

# Making your data FAIR



Kristina Hettne

08/05/2020



All content is licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/) logo's excluded and unless specified otherwise in the caption of an image.

Universiteit Leiden  
The Netherlands

# Findable, Accessible, Interoperable, 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) <https://doi.org/10.1038/sdata.2016.18>

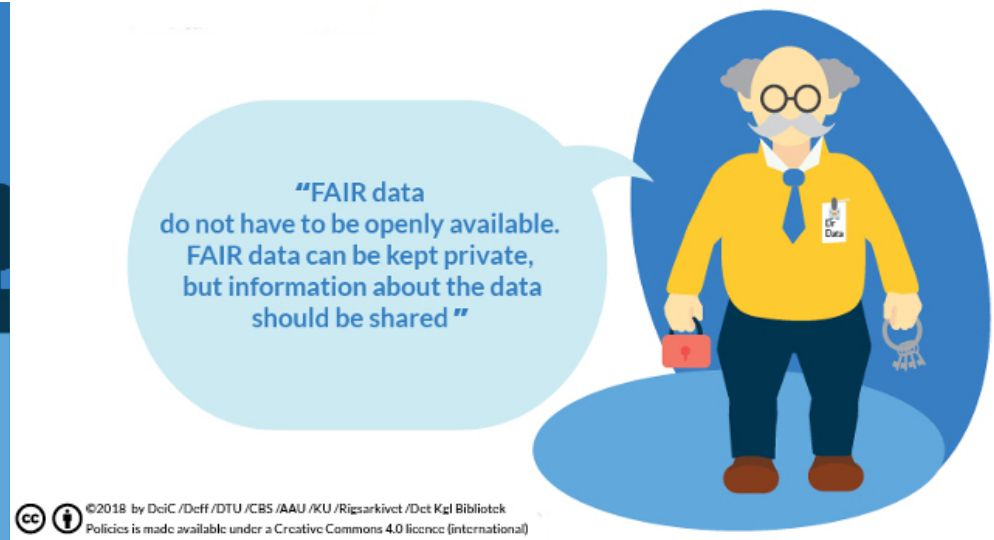
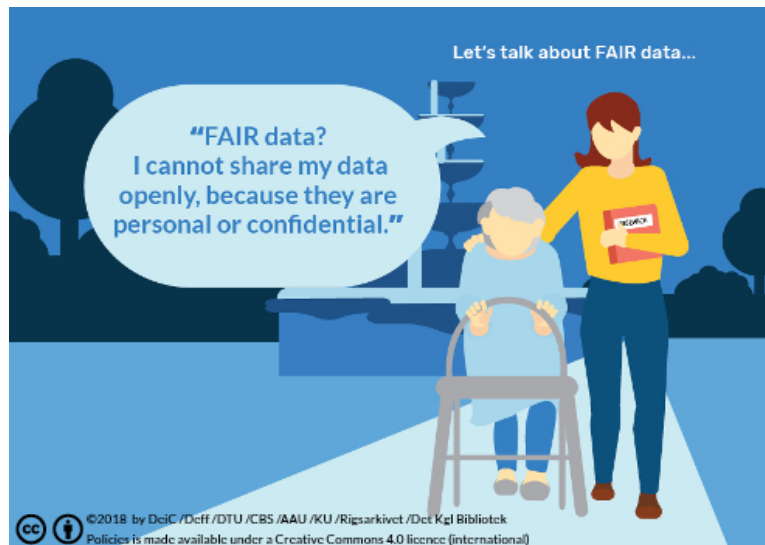
Research data needs to:

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

Privacy-sensitive data can meet the FAIR principles

# By making your data FAIR you...

- Create opportunities for sharing and reuse
- Enlarge your exposure
- Enhance your impact
- Show your future employer what you have done
- Avoid issues about verification
- Comply with requirements from funders



# It's all about 15 principles

## Box 2 | The FAIR Guiding Principles

### To be Findable:

- 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 it describes
- F4. (meta)data are registered or indexed in a searchable resource

### To be Accessible:

- A1. (meta)data are retrievable by their identifier using a standardized communications protocol
  - A1.1 the protocol is open, free, and universally implementable
  - A1.2 the protocol allows for an authentication and authorization procedure, where necessary
- A2. metadata are accessible, even when the data are no longer available

### To be 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

### To be Reusable:

- 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

<http://www.nature.com/articles/sdata201618>

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

# The FAIRification process

## Pre-FAIRification

### 1. identify FAIRification objective

*e.g. increase interoperability and define driving user question(s), or increase findability with metadata.*

### 2. analyze data

*e.g. investigate the representation (format) and meaning (semantics) of the data, or assess FAIR status.*

### 3. analyze metadata

*e.g. analyze availability of (or gather) metadata such as license and provenance information, or assess FAIR status.*

## FAIRification

### 4a. define semantic data model

*Reuse existing model, or generate a model through conceptual modelling and searching for ontology terms.*

### 5a. make data linkable

*Transform data into a machine-readable knowledge graph representation by using a semantic model.*

### 4b. define semantic metadata model

*Reuse existing model for generic items and define a model for domain-specific items.*

### 5b. make metadata linkable

*Transform metadata into a machine-readable knowledge graph representation by using a semantic model.*

## Post-FAIRification

### 7. assess FAIR data

*Assess if the objective is met e.g. answer driving user question(s), or assess FAIR status.*

### 6. host FAIR data

*Make FAIR data and metadata available for human and machine use via e.g. a FAIR Data Point.*

# FAIR-ness: what is in the hands of the researcher?

## Box 2 | The FAIR Guiding Principles

### To be Findable:

- 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 it describes
- F4. (meta)data are registered or indexed in a searchable resource

### To be Accessible:

- A1. (meta)data are retrievable by their identifier using a standardized communications protocol
  - A1.1 the protocol is open, free, and universally implementable
  - A1.2 the protocol allows for an authentication and authorization procedure, where necessary
- A2. metadata are accessible, even when the data are no longer available

### To be Interoperable:

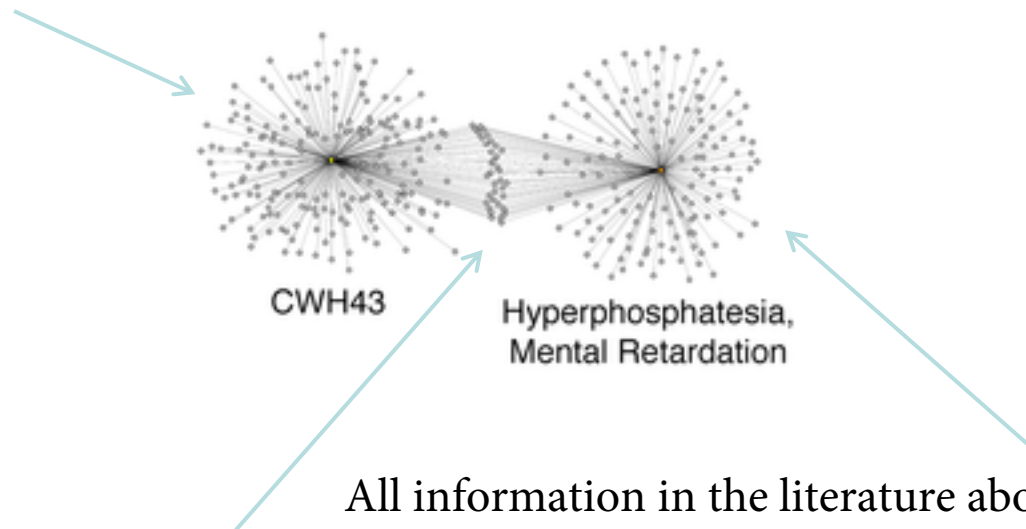
- I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (meta)data use vocabularies that follow FAIR principles
- I3. (meta)data include qualified references to other (meta)data

### To be Reusable:

- 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

# Example 1: finding new disease genes mined from scientific literature

All information in the literature about a gene



Overlapping information

=

New evidence for associating a gene with a disease

Thanks to data reuse:  
~204.000.000 novel gene-  
disease associations

Hettne KM, Thompson M, van Haagen HHHBM, van der Horst E, Kaliyaperumal R, 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>

<https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0149621>

**Data from: The implicitome: a resource for rationalizing gene-disease associations**



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, Bruskiwich 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.  

<b>Title</b>	<b>All associations as nanopublications</b>
<b>Downloaded</b>	49 times
<b>Description</b>	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.
<b>Download</b>	<a href="#">gda-np.nq.gz (29.22 Gb)</a>
<b>Details</b>	<a href="#">View File Details</a>

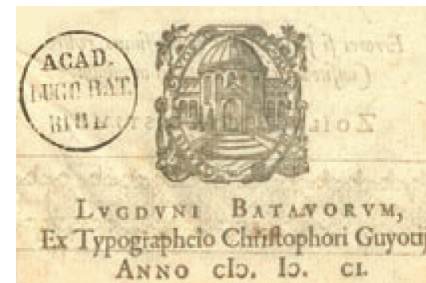
<b>Title</b>	<b>All associations as CSV</b>
<b>Downloaded</b>	50 times
<b>Description</b>	All ~204 million associations (explicit and implicit) listed as a CSV text file. Note that concept pairs are specified by concept ID and its first label according to our thesaurus.
<b>Download</b>	<a href="#">README.txt (228 bytes)</a>
<b>Download</b>	<a href="#">matchscores.release.csv.gz (7.009 Gb)</a>
<b>Details</b>	<a href="#">View File Details</a>

F  
A  
I  
R



# Example 2: Book trade history in Leiden

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



**Peter Verhaar**  
Digital Scholarship Librarian

Leiden University Fund (LUF) teaching grant

# Raw data (Microsoft Word)

**Aa, Cornelis van der** (\* 1749?; † ?; w. 1767-?)

Boekverkoper.

Geen lid van de grote boekverkoopersfamilie 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.

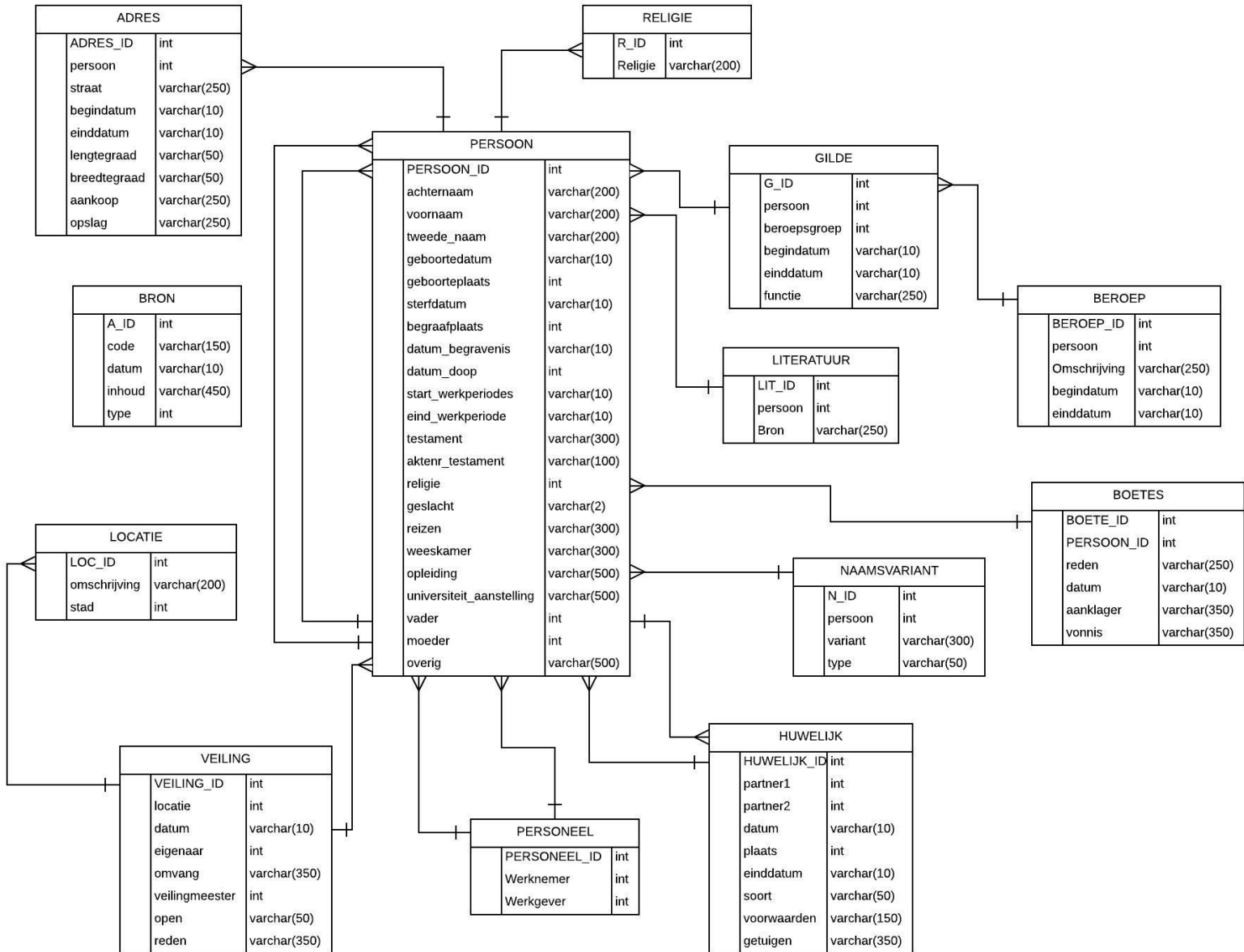
# Data cleaning and structuring

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

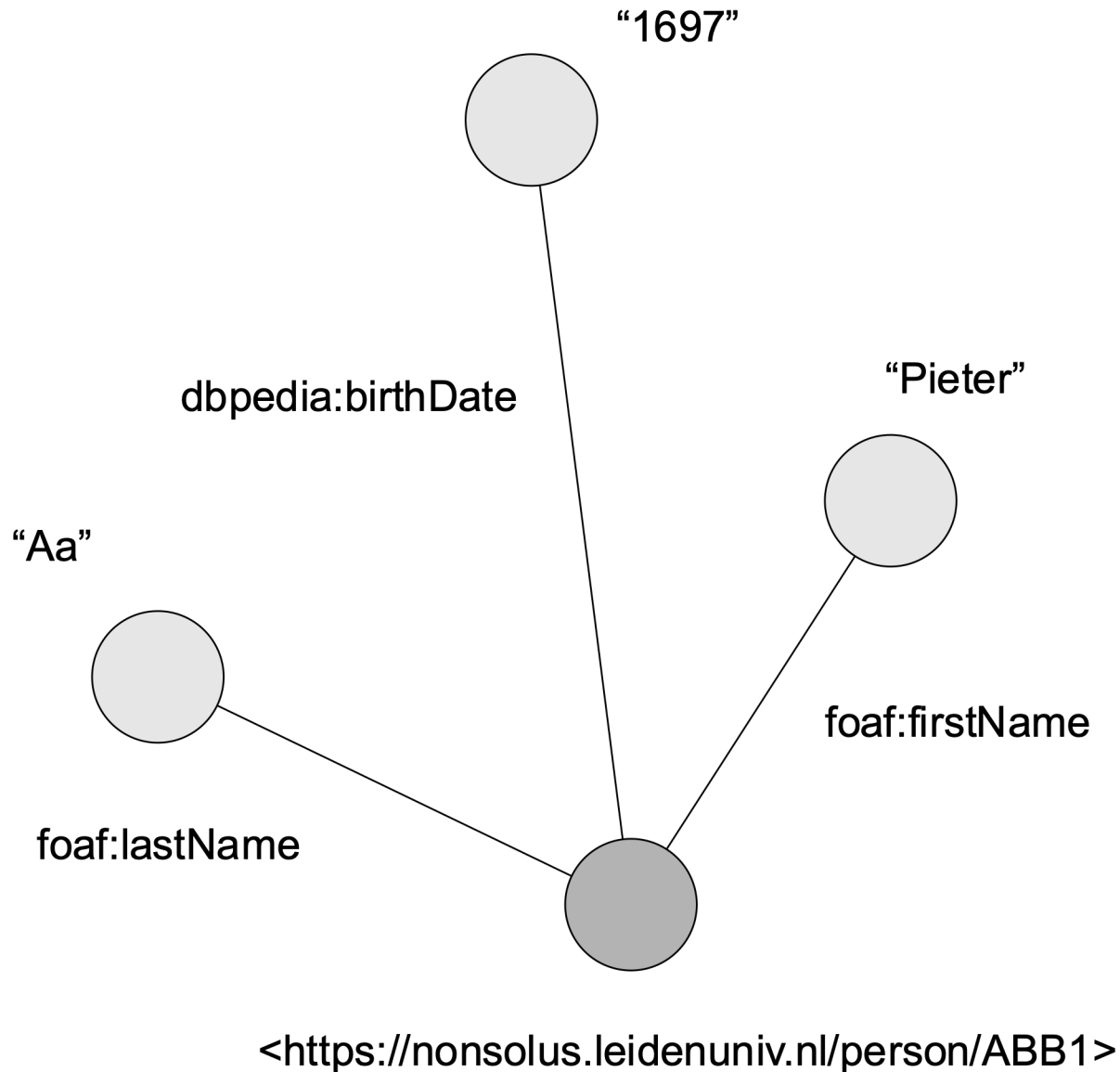
Aa, Pieter Jansz van der (\* Leiden 1697; † 2-8-1751 [begr. PK 31-7/7-8-1751]; w. 1719-36)

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

# Define data model



# Define the semantic data model: a person





- Main page
- Community portal
- Project chat
- Create a new item
- Create a new lexeme
- Recent changes
- Random item
- Query Service
- Nearby
- Help
- Donate

Tools

- What links here
- Related changes
- Special pages
- Permanent link
- Page information
- Concept URI
- Cite this page

Item [Discussion](#)

Read [View history](#)

# Abraham Elzevir (Q2733599)

Dutch printer

[edit](#)

Abraham Elsevier

[In more languages](#) [Configure](#)

Language	Label	Description	Also known as
English	Abraham Elzevir	Dutch printer	Abraham Elsevier
Dutch	Abraham Elsevier	No description defined	
German	No label defined	No description defined	
French	Abraham Elzevir	No description defined	

[All entered languages](#)

## Statements

instance of human [edit](#)

[2 references](#)

[+ add value](#)

# Make data linkable: transform to RDF

**OpenRefine** Bookkeepers in Leiden Kristina [Permalink](#) Open... Export ▾ Help

Facet / Filter Undo / Redo 10 / 10 **1311 rows** Extensions: RDF ▾ Wikidata ▾

« first < previous **1 - 10** next > last »

**RDF Schema alignment**

The RDF schema alignment skeleton below specifies how the RDF data that will get generated from your grid-shaped data. The cells in each record of your data will get placed into nodes within the skeleton. Configure the skeleton by specifying which column to substitute into which node.

**Base URI:** <http://h2676137.stratoserver.net:3333/> [Edit](#)

[RDF skeleton](#) [RDF Preview](#)

This is a sample `Turtle` representation of (up-to) the first 10 rows

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .

<http://h2676137.stratoserver.net:3333/Person/1f04be3baaa7921d4ab9a7095782cddb> a <http://purl.obolibrary.org/obo/NCBITaxon_9606> ;
  rdfs:label "Person" ;
  foaf:firstName "Boudewijn" ;
  foaf:lastName "Aa" ;
  <http://semanticscience.org/resource/SIO_001317> "Boudewijnsz van der" .

<http://h2676137.stratoserver.net:3333/Birthdate/1f04be3baaa7921d4ab9a7095782cddb> a <http://dbpedia.org/ontology/birthDate> ;
  rdfs:label "Birthdate" ;
  <http://semanticscience.org/resource/SIO_000300> "1676" .

<http://h2676137.stratoserver.net:3333/Person/1f04be3baaa7921d4ab9a7095782cddb> <http://semanticscience.org/resource/SIO_000008> <http://purl.obolibrary.org/obo/ERO_0001966> <http://purl.bioontology.org/ontology/SNOMEDCT/703117000> .

<http://h2676137.stratoserver.net:3333/Person/ca8dd149c303f616ebd658960458a051> a <http://purl.obolibrary.org/obo/NCBITaxon_9606> ;
  rdfs:label "Person" .
```

BirthName	placeOfDeath	placeOfDeathNa
NULL	NULL	N
PLACE1	leiden	N
NULL	NULL	N
NULL	NULL	N
NULL	NULL	N
NULL	NULL	N
NULL	NULL	N
NULL	NULL	N
NULL	NULL	N
NULL	NULL	N

This was done with the OpenRefine RDF plugin, but there are many other ways...

# Host FAIR data



Universiteit  
Leiden

Onderzoek

Onderwijs

Wetenschappers

Over ons

Faculiteiten

Campus Den Haag

Alumni

Bibliotheek

## Database over het Leidse Boek

Kopieer om de database te bevragen de genoemde zoekvraag in het tekstveld en klik op "Zoek!"

### Alle personen die geboren zijn buiten Leiden:

```
SELECT achternaam , voornaam , tweede_naam , geboortedatum , s.Naam  
FROM PERSOON p , STAD s  
WHERE p.geboorteplaats = s.S_ID  
AND s.Naam != 'leiden'  
ORDER BY p.geboortedatum
```

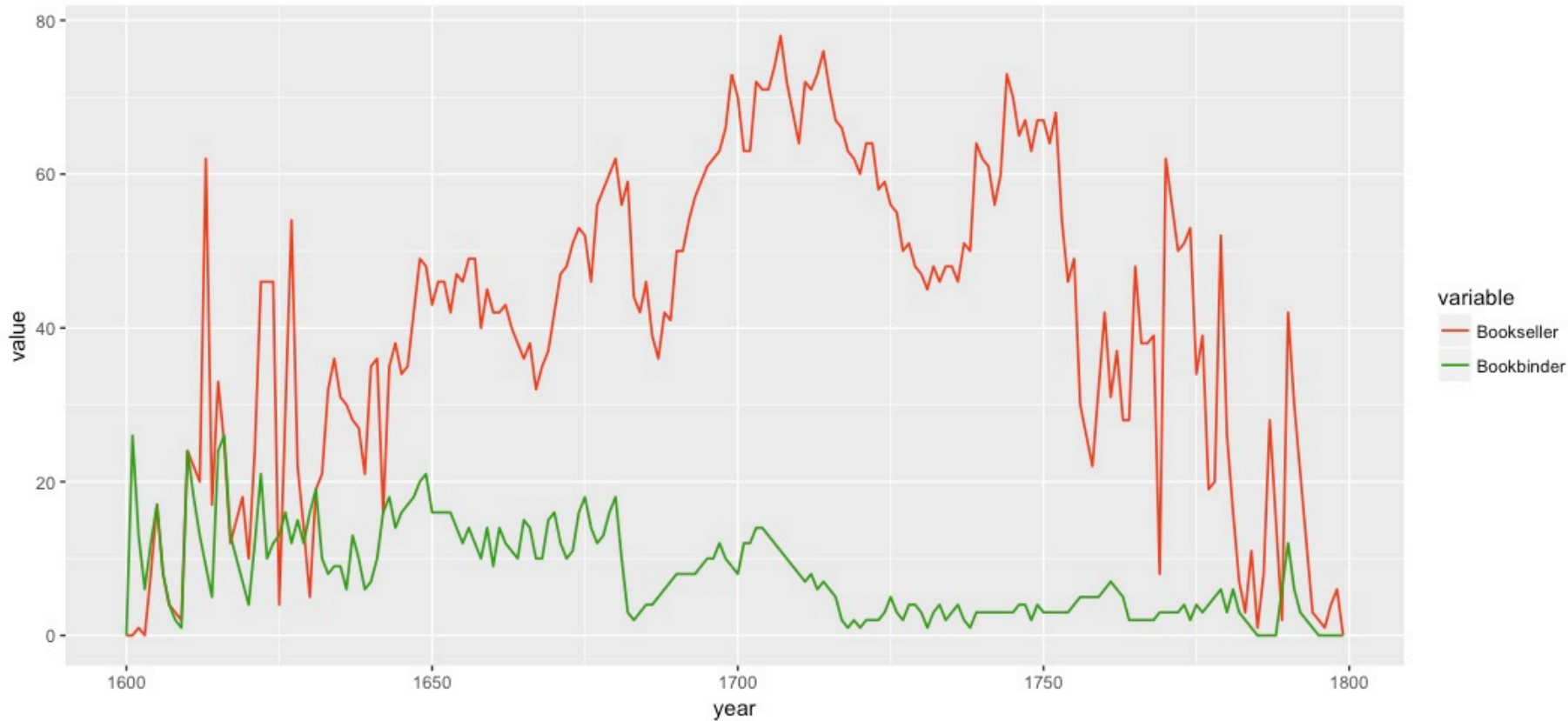
Zoek!

<https://bookandbyte.universiteitleiden.nl/Boekgeschiedenis/>



# BA student projects “Boekgeschiedenis in de praktijk” spring 2019

- The database has been used to study the following topics:
  - The history of the guild of booksellers in Leiden
  - The network of family relations
  - The influx of foreigners in the Leiden book trade
  - The occurrences of banned books on the catalogues of book actions organised in Leiden
- Insights from these projects led to improvements of the database
- Ongoing project

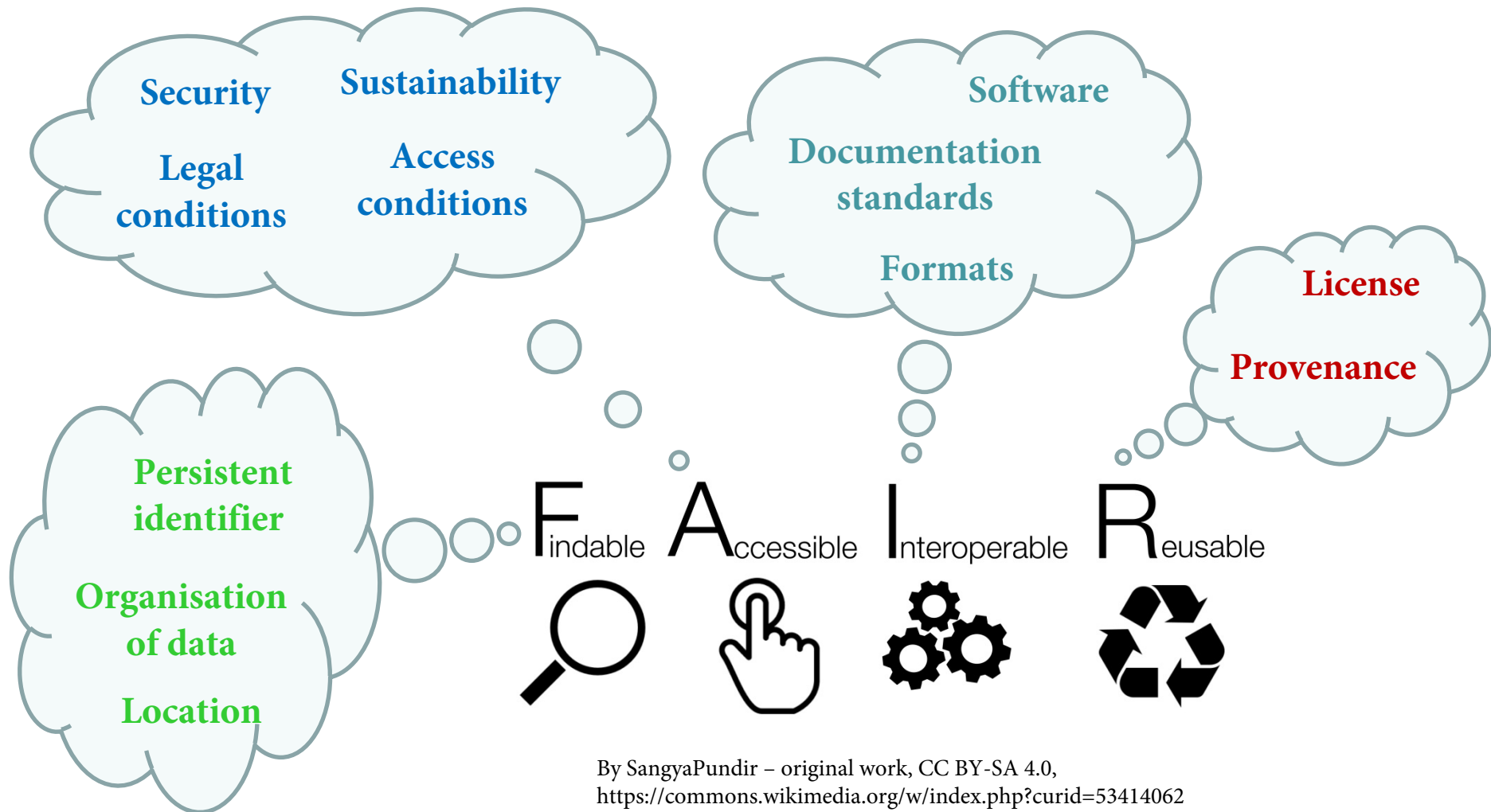


## Number of booksellers and bookbinders, 1600-1800

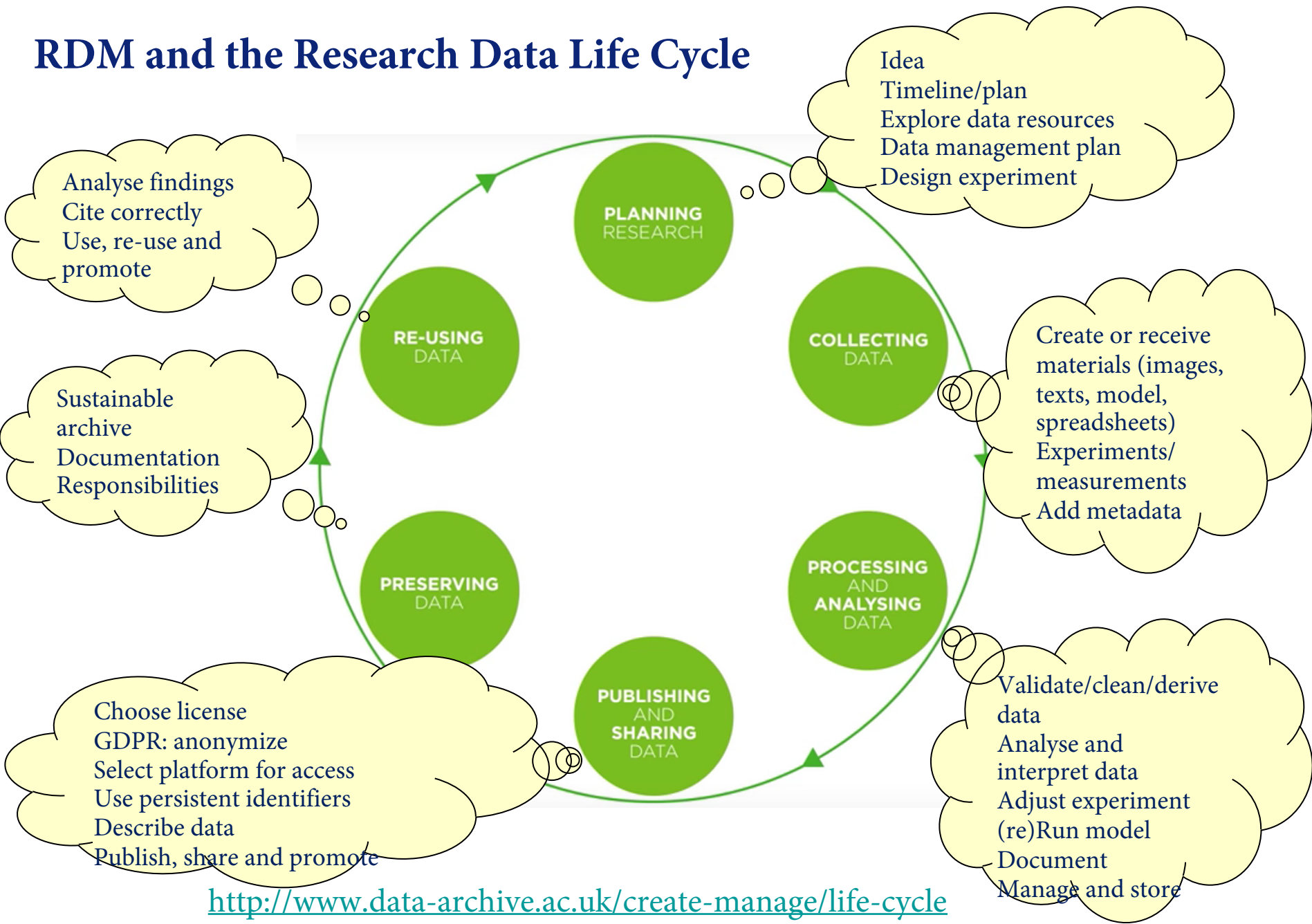


## Locations of booksellers in 1700

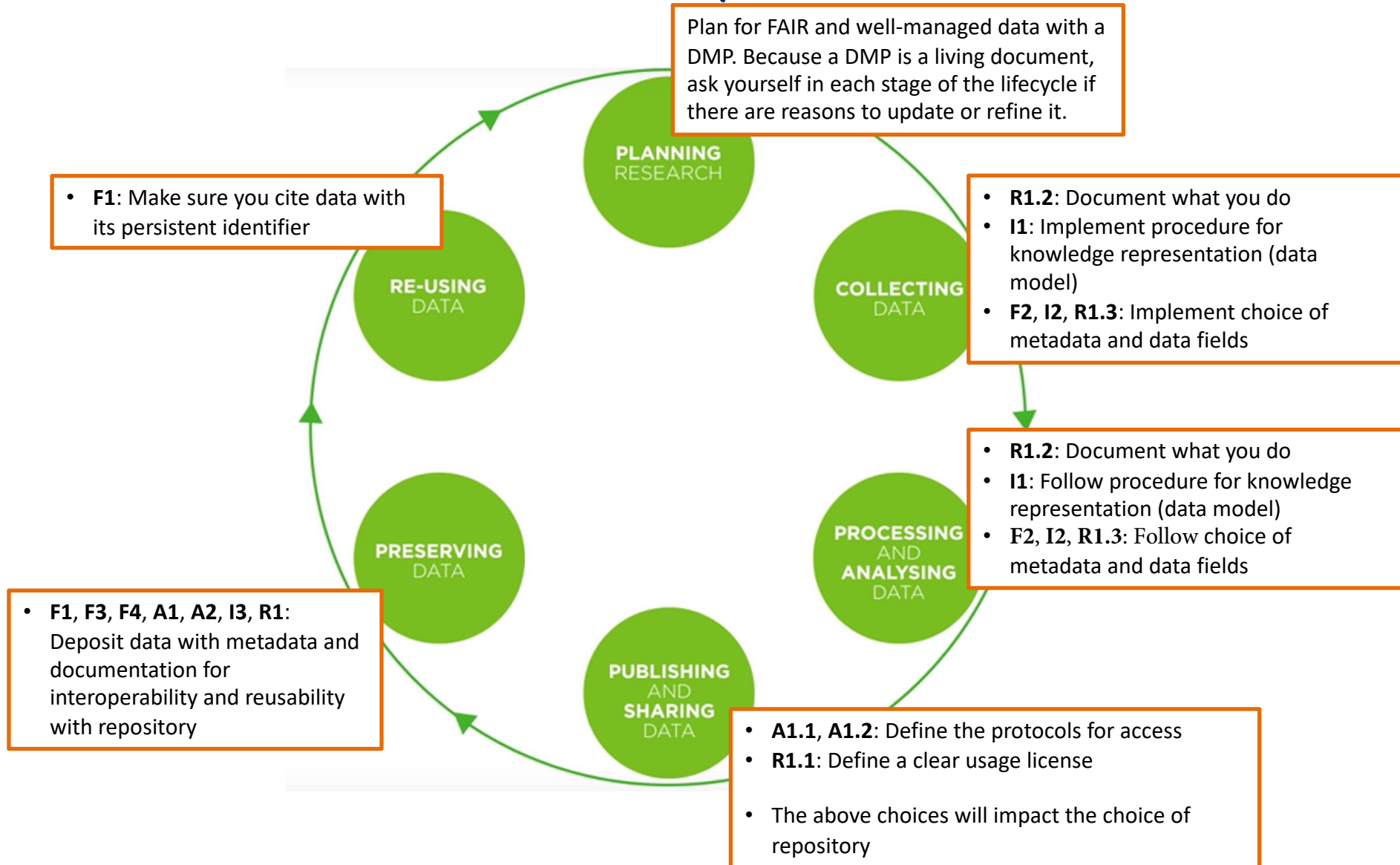
# FAIR data management planning



# RDM and the Research Data Life Cycle



# FAIR and the Research Data Life Cycle



<http://www.data-archive.ac.uk/create-manage/life-cycle>

# 4 steps to make your data more FAIR

- F: Put your data in a repository
- A: Make sure there is a data access protocol
- I: Describe your data using the metadata scheme offered by the repository
- R: Choose a license

# Dataset FAIR-ness

- As a researcher you do a good job if you:
  - Fill out as much as possible of the repository fields when you submit your data
- As a researcher you do a great job if you:
  - Use metadata standards to record metadata elements (this is hard, ask colleagues or a data steward for help!)
- As a researcher you do an amazing job if you:
  - Use standard vocabularies to record data elements (this is hard, ask a data steward for help!)
  - Save your data in an FAIR interoperable format such as XML or RDF (this is hard, ask a data steward for help!)



# Want to learn more?

## FEATURED ARTICLES

MORE +



### The FAIR Principles: First Generation Implementation Choices and Challenges

Author : Barend Mons; Erik Schultes; Fenghong Liu; Annika Jacobsen

Institution : Leiden University Medical Center, Leiden 2333 ZA, The Netherlands; GO FAIR Inter...

Doi : 10.1162/dint\_e\_00023

[Abstract \(views 93 \)](#) | [Full Text PDF](#) | [Full Text HTML](#)



### FAIR Principles: Interpretations and Implementation Considerations

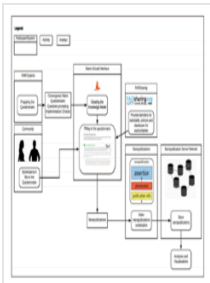
Author : Annika Jacobsen; Ricardo de Miranda Azevedo; Nick Juty; Dominique Batista; Simon C...

Institution : Leiden University Medical Center, Leiden 2333 ZA, The Netherlands; Institute of Dat...

Keywords : FAIR guiding principles; FAIR implementation; FAIR convergence; FAIR communitie...

Doi : 10.1162/dint\_r\_00024

[Abstract \(views 105 \)](#) | [Full Text PDF](#) | [Full Text HTML](#)



### FAIR Convergence Matrix: Optimizing the Reuse of Existing FAIR-Related Resources

Author : Hana Pergl Sustkova; Kristina Maria Hettne; Peter Wittenburg; Annika Jacobsen; Tobia...

Institution : GO FAIR International Support and Coordination Office, Leiden, The Netherlands; C...

Keywords : FAIR Implementation Choices and Challenges; Convergence; FAIR Communities

Doi : 10.1162/dint\_a\_00038

[Abstract \(views 114 \)](#) | [Full Text PDF](#) | [Full Text HTML](#)



### Data Intelligence

Host: National Science Library,  
Chinese Academy of Sciences

Publisher: National Science Library,  
Chinese Academy of Sciences

#### Co-Editors-in-Chief:

James Hendler , Huizhou Liu , Ying Ding

#### Executive Editors-in-Chief:

Guilin Qi , Yan Zhao



29 original papers  
Official issue out in 2020

<http://www.data-intelligence-journal.org>

# Maximum Exposure for yourself, your publications and your data

## Total citations

367

382

416

## Online attention



Altmetric score (what's this?)

Tweeted by 177

On 7 Facebook pages

Mentioned in 2 Google+ posts

NBDC [Credits] [Japanese | English]

Life Science Database Archive

Search across datasets...

Home About Archive Update History Data List Contact us

**FANTOM5**

About This Database

- Database Description
- Download
- License
- Update History of This Database

**Database Description**

General information of database

Database name	FANTOM5 <a href="#">vteegbio.jp</a>
Alternative name	Functional Annotation of the Mammalian Genome
DOI	10.18988/lsdb.nbd01389-000.V002
Version	1.0

nature International weekly journal of science

Home News & Comment Research Careers & Jobs Current Issue Archive Audio & Video For Authors

Archive Volume 507 Issue 7493 Articles Article

NATURE | ARTICLE

日本語要約

A promoter-level mammalian expression atlas

The FANTOM Consortium and the RIKEN PMI and CLST (DGT)

Editor's summary العربية

FANTOM5 (standing for functional annotation of the mammalian genome 5) is the fifth major stage of a major international collaboration that aims to dissect the transcriptional regulatory networks that...

Associated links

Scientific Data | Data Descriptor  
FANTOM5 CAGE profiles of human and mouse reprocessed for GRCh38 and GRCm38 genome assemblies  
Imad Abugessaia *et al.*

Scientific Data | Data Descriptor  
FANTOM5 CAGE profiles of human and mouse samples  
Shuhei Noguchi *et al.*

nature.com > scientific data > data descriptors > article

SCIENTIFIC DATA

Altmetric: 5 Citations: 2 More detail >>

Data Descriptor | OPEN

FANTOM5 CAGE profiles of human and mouse samples

Shuhei Noguchi, Takahiro Arakawa [...] Yoshihide Hayashizaki

Scientific Data 4, Article number: 170112 (2017)  
doi:10.1038/sdata.2017.112  
Download Citation

Received: 06 December 2016  
Accepted: 25 April 2017  
Published online: 29 August 2017

Cell biology  
Computational biology and bioinformatics  
Developmental biology Molecular biology  
Systems biology

Abstract

In the FANTOM5 project, transcription initiation events across the

Sections Figures References

26 March 2014

PDF Share Tools

Associated Content

Collection  
The FANTOM5 project

Nature | Article  
A promoter-level mammalian expression atlas

The FANTOM Consortium and the RIKEN PMI and CLST (DGT), Alistair R. R. Forrest [...] Yoshihide Hayashizaki

Scientific Data | Comment | OPEN  
The FANTOM5 collection, a data series underpinning mammalian transcriptome atlases in diverse cell types

Hideya Kawaji, Takeya Kasukawa [...] Yoshihide Hayashizaki

thresholding • Known gene tissue specificity • promoters • Key cell-type • Methods • Accession files and tables •

a natureresearch event

**Ferroc Materials:**  
Challenges and opportunities  
October 25-27, 2017 | Xi'an, China

spatial organization of cDNA sequencing, we use primary cells, cell the express s many ma TSSs, with nt cell types evolve at

ferences

Data repository

Data journal

Journal article

# Centre for Digital Scholarship



## Data management

In short, data management can be defined as the creation, storage, maintenance, disclosure, archiving and sustainable preservation of research data.



### RDM checklist

Guide to sound data management



### Training

Courses, workshops, and hands-on instruction



### Leiden University Data Management Network

Brings together research and support

#### RDM experts at the CDS



**Fieke Schoots**  
Digital Scholarship Librarian



**Michelle van den Berk**  
Digital Scholarship Librarian



**Kristina Hettne**  
Digital Scholarship Librarian



**Joanne Yeomans**  
Digital Scholarship Librarian

#### Data Management Calendar



Webinar  
**Online workshop: How to write a Data Management Plan (DMP)**  
Fieke Schoots



Network meeting  
**Data Network Meeting: Convening event and drinks**  
Joanne Yeomans



Course  
**How to publish your data**  
Michelle van den Berk

#### Useful links



University DMP template



Leiden University Data management Regulations

<https://www.library.universiteitleiden.nl/researchers/data-management>