Bring Your Own Data (BYOD) workshop for Data Science

Kristina Hettne, Ben Companjen, Peter Verhaar

18 June 2019

Universiteit Leiden
The Netherlands

Learning objectives

- Basic knowledge of the FAIRification process
- Basic knowledge of conceptual data modelling
- Basic knowledge of semantic data modelling
- Basic skills in OpenRefine with RDF and Wikidata plugin
- Basic knowledge of data publication

Schedule

Time	Subject	Details	Lead by
09.00 -09.15 h	**Introduction**	• The FAIR principles • The FAIRification process	Kristina Hettne
09.15 -10.00 h	**You and your data**	• Round of introductions about you and your data	Kristina Hettne
10.00 -10.30 h	**Tutorial - part 1**	• What is semantic modelling? • How to create a conceptual semantic model	Kristina Hettne/ Peter Verhaar /Ben Companjen
10.30 -10.45 h	Break		

10.45 -12.15 h	**Tutorial - part 2**	• How to find ontologies using online sources
		• How to populate the semantic model
		• How to use the FAIRifier to wrangle the data
		  ○ Assign ontologies
		  ○ Export RDF
			Kristina Hettne/ Peter Verhaar/ Ben Companjen

12.15 -12.45 h	Break		

12.45-14.45	**Hands-on**	• Hands-on FAIRification training with participants own data:
		  ○ Create a semantic model
		  ○ Work with the FAIRifier
			Kristina Hettne/Peter Verhaar/Ben Companjen

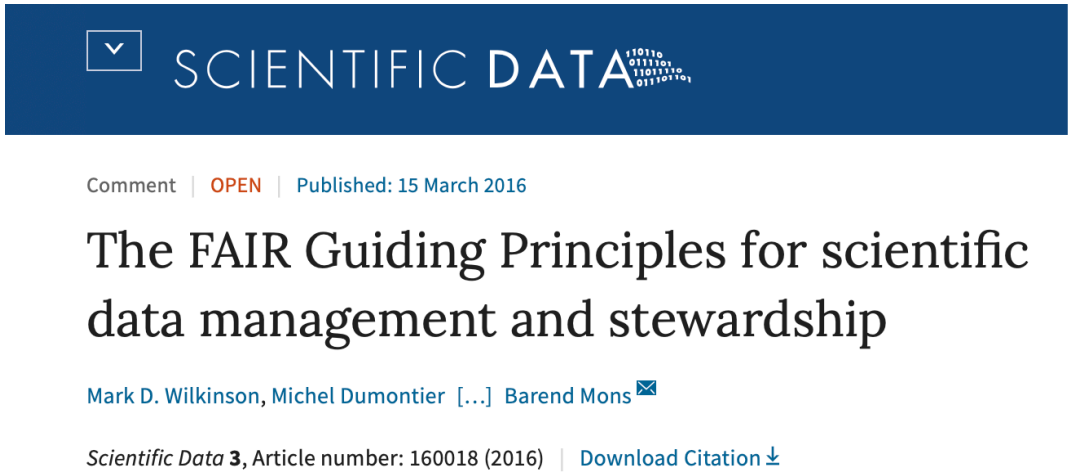
14.45 -15.30 h	Break + discussion	Room for plenary discussion about the results	
15.30-16.30 h	**Share and publish FAIR data**	• Data citation • Data repositories • FAIR Data Point	Kristina Hettne
16.30-17:00 h	**What did we learn?**	Questions and feedback on training	Kristina Hettne

Introduction

Universiteit
Leiden
The Netherlands

Discover the world at Leiden University

Findable, Accessible, Interoperable and Reusable (FAIR)



SCIENTIFIC DATA

Comment | OPEN | Published: 15 March 2016

The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson, Michel Dumontier [...] Barend Mons

Scientific Data **3**, Article number: 160018 (2016) | Download Citation

Research data needs to:

- Be accessible under clear conditions and licenses
- With clear references
- With rich metadata

Privacy-sensitive data can meet the FAIR principles

https://doi.org/10.1038/sdata.2016.18

Findable

The first step in (re)using data is to find them. Metadata and data should be easy to find for both humans and computers. Machine-readable metadata are essential for automatic discovery of datasets and services, so this is an essential component of the **FAIRification process**.

F1. (Meta)data are assigned a globally unique and persistent identifier

F2. Data are described with rich metadata (defined by R1 below)

F3. Metadata clearly and explicitly include the identifier of the data they describe

F4. (Meta)data are registered or indexed in a searchable resource

https://www.go-fair.org/fair-principles/

Not dealt with in this BYOD

Accessible

Once the user finds the required data, she/he needs to know how can they be accessed, possibly including authentication and authorisation.

A1. (Meta)data are retrievable by their identifier using a standardised communications protocol

A1.1 The protocol is open, free, and universally implementable

A1.2 The protocol allows for an authentication and authorisation procedure, where necessary

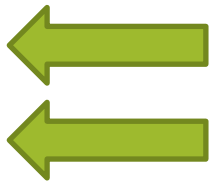
A2. Metadata are accessible, even when the data are no longer available

https://www.go-fair.org/fair-principles/

Interoperable

The data usually need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing.

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

https://www.go-fair.org/fair-principles/

Reusable

The ultimate goal of FAIR is to optimise the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings.

R1. Meta(data) are richly described with a plurality of accurate and relevant attributes

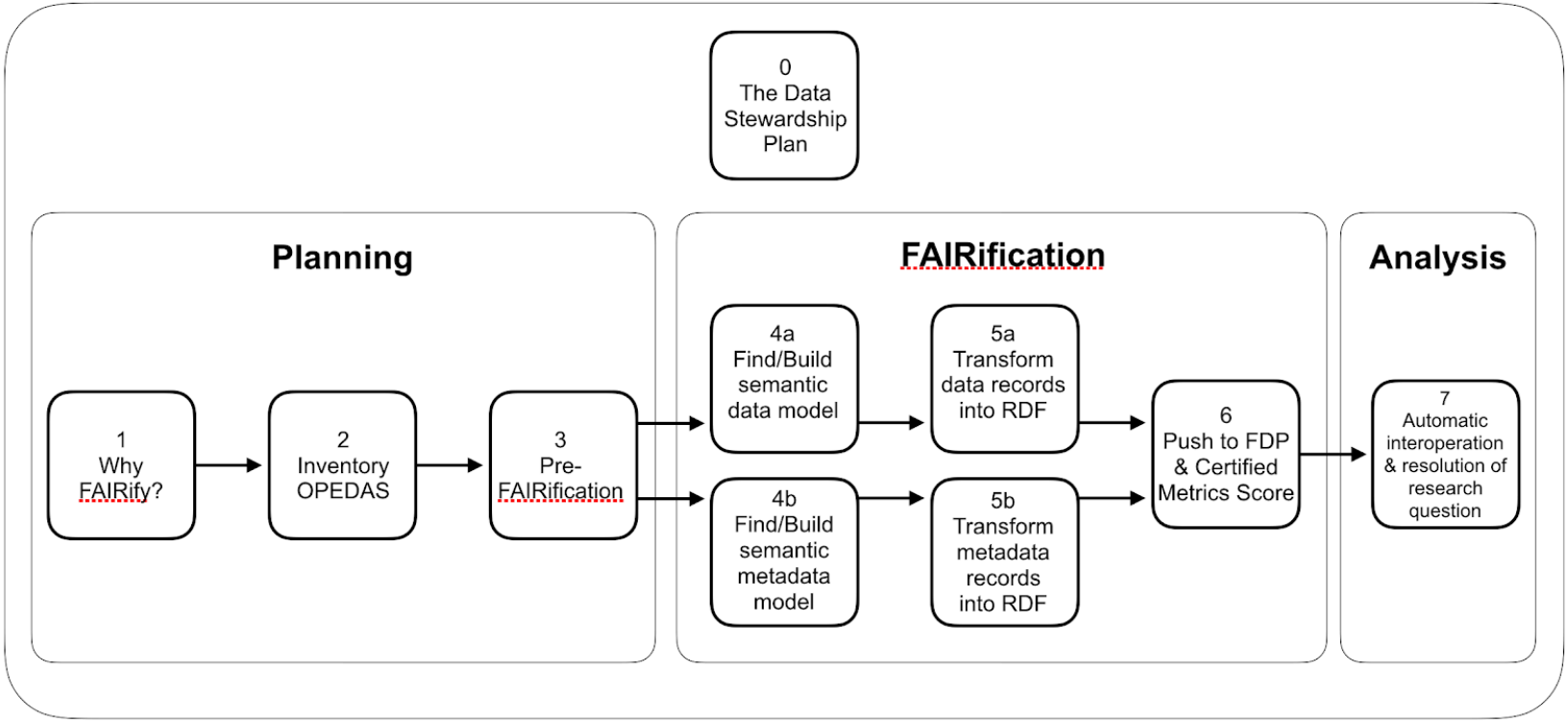
R1.1. (Meta)data are released with a clear and accessible data usage license

R1.2. (Meta)data are associated with detailed provenance

R1.3. (Meta)data meet domain-relevant community standards

https://www.go-fair.org/fair-principles/

The FAIRification process

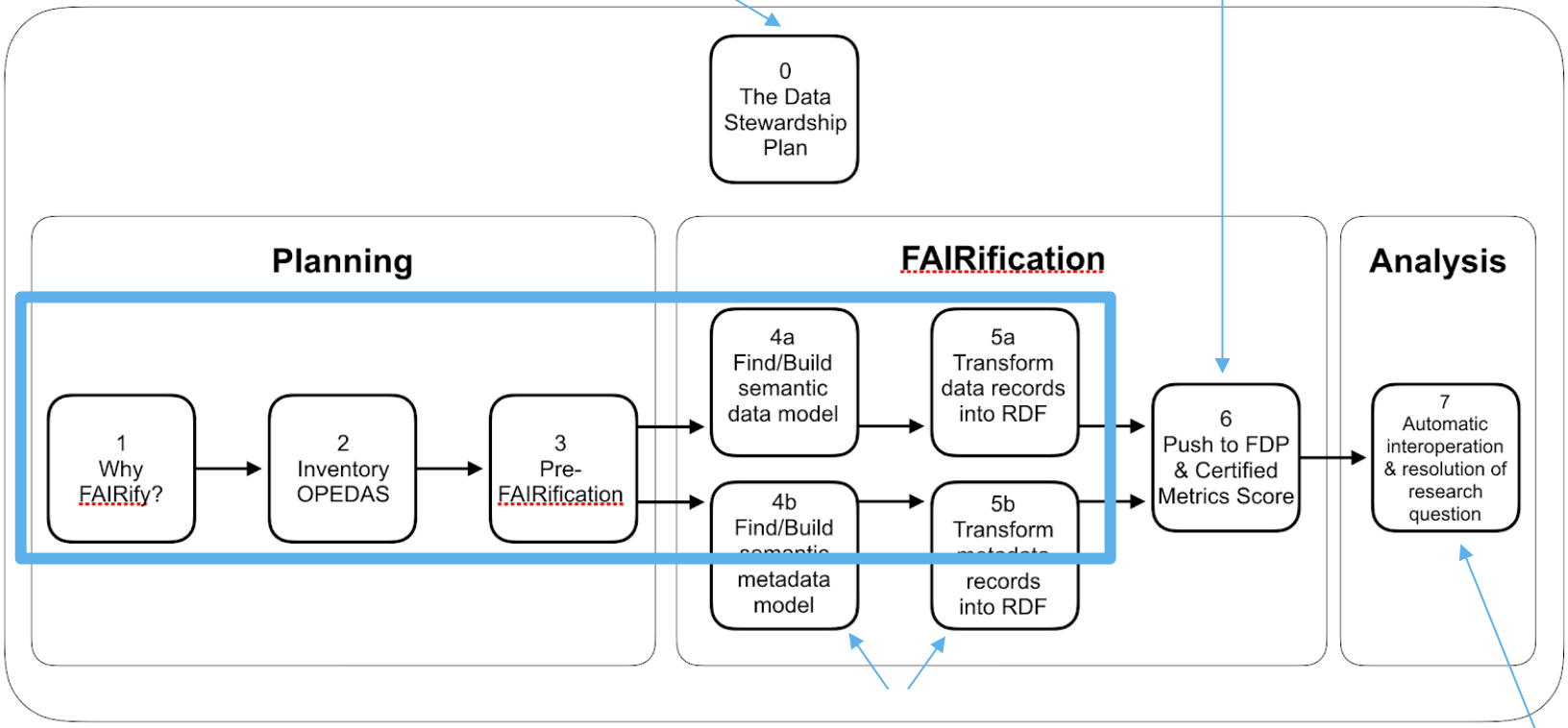


OPEDAS = Other People's Existing Data and Services

The FAIRification process

Covered in DMP workshop

Partly covered. Data publishing workshop (to come end 2019)



Covered in data publishing workshop (to come end 2019)

OPEDAS =
Other People's Existing Data and Services

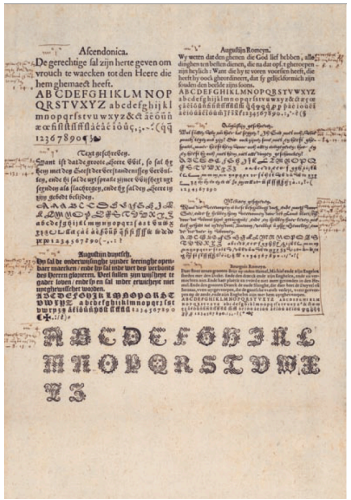
Consultancy/Separate workshop (to come 2020)

1. Why FAIRify?

- I want my data to be more FAIR, because I need to satisfy some requirements (show compliance)
- I want my data to be more FAIR, because I need citations (show impact)
- I want my data to be more interoperable (driving user questions – interoperable with what?)
- I want my data to be more reusable, e.g. by defining access/use conditions in metadata (show reusability)
- I want other people to be able to find my data (focus on metadata)
- I need more users for my software/framework/repository, otherwise it will be dead by the end of the year
- I want my data to be FAIR (Reusable) because I am the first re-user of my own data, notably when (a) I need to reproduce results tomorrow (b) I need to answer questions from referees of my paper (c) my postdoc leaves and the work is picked up by his/her successor

Use case: book trade history in Leiden

- Scholarly archive of book historian Prof. dr. Paul Hoftijzer
- Archive contains descriptions of people, organisations and events relevant to Leiden book trade in the early modern period
- Why FAIRify?
  - Teach Digital Humanities techniques
  - Open up for possible reuse



Aa, Cornelis van der (* 1749?; † ?; w. 1767-?)
Boekverkoper.
Geen lid van de grote boekverkopersfamilie Van der Aa.
Gilde: Pre-1765 L bij Johannes Lemair; 5-8-1765 bij Jacobus van der Spijck voor 4 jaar. 2-10-1767 Vrijmeester (AB 83a, f. 34v).
1796 te Haarlem, 1816 te Amsterdam.
GAL, prentverzameling 46601 portret Cornelis van der Aa, boekverkoper, geb. Leiden 1749, gegraveerd door Reinier Vinkeles naar C. van Geulen.
Veilingen: 11-11-1783 Veiling, met Vincent van der Vinne, van de collecties van Cornelius Asconius van Sypesteyn en C. en G. Schertzer, waaronder ook kunstvoorwerpen.
Lit.: Ledeboer.

Aa, Hillebrand van der (* 1661 (doop 22-3); † ± 1721; w. 1697-17?)
Plaatsnijder en beeldhouwer (bij zijn eerste huwelijk). Luthers. Zoon van Boudewijn Pietersz van der Aa en Annetje Poortemuller, broer van Pieter van der Aa. Getuigen bij zijn doop Roocks Immerseel, Tönjes Lockers en Elsche Heinrichsen.
Adres: 1683 Nieuwsteeg; 1685 Salomonssteeg (ouderlijk huis); 1696-99 Rapenburg?; 1701 Kloksteeg.
Huwelijk: 1. SH 28-4-1683 Maria Badde uit Haarlem (getuige zijn broer Pieter; † 29-1-1684, begr. PK); 2. SH 23-6-1684 Catharina Oesinger (Pieter van der Aa in de Nieuwsteeg zijn getuige; † 29-11-1749; zij werd begraven buiten Leiden; haar adres is Haarlemmerstraat bij de Turfmarkt); kinderen: Maria en Balduinus, de laatste werd predikant in de Leidse Lutherse gemeente.

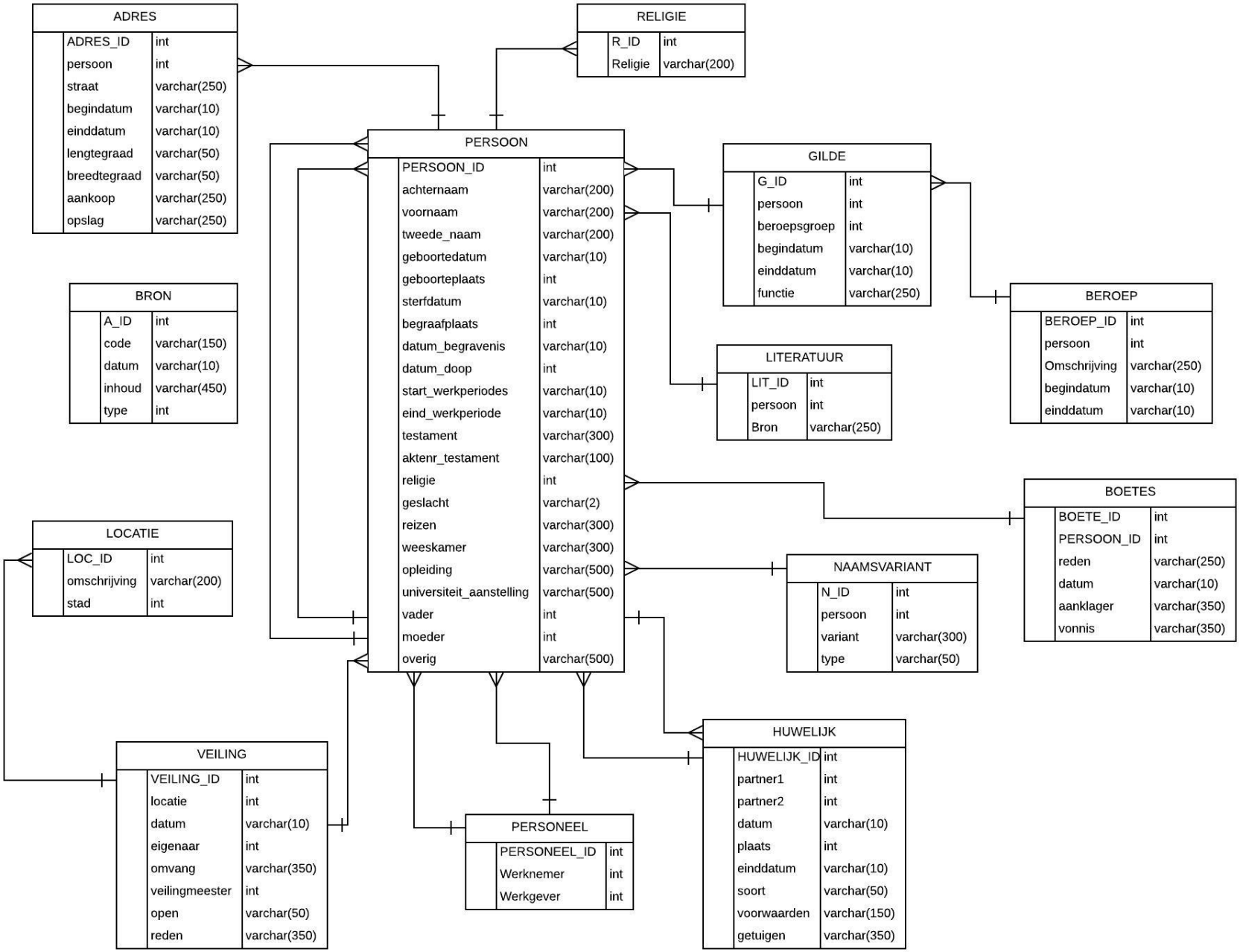
2. Inventory OPEDAS

- Metadata standards
- Data standards
- Licenses
- Data management plan
- Storage/hosting/Repository

OPEDAS = Other People's Existing Data and Services

3. Pre-FAIRification

- Create a conceptual model (data model)
  - Define relationships between the data elements while leaving out implementation details such as specific ontologies



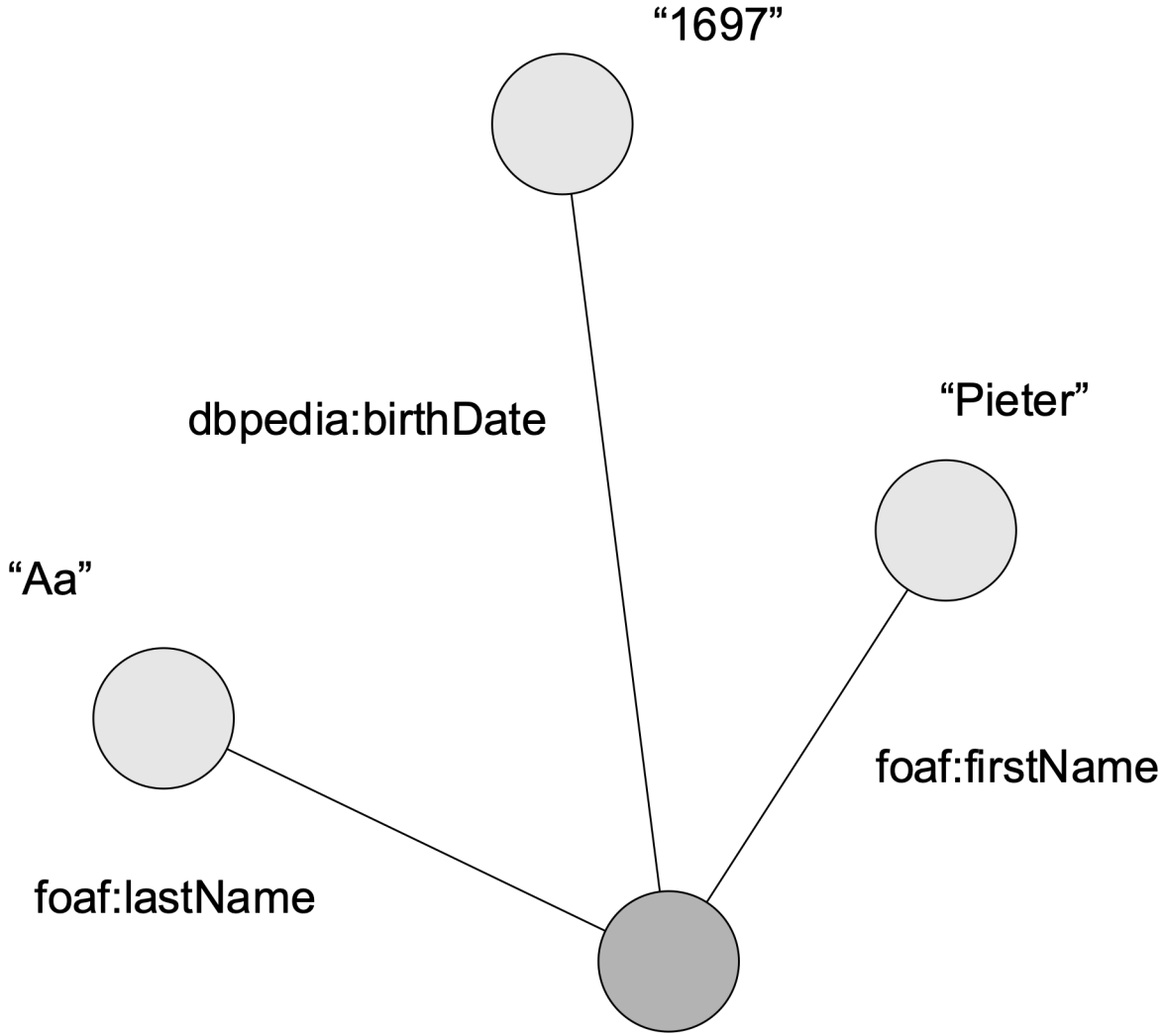
3. Pre-FAIRification

- Organize/Structure data before it is rendered machine-readable

Aa, Pieter Jansz van der (* Leiden 1697; † 2-8-1751 [begr. PK 31-7/7-8-1751]; w. 1719-36)

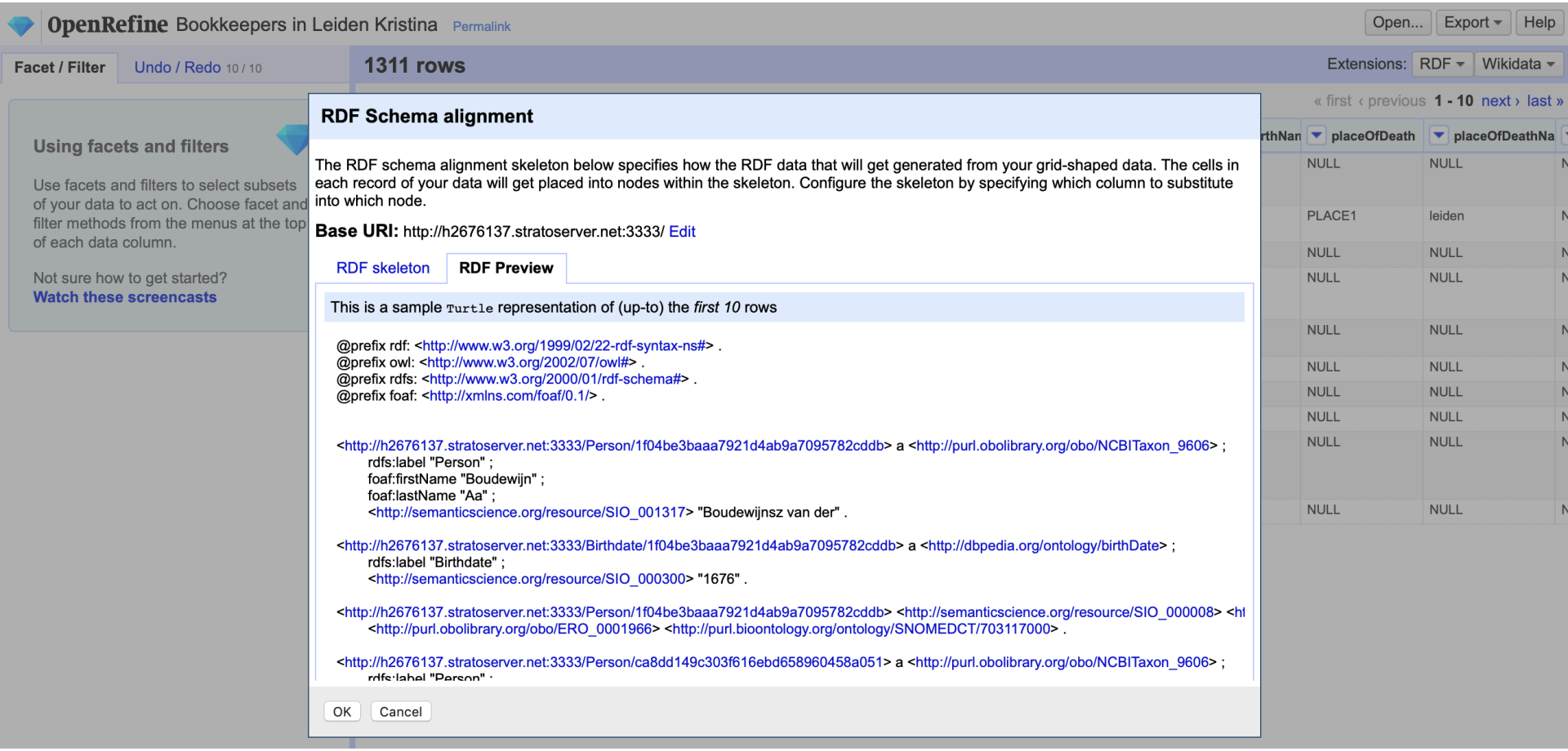
firstName	lastName	secondName	dateOfBirth	dateOfDeath	placeOfBirth
Pieter	Aa	Jansz van der	1697	1751-08-02	leiden
Boudewijn	Aa	Jansz van der	1692	NULL	leiden
Cornelis	Aa	van der	1749	NULL	NULL
Hillebrand	Aa	van der	1661	1721	NULL

4a. Define the semantic data model



<https://nonsolus.leidenuniv.nl/person/ABB1>

5a. Transform data records to RDF



You and your data

Universiteit
Leiden
The Netherlands

Discover the world at Leiden University

You and your data

- What is your research question?
- Did you fill out a Data Management Plan?
- In what format is your data right now?

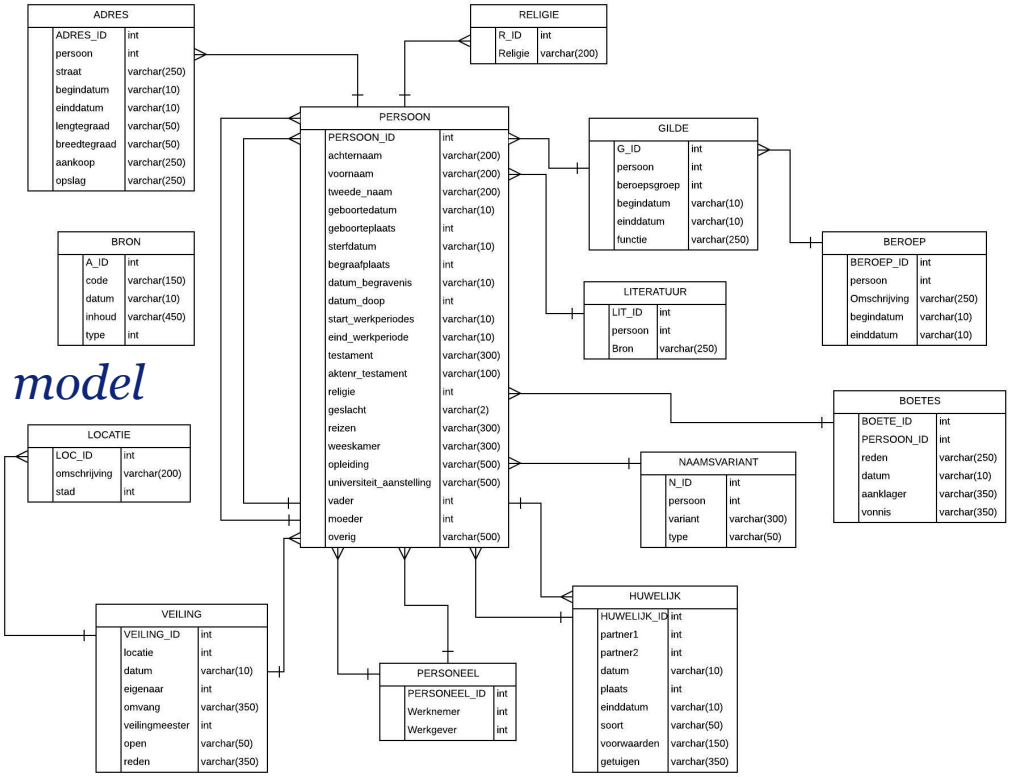
Tutorial - part 1

Universiteit
Leiden
The Netherlands

What is a data model?

- "A good data model requires you to be an expert in your field of study, but also to think clearly like a logician or philosopher." – Erik Schultes and Rajaram Kaliaperumal
- Central to data interoperability
- Shows relations between entities

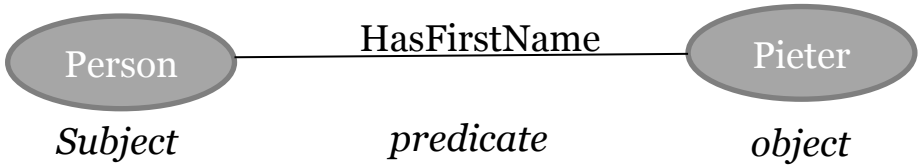
Example: Peter's book trade history model



Resource Description Framework

The Resource Description Framework (RDF) is a very generic model for describing things.

- Statement = subject, predicate, object (a.k.a. triple)



- Use Uniform Resource Identifiers (URIs) as globally unique identifiers for instances
  - Then computers can match them across datasets!

- Use URIs for properties and classes too
  - Data models can themselves be described in RDF using RDF Schema and OWL
  - Then computers can reason about the meaning of data beyond your dataset

- Reuse URIs and data models when possible

- How to go from tabular data to RDF: plugin in OpenRefine
  - Other options available: e.g. CSV on the Web metadata, RDF Mapping Language

How to create a data model

- First try to answer these questions: What is your research question? Why did you collect this data?
- With your research question in mind, look at your data: which are the main entities?
  - If you have a spreadsheet, what do your rows represent? This is probably one of your central entities in your data model.
  - Are there other main “Entities” in your data?
- When you have figured out your entities: what are the relationships between them?
  - Do not be too specific in this stage, it only hampers the creative process

- Tools: GRAFO (https://gra.fo), pen and paper, PowerPoint or similar…

Exercise: model this spreadsheet in triple format

id	firstName	lastName	secondName	placeOfBirthName	gender
ABB1	Boudewijn	Aa	Boudewijnsz van der	Leiden	M
ABJ2	Boudewijn	Aa	Jansz van der	Leiden	M
ACV3	Cornelis	Aa	van der	NULL	M

Tutorial - part 2

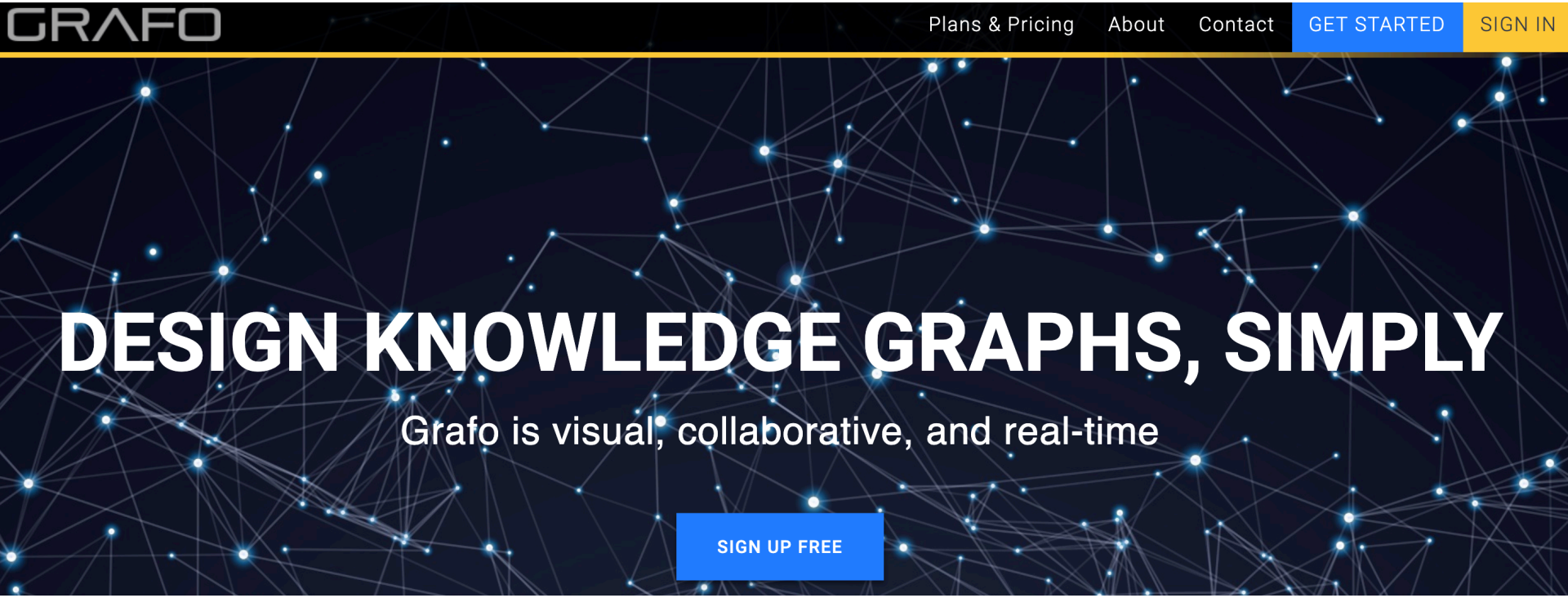
Universiteit
Leiden
The Netherlands

How to find ontologies using online resources

- Collections of resources
  - FAIRsharing: https://fairsharing.org/standards/
  - Linked Open Vocabularies: https://lov.linkeddata.es/dataset/lov/
- Databases of entities
  - WikiData: https://www.wikidata.org
  - DBPedia: https://wiki.dbpedia.org
  - WikiPedia: https://www.wikipedia.org
  - Schema.org: https://schema.org
- Relations
  - Semantic Science ontology: http://sio.semanticscience.org (can be searched through Linked Open Vocabularies search engine above)
  - Relation Ontology (see OBO Foundry link below)
- Your own domain
  - Example Life Sciences: BioPortal http://bioportal.bioontology.org, OBO Foundry: http://www.obofoundry.org

How to populate the data model – demo GRAFO

https://gra.fo/

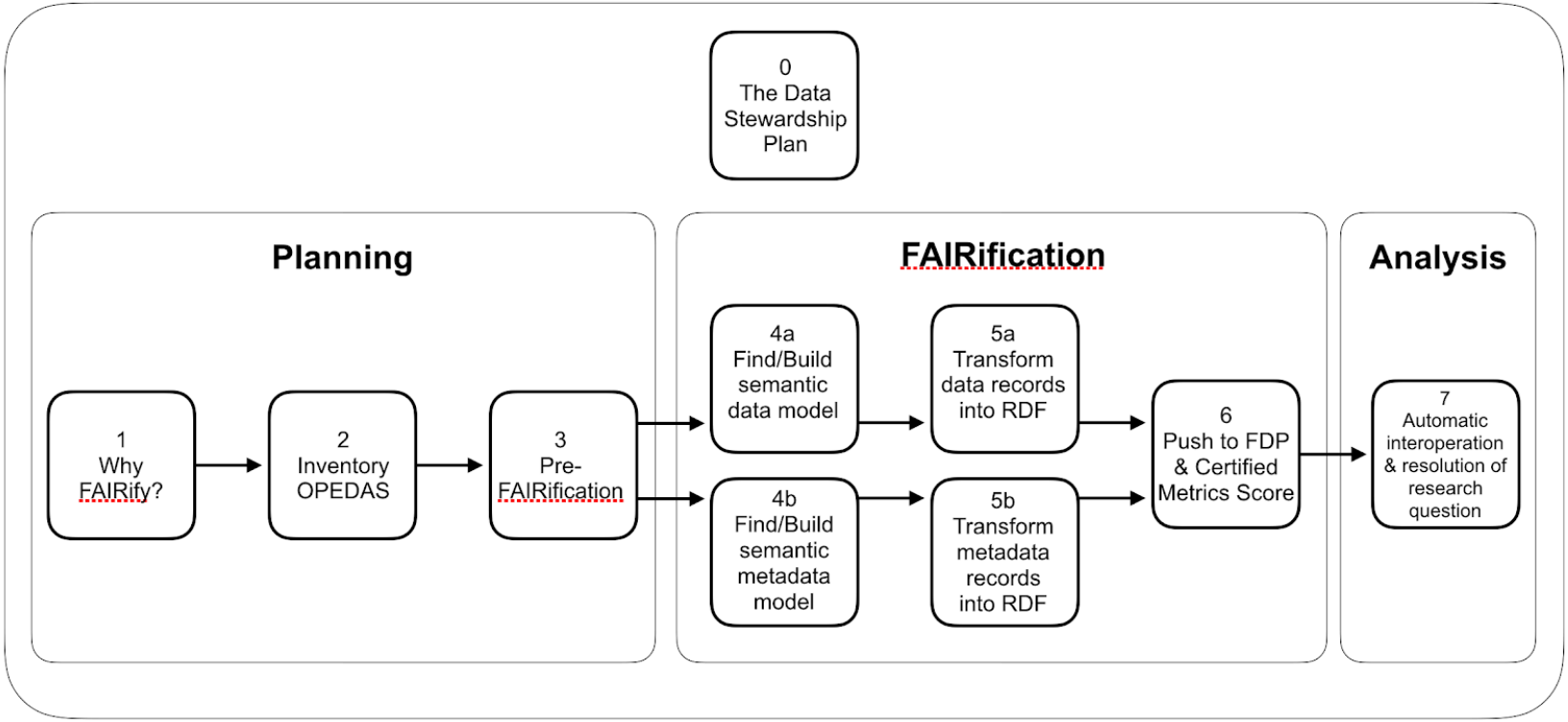


How to use the FAIRifier/OpenRefine to wrangle the data – tutorial

Hands-on

Universiteit
Leiden
The Netherlands

The FAIRification process



OPEDAS = Other People's Existing Data and Services

Share and publish FAIR data

Universiteit Leiden
The Netherlands

Discover the world at Leiden University

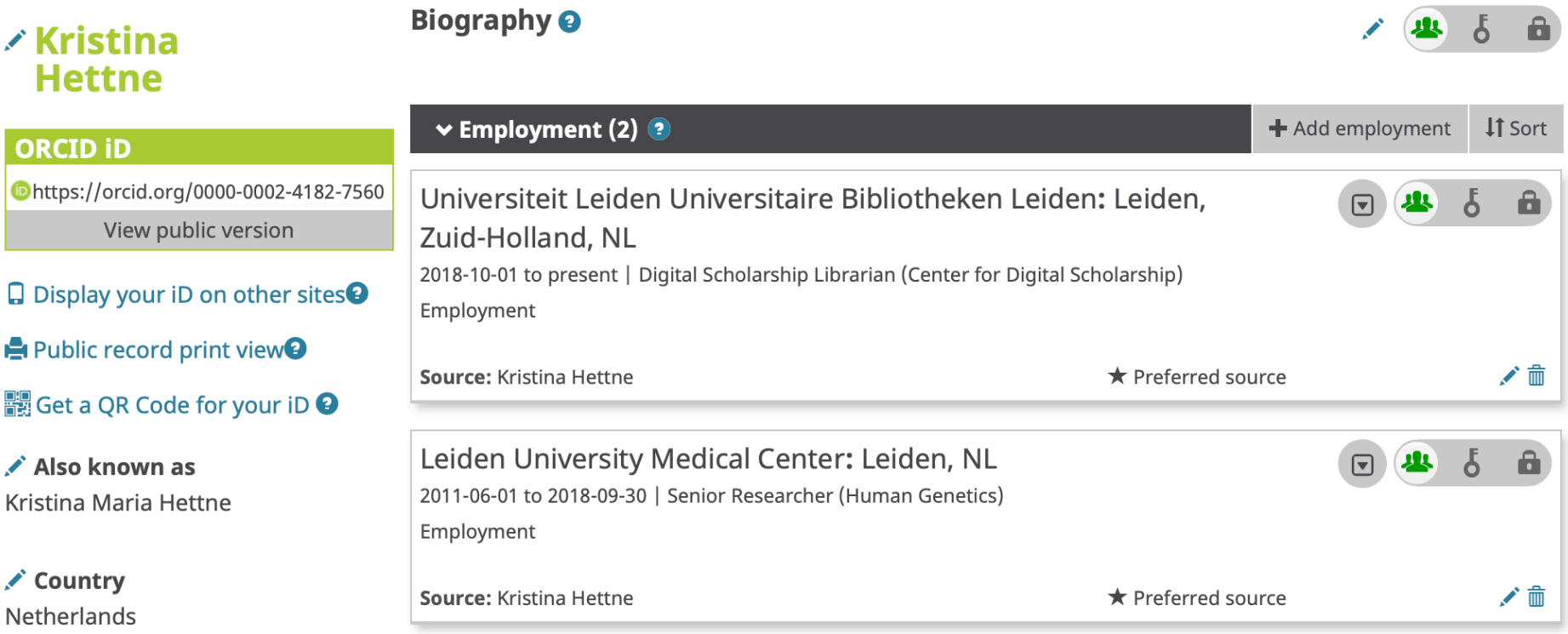
Question:

- Do you already have experience with data publishing?

Most important lessons about data publishing

- Publish you data in a repository and get a persistent identifier (for example a DOI)
- Get an ORCID: that way you can connect yourself with your publications and your data

https://orcid.org/my-orcid



Which repository to choose?

Science

✓ Meets all requirements ? Partly meets all requirements × Does not meet all requirements Not applicable **Recommended by this faculty**

Before

- ✓ DMP Online *(International)*
- Essentials 4 Data Support *(National)*
- MANTRA *(International)*
- ✓ NWO datamanagementplan *(National)*
- ✓ Template DMP Leiden *(Local)*

During

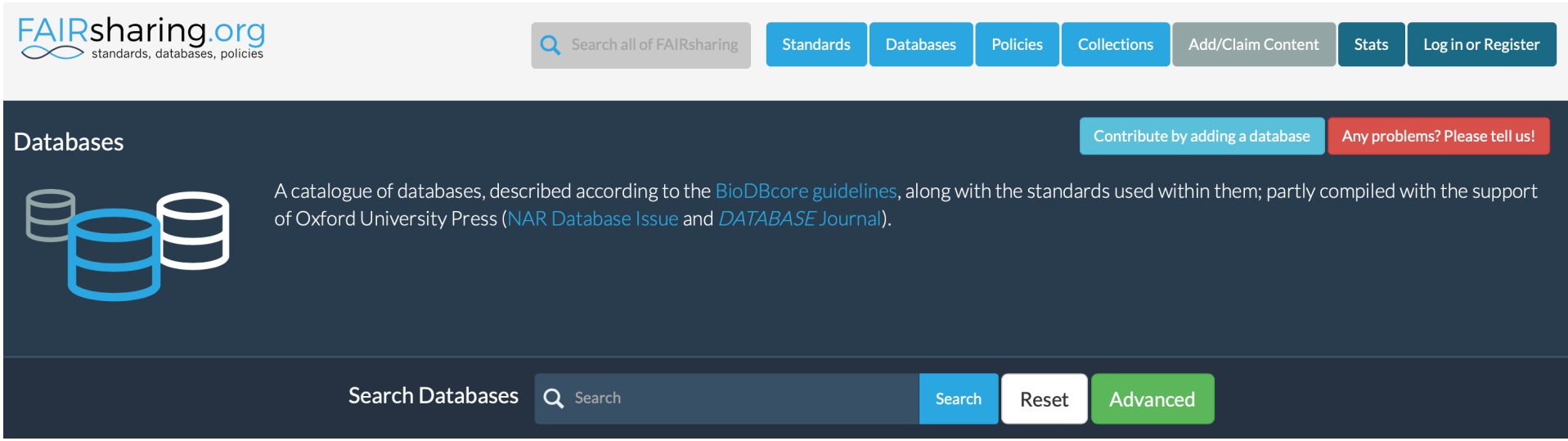
- ? B2DROP *(International)*
- ? B2SAFE *(International)*
- ? B2SHARE *(International)*
- ? B2STAGE *(International)*
- ✓ Bulkstorage *(Local)*
- ✓ Data Verse Network *(International)*
- ? Dataopslag Cell Observatory *(Local)*
- ✓ DCCD *(International)*
- × DDMoRe - Drug Disease Model Resources *(International)*
- ✓ Departments *(Local)*
- ✓ Dutch Dataverse Network (DDN) *(National)*
- Electronic Lab Journal *(-- Unknown --)*
- ? Figshare *(International)*
- × Infrared Space Observatory data archive *(International)*
- ✓ ISRIC - World Soil Information *(International)*
- Open Machine Learning (OpenML) *(International)*
- ✓ SeaDataNet *(International)*
- ✓ SURF Data Archive *(National)*
- ? SURFdrive *(National)*
- ✓ SURFfilesender *(National)*
- ✓ Virtual Research Environments *(Local)*
- ✓ Workgroups *(Local)*
- ✓ Zenodo *(International)*

After

- ✓ 4TU.ResearchData *(National)*
- × B2FIND *(International)*
- × B2SAFE *(International)*
- × B2SHARE *(International)*
- ? DataFirst *(International)*
- × DCCD *(International)*
- ? Dryad *(International)*
- × Figshare *(International)*
- ? ISRIC - World Soil Information *(International)*
- × MycoBank *(International)*
- Open Machine Learning (OpenML) *(International)*
- × OpenNeuro *(International)*
- ? SeaDataNet *(International)*
- × SURF Data Archive *(National)*
- ✓ TalkBank *(International)*
- ? Zenodo *(International)*

https://vre.universiteitleiden.nl/vre/lrd/Pages/FilterByPhase.aspx?model=discipline&phase=6

Alternative catalogues



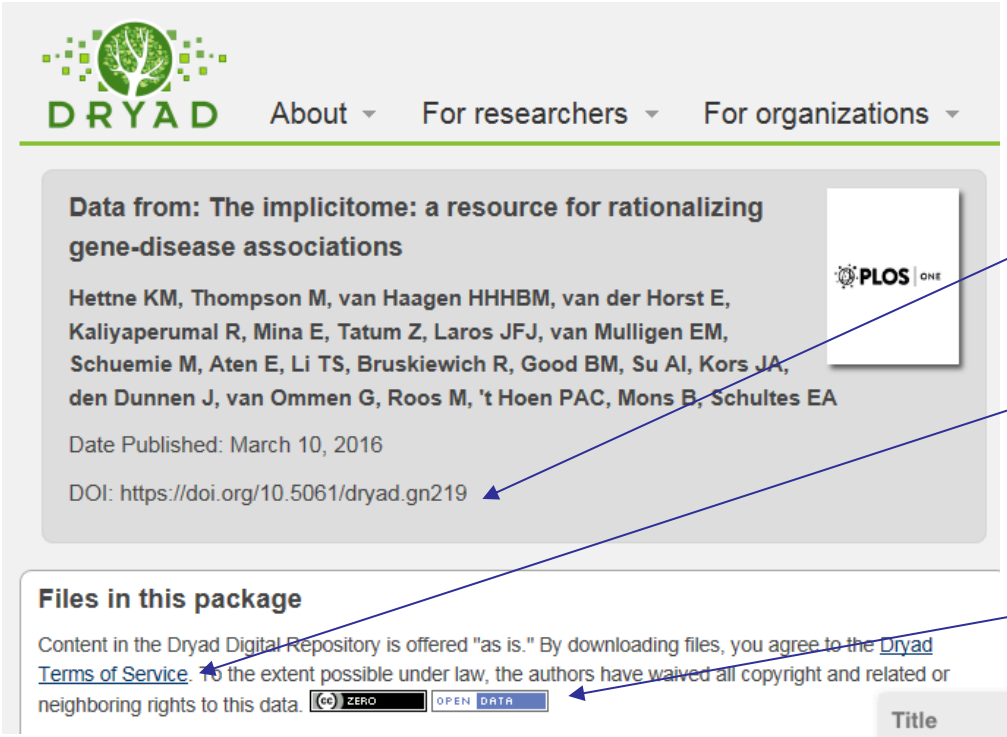
Sometimes a publisher requires a specific repository

Citation: Hettne KM, Thompson M, van Haagen HHHBM, van der Horst E, Kaliyaperumal R, Mina E, et al. (2016) The Implicitome: A Resource for Rationalizing Gene-Disease Associations. PLoS ONE 11(2): e0149621. https://doi.org/10.1371/journal.pone.0149621

Editor: Tudor Groza, Garvan Institute of Medical Research, AUSTRALIA

Received: September 22, 2015; **Accepted:** February 3, 2016; **Published:** February 26, 2016

Data Availability: All data files relevant to this publication have been submitted to the DataDryad repository with DOI: doi:10.5061/dryad.gn219.


DRYAD
About
For researchers
For organizations
Data from: The implicitome: a resource for rationalizing gene-disease associations
PLOS ONE
Hettne KM, Thompson M, van Haagen HHHBM, van der Horst E, Kaliyaperumal R, Mina E, Tatum Z, Laros JFJ, van Mulligen EM, Schuemie M, Aten E, Li TS, Bruskiewich R, Good BM, Su AI, Kors JA, den Dunnen J, van Ommen G, Roos M, 't Hoen PAC, Mons B, Schultes EA
Date Published: March 10, 2016
DOI: https://doi.org/10.5061/dryad.gn219
Files in this package
Content in the Dryad Digital Repository is offered "as is." By downloading files, you agree to the Dryad Terms of Service. To the extent possible under law, the authors have waived all copyright and related or neighboring rights to this data.
CC ZERO
OPEN DATA
F
A
R
I
Title
All associations as nanopublications
Downloaded
49 times
Description
The complete set of all ~204 million associations (explicit and implicit) as nanopublications. Each nanopublication asserts an association between a gene and a disease concept and the percentile rank of the match score.
Download
gda-np.nq.gz (29.22 Gb)
Details
View File Details
Title
All associations as CSV
Downloaded
50 times
Description
All ~204 million associations (explicit and implicit) listed as a CSV text file. Note that concept pairs are specified by concept ID and its first label according to our thesaurus.
Download
README.txt (228 bytes)
Download
matchscores.release.csv.gz (7.009 Gb)
Details
View File Details

However…

- Even though a repository might accept RDF, they do not offer a searchable interface to the individual data points

CSV on the Web

- Tabular data can easily be shared as CSV
- To describe the tables and explain meaning of columns for both humans and computers, use tabular metadata
- You can express mappings from columns to subject, predicate and object URIs or literal values using such tabular metadata

```
"country","country group","name (en)”
"at","eu","Austria”
"be","eu","Belgium”
```

```
{
  "titles": "country",
  "name": "country",
  "valueUrl": "http://example.org/country/{country}",
  "propertyUrl": "schema:url”
}
```

```
<#at> schema:url <http://example.org/country/at>
```

Data *set* modelling: DCAT

- Data Catalog Vocabulary (DCAT) is a data model for datasets (and catalogues and distributions)
  - "A **dataset** in DCAT is defined as a "collection of data, published or curated by a single agent, and available for access or download in one or more serializations or formats". A dataset is a conceptual entity, and can be represented by one or more **distributions** that serialize the dataset for transfer." – DCAT 2 Scope (draft)
- DCAT is widely used in government data portals and data repositories
  - Google uses DCAT metadata when indexing datasets in Google Dataset Search

- A Catalog describes Datasets
- A Dataset is (usually) available as one or more Distributions

- The FAIR DataPoint software developed at the LUMC and CKAN data publishing software work with DCAT

Demo FAIR DataPoint

What did we learn?

Universiteit
Leiden
The Netherlands

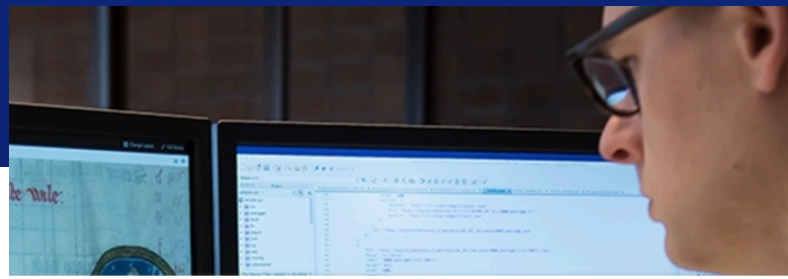
Learning objectives - recap

- Basic knowledge of the FAIRification process
- Basic knowledge of conceptual data modelling
- Basic knowledge of semantic data modelling
- Basic skills in OpenRefine with RDF and Wikidata plugin
- Basic knowledge of data publication

Questions

- What was the most useful thing you learned today?
- What was the least useful thing you learned today?

Thank you!



Blog: https://digitalscholarshipleiden.nl/

Universiteit Leiden
The Netherlands

Centre for Digital Scholarship

https://www.library.universiteitleiden.nl/research-and-publishing/centre-for-digital-scholarship