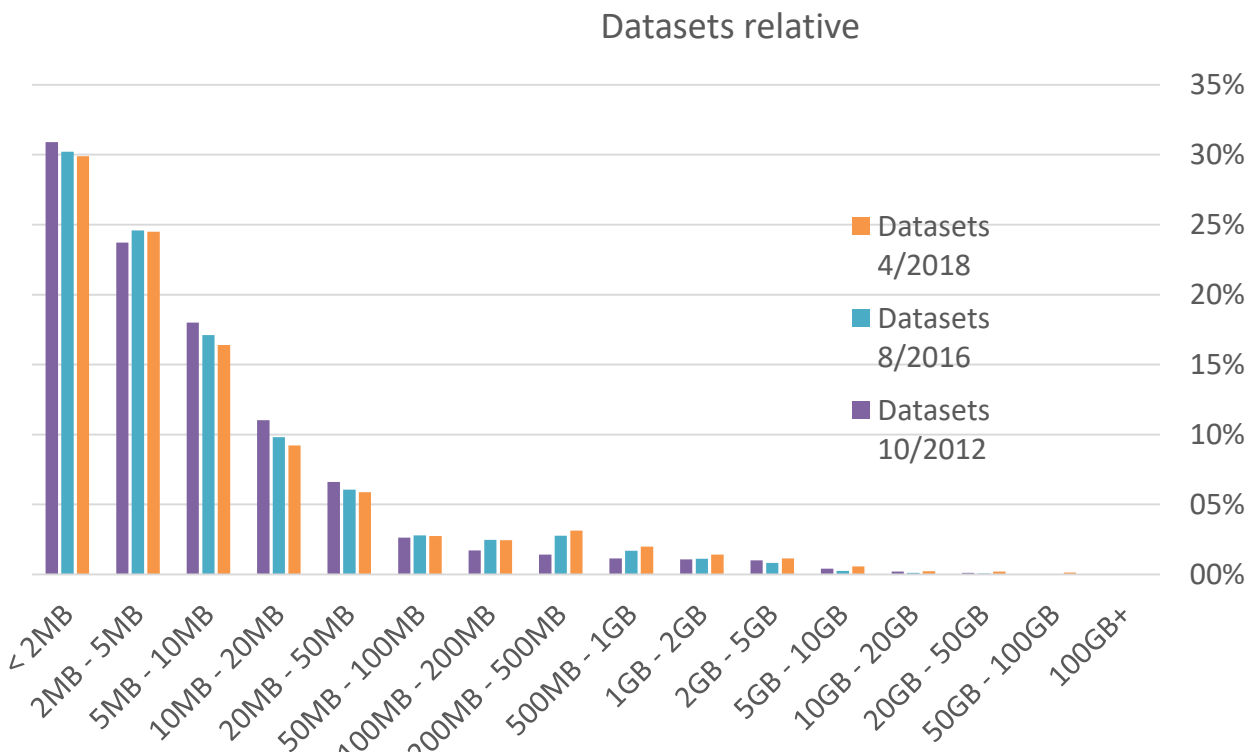
netherlands eScience center

DANS

The message is…

- Denk aan de “long tail of science”
- De vele vakgebieden waar “data” niet de boventoon voeren bij het onderzoek of het onderwerp
- Die evenzeer aandacht behoeven, en waarschijnlijk meer arbeidsintensiviteit vragen als het gaat om wat de “I” daar zou kunnen betekenen
- Accepteer dat het ook gewoon een non-issue kan zijn

Datasets in DANS EASY archive according to size

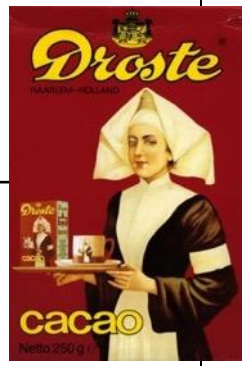


The long tail of research data

	Datasets 10/2012	Datasets 8/2016	Datasets 4/2018
> 1 Gb	2,8%	2,5%	3,8%
> 2Gb	1,8%	1,3%	2,3%

Operationalizing the I in FAIR

FAIR Interoperable Principle	Remarks
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.	Understandable both for humans and for computers
I2. (meta)data use vocabularies that follow FAIR principles.	Droste effect: circularity How to interpret “use vocabularies”?
I3. (meta)data include qualified references to other (meta)data.	Goal is to create as many meaningful linkages as possible between (meta)data resources to enrich the contextual knowledge about the data. Scientific links between datasets need to be described



For FAIR Help see: https://www.dtls.nl/fair-data/fair-principles-explained/

DANS operationalisation of I

FAIR 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*.

Tim-Berners Lee 5-star scheme

★	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]

DANS FAIR "Light" Interoperable

1. Proprietary (privately owned), non-open format data
2. (Part of) the files in a multi-file dataset are in a proprietary format
3. Non-proprietary, open format = 'preferred format' in Trusted Repository
4. As well as in a preferred format, data is standardised using a standard vocabulary (for the research field to which the data pertain)
5. Data additionally linked to other data to provide context

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

How to interpret interoperability 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?

?

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%

FAIR Internationaal

- Bedenk dat WIJ de pioniers en voorlopers zijn in Europa (en de rest van de wereld)
- Met onze gedachten over de operationalisering van de elementen uit FAIR geven wij een voorbeeld voor de rest van Europa en “de beweging”
- In meerdere landen van Europa is men nog huiverig om FAIR-ness op te leggen als minimum voor open data
- Dus werk als gids, maar niet overhaast

My thoughts

- Ik begrijp waarom in sommige domeinen de I van FAIR een bijzondere betekenis heeft
- Ik begrijp waarom full computer searchable and evaluatable wenselijk kan zijn
- Dus bepaal wat “I” is, vanuit de behoefte van het domein
- Vergelijk met de Digital Academic Repositories, waar binnen het concept van de Dublin Core nadere afspraken werden gemaakt om de uitwisselbaarheid en harvesting voor een specifiek doel beter mogelijk te maken. Ook zonder die afspraken was het officiële format nog stees Dublin Core
- Bedenk dat als aan de F en A voldoende aandacht is besteed, investeren in AI (Artificial Intelligence) een betere, duurzamere optie kan zijn
- Behoud je relativeringsvermogen, ook bij beleid maken

And finally…

- Vergeet de software niet!!
- Sustainability, etc.
- FAIR for software
  - Ook daar hebben F, A, I, R betekenis, maar lichtelijk aangepast
- Van 29 oktober-1 november 2018 wordt in Amsterdam de IEEE eScience Conferentie gehouden. Op 29 Oktober daarin organiseert PLAN-E de workshop eScience is FAIR Science
- DANS en NLeSC werken aan een uniek voorschrift voor het onderbrengen van software ontwikkeld in het academisch domein

The End

netherlands eScience center

DANS