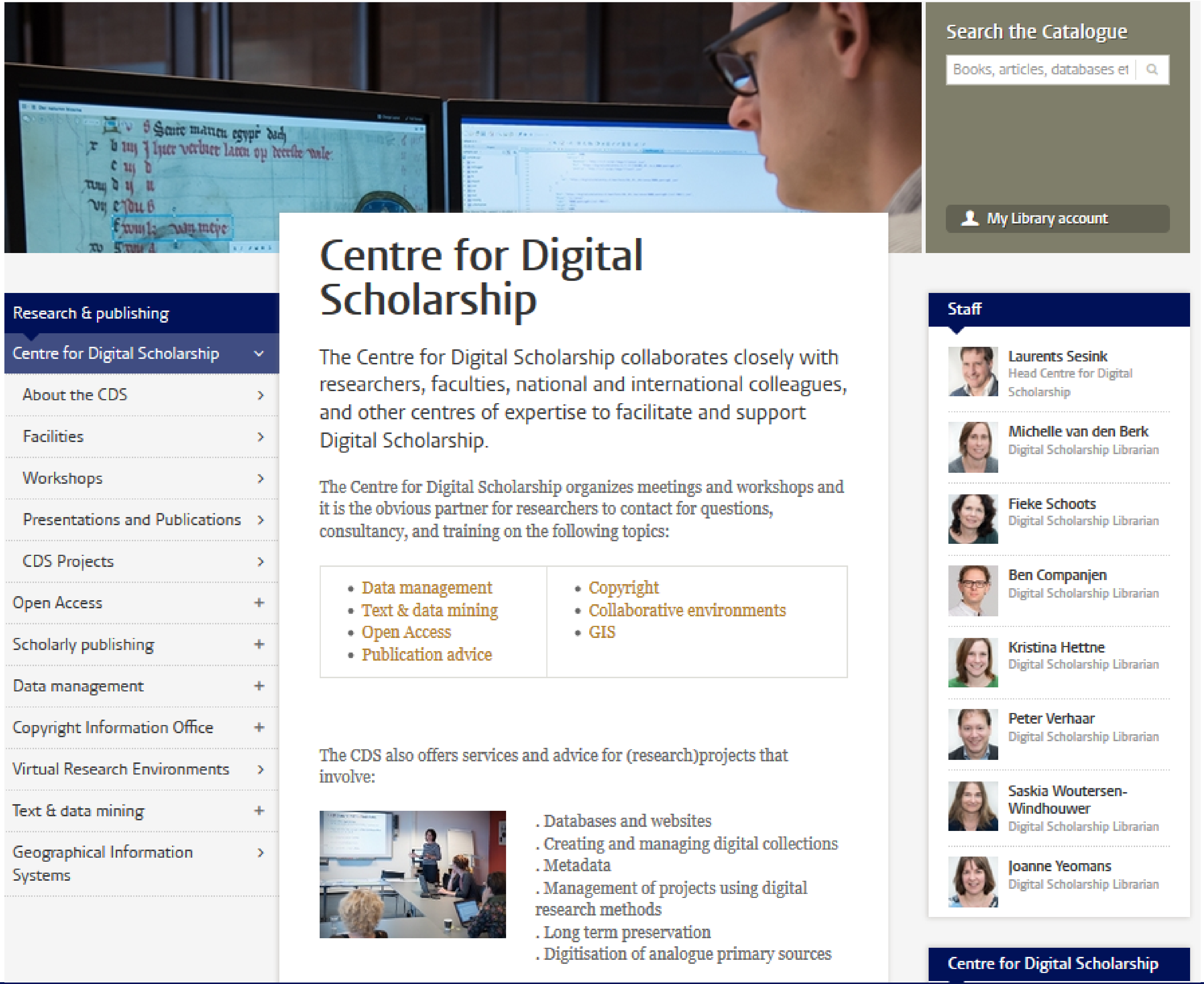
FAIR before FAIR: a case study in reproducible research

Kristina Hettne | Data Conversations

17 June 2019

Universiteit Leiden
The Netherlands



Search the Catalogue

Books, articles, databases et

My Library account

Research & publishing

- Centre for Digital Scholarship
  - About the CDS
  - Facilities
  - Workshops
  - Presentations and Publications
  - CDS Projects
- Open Access
- Scholarly publishing
- Data management
- Copyright Information Office
- Virtual Research Environments
- Text & data mining
- Geographical Information Systems

Centre for Digital Scholarship

The Centre for Digital Scholarship collaborates closely with researchers, faculties, national and international colleagues, and other centres of expertise to facilitate and support Digital Scholarship.

The Centre for Digital Scholarship organizes meetings and workshops and it is the obvious partner for researchers to contact for questions, consultancy, and training on the following topics:

• Data management • Text & data mining • Open Access • Publication advice	• Copyright • Collaborative environments • GIS

The CDS also offers services and advice for (research)projects that involve:

. Databases and websites
. Creating and managing digital collections
. Metadata
. Management of projects using digital research methods
. Long term preservation
. Digitisation of analogue primary sources

Staff

Laurents Sesink
Head Centre for Digital Scholarship

Michelle van den Berk
Digital Scholarship Librarian

Fieke Schoots
Digital Scholarship Librarian

Ben Companjen
Digital Scholarship Librarian

Kristina Hettne
Digital Scholarship Librarian

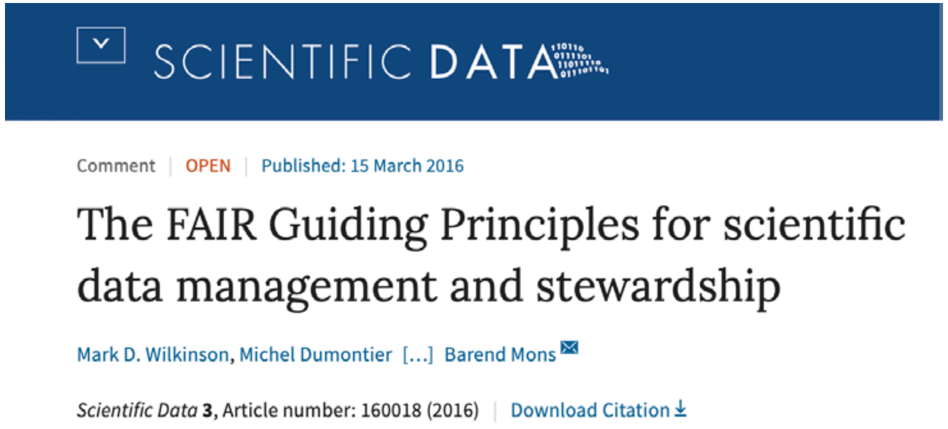
Peter Verhaar
Digital Scholarship Librarian

Saskia Woutersen-Windhouwer
Digital Scholarship Librarian

Joanne Yeomans
Digital Scholarship Librarian

Centre for Digital Scholarship

Findable, Accessible, Interoperable and Reusable (FAIR) – not “open”



SCIENTIFIC DATA

Comment | OPEN | Published: 15 March 2016

The FAIR Guiding Principles for scientific data management and stewardship

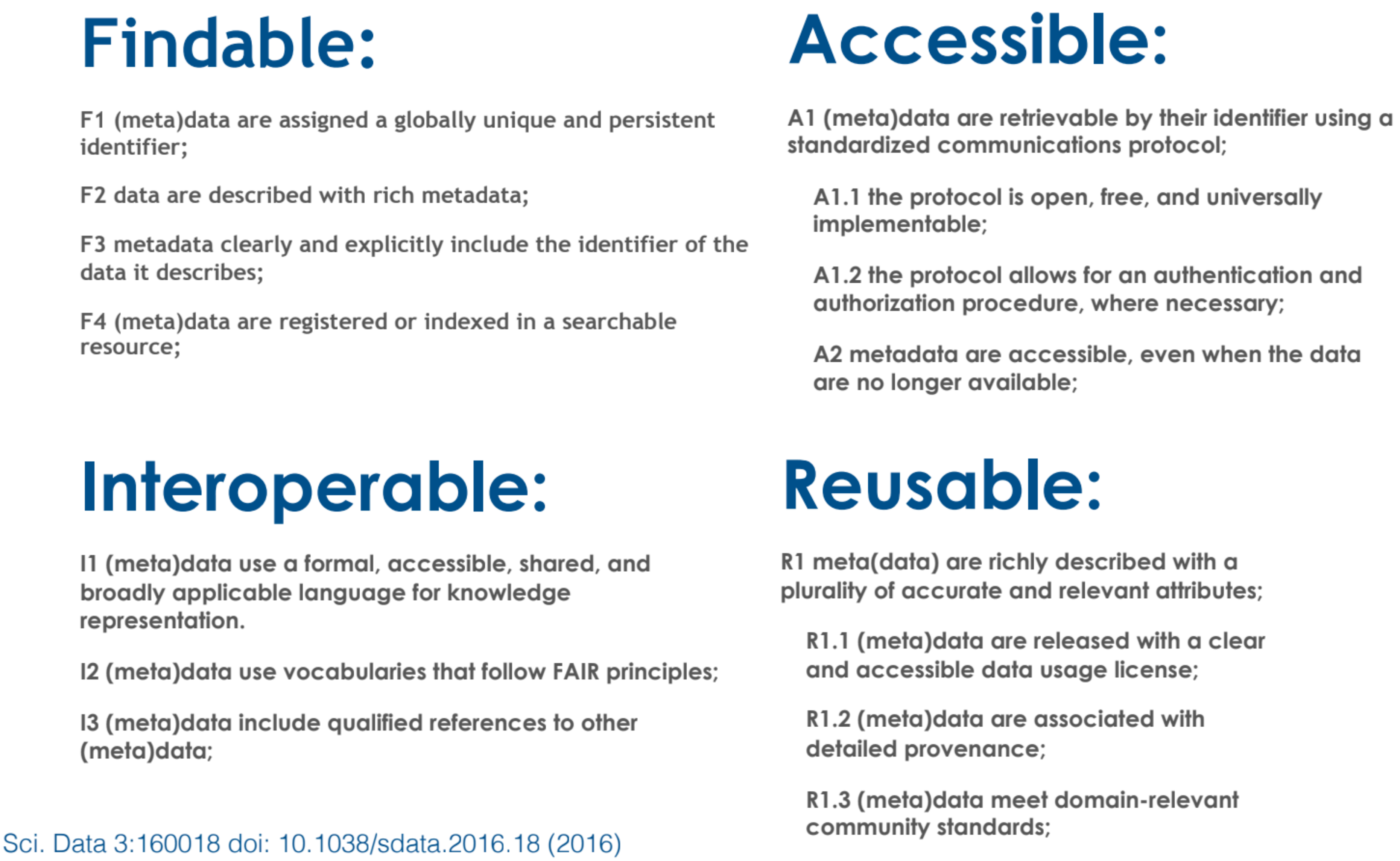
Mark D. Wilkinson, Michel Dumontier [...] Barend Mons

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



Findable:

F1 (meta)data are assigned a globally unique and persistent identifier;

F2 data are described with rich metadata;

F3 metadata clearly and explicitly include the identifier of the data it describes;

F4 (meta)data are registered or indexed in a searchable resource;

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;

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;

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;

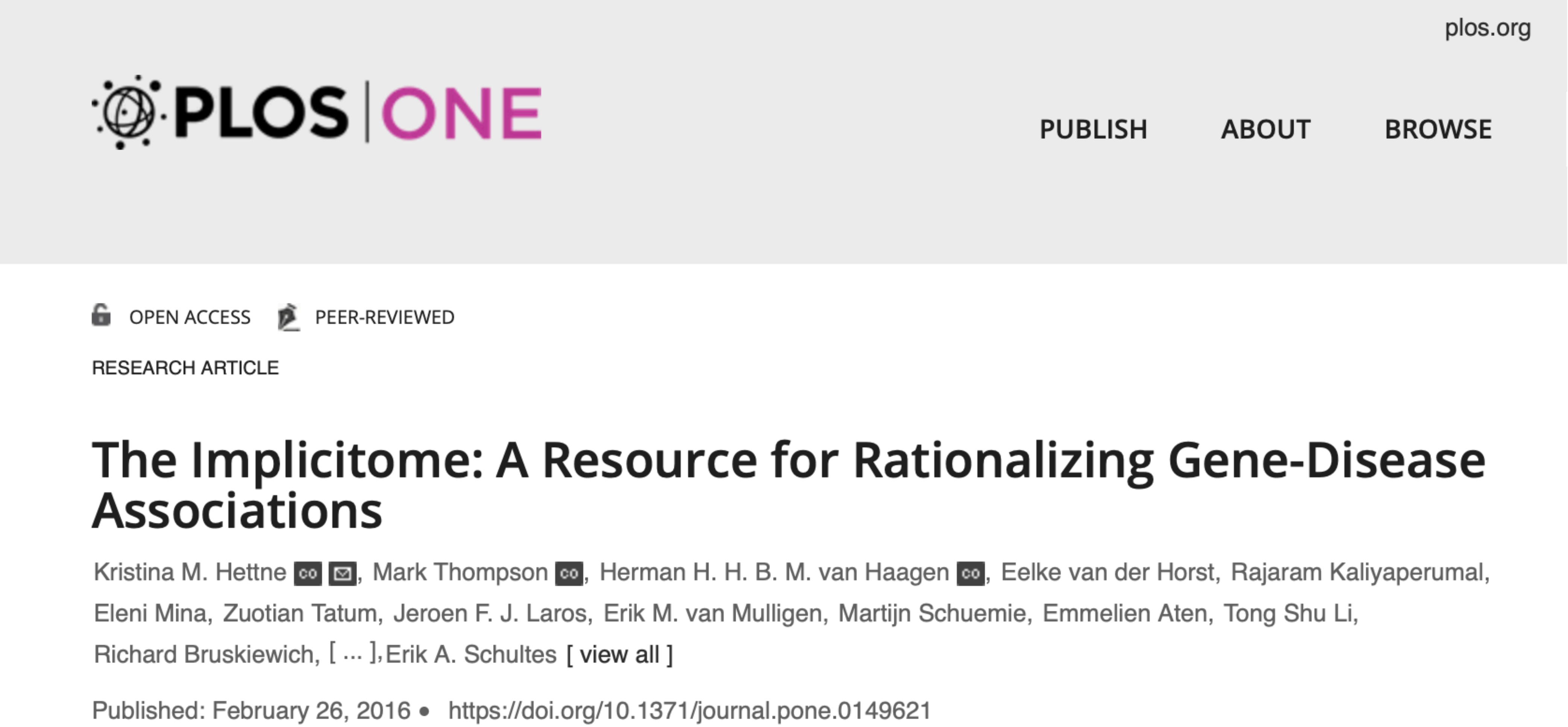
R1.2 (meta)data are associated with detailed provenance;

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

Sci. Data 3:160018 doi: 10.1038/sdata.2016.18 (2016)

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

Implementing FAIR before FAIR...



OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

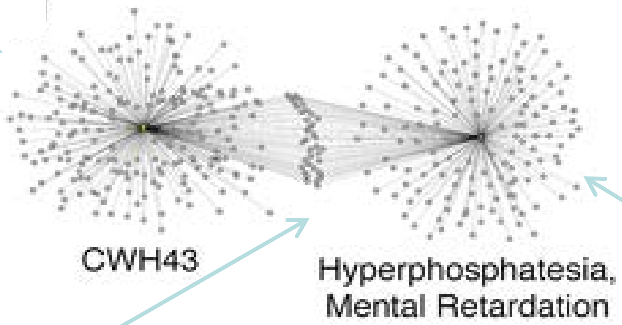
The Implicitome: A Resource for Rationalizing Gene-Disease Associations

Kristina M. Hettne, Mark Thompson, Herman H. H. B. M. van Haagen, Eelke van der Horst, Rajaram Kaliyaperumal, Eleni Mina, Zuotian Tatum, Jeroen F. J. Laros, Erik M. van Mulligen, Martijn Schuemie, Emmelien Aten, Tong Shu Li, Richard Bruskiewich, [...], Erik A. Schultes [view all]

Published: February 26, 2016 • https://doi.org/10.1371/journal.pone.0149621

Text mining for gene-disease associations

All weighted information in the literature about a gene



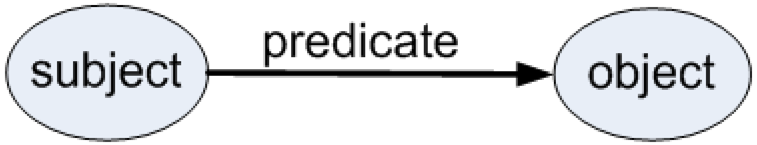
Overlapping and prioritized information

=

New evidence for associating a gene with a disease

All weighted information in the literature about a disease

Data reuse: ~204.000.000 new gene-disease associations modelled as semantic triples

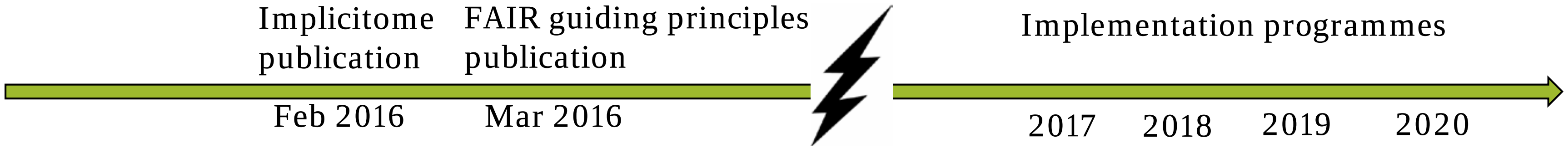


Implementing FAIR takes time and effort

Research time: ~80%
FAIRification: ~20%

(Second time: probably 5%)

https://goo.gl/UPFdhx



Implicitome publication

Feb 2016

FAIR guiding principles publication

Mar 2016

Implementation programmes

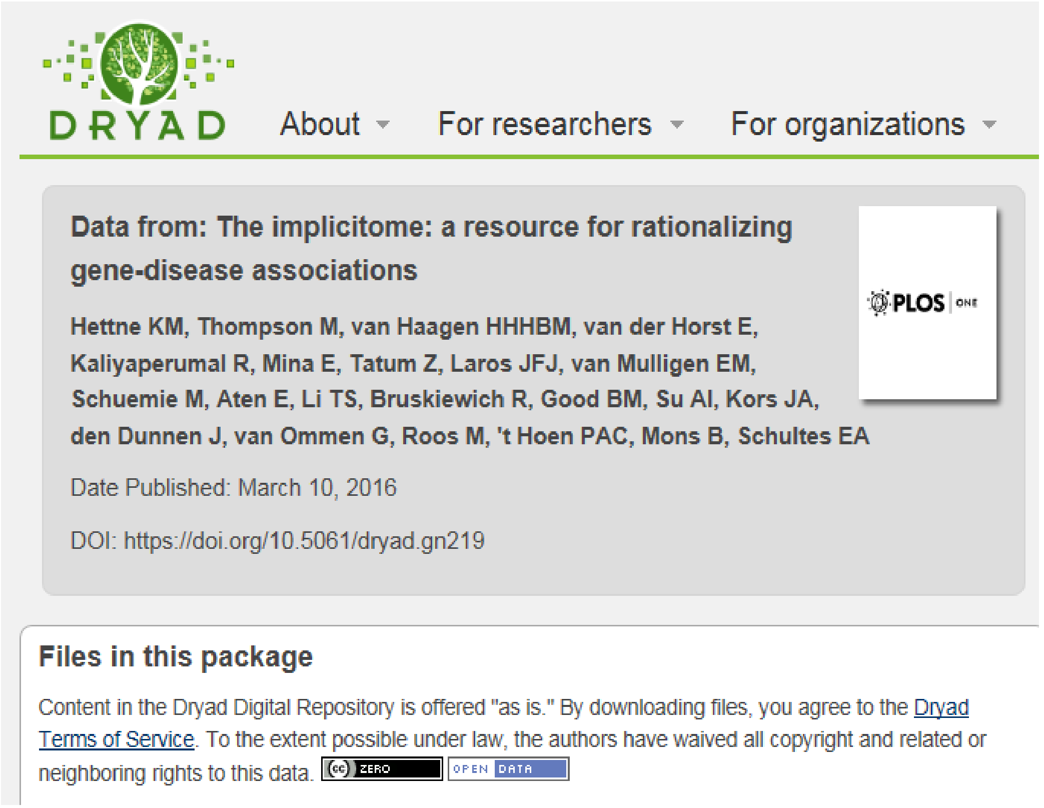
2017

2018

2019

2020

Used repositories and licenses (F, A, R)



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, 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

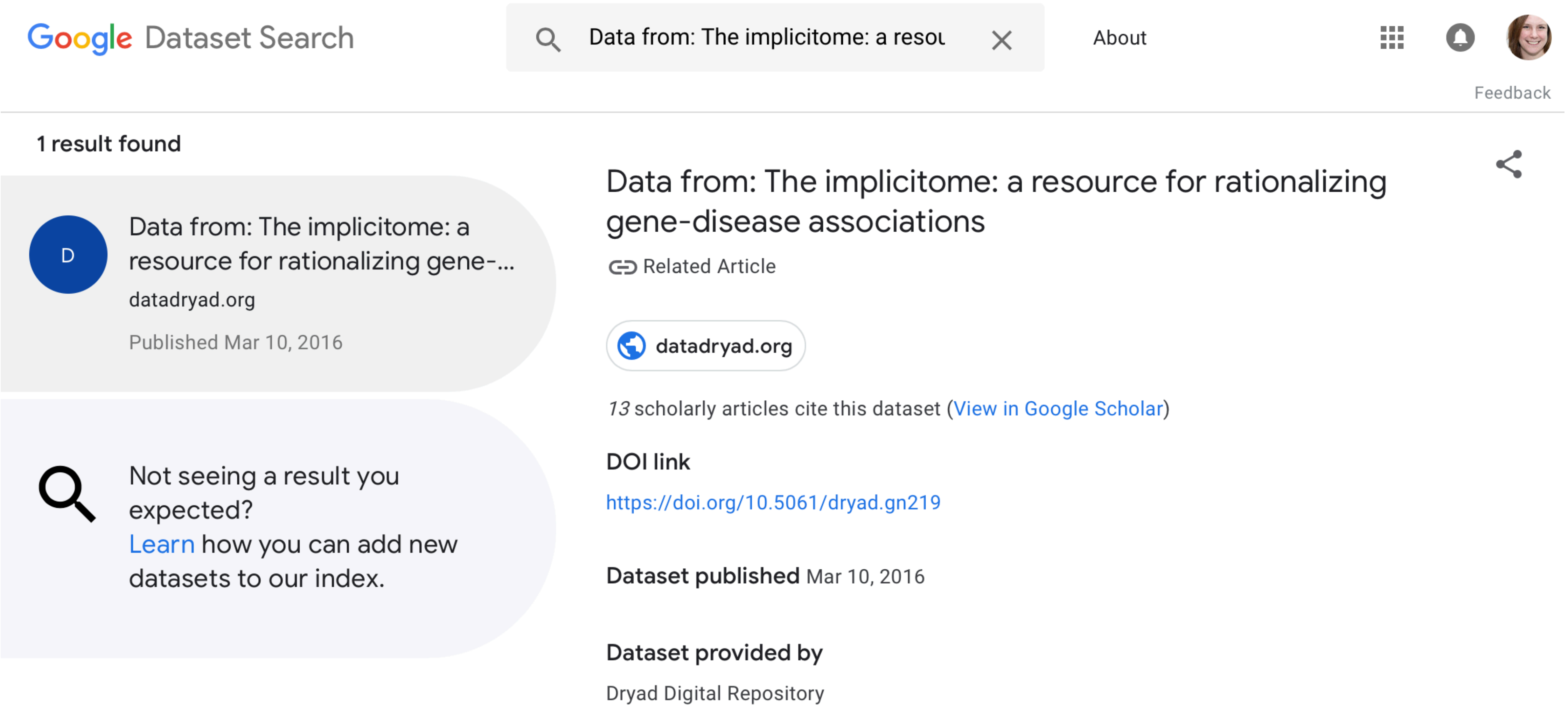
- **Data:**
- After: DataDryad
  - PLoS ONE preferred repository
- During: BeeHub (now Surfdrive)

- **Code:**
- After: GitHub
  - Pipeline: General Public License(s)
  - Figures for publication: General Public License(s)
  - MEDLINE and Thesaurus: National Library of Medicine license
- During: Local solutions

http://datadryad.org/resource/doi:10.5061/dryad.gn219
http://beehub.nl/biosemantics/gene-disease%20resources/

https://github.com/BiosemanticsDotOrg/GeneDiseasePaper

Findable by people and machines



1 result found

Data from: The implicitome: a resource for rationalizing gene-...

datadryad.org

Published Mar 10, 2016

Not seeing a result you expected? Learn how you can add new datasets to our index.

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

Related Article

datadryad.org

13 scholarly articles cite this dataset (View in Google Scholar)

DOI link

https://doi.org/10.5061/dryad.gn219

Dataset published Mar 10, 2016

Dataset provided by

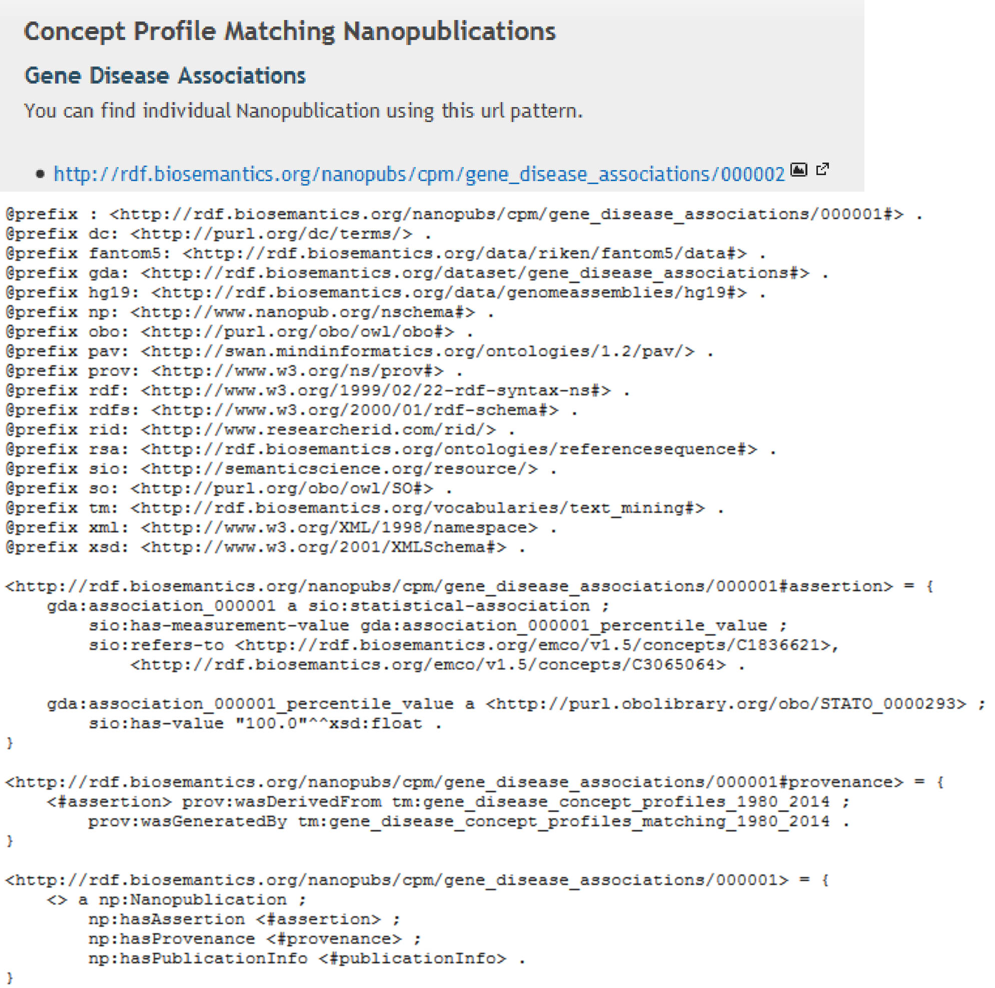
Dryad Digital Repository

Data modelling and reuse (I, R)

- CSV with identifiers
- Machine-readable Nanopublications
  - rdf.biosemantics.org
- SCRIPPS: knowledge.bio
  - Bitbucket
  - MIT license
  - Database and web interface

http://rdf.biosemantics.org/

https://bitbucket.org/sulab/kb1



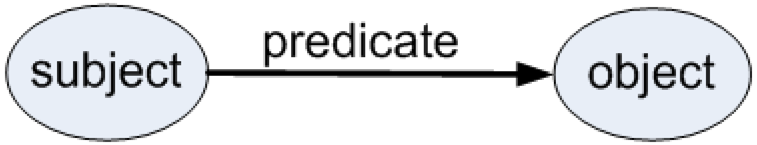
Concept Profile Matching Nanopublications

Gene Disease Associations

You can find individual Nanopublication using this url pattern.

- http://rdf.biosemantics.org/nanopubs/cpm/gene_disease_associations/000002

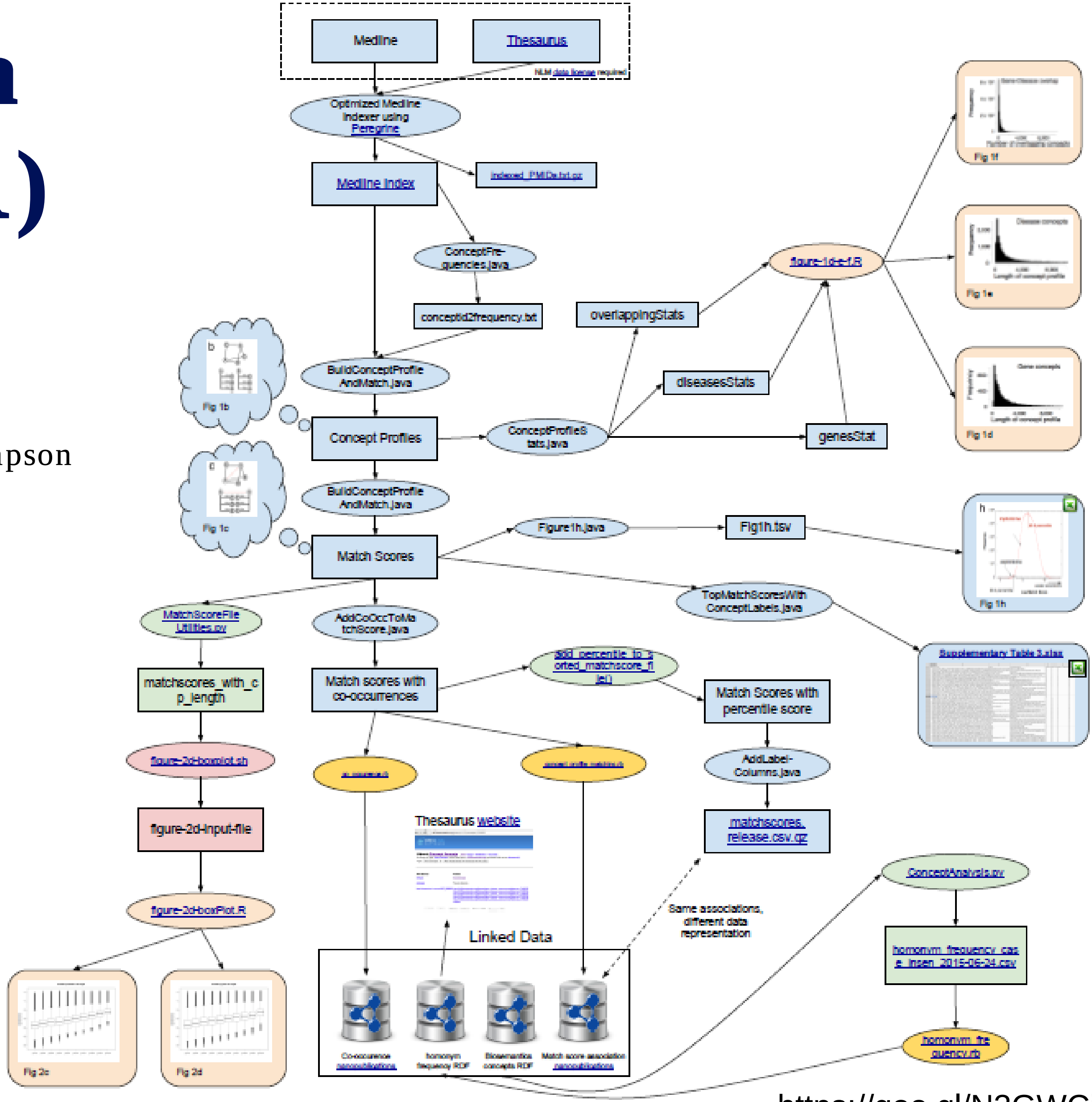
```
@prefix : <http://rdf.biosemantics.org/nanopubs/cpm/gene_disease_associations/000001#> .
@prefix dc: <http://purl.org/dc/terms/> .
@prefix fantom5: <http://rdf.biosemantics.org/data/riken/fantom5/data#> .
@prefix gda: <http://rdf.biosemantics.org/dataset/gene_disease_associations#> .
@prefix hg19: <http://rdf.biosemantics.org/data/genomeassemblies/hg19#> .
@prefix np: <http://www.nanopub.org/nschema#> .
@prefix obo: <http://purl.org/obo/owl/obo#> .
@prefix pav: <http://swan.mindinformatics.org/ontologies/1.2/pav/> .
@prefix prov: <http://www.w3.org/ns/prov#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix rid: <http://www.researcherid.com/rid/> .
@prefix rsa: <http://rdf.biosemantics.org/ontologies/referencesequence#> .
@prefix sio: <http://semanticscience.org/resource/> .
@prefix so: <http://purl.org/obo/owl/SO#> .
@prefix tm: <http://rdf.biosemantics.org/vocabularies/text_mining#> .
@prefix xml: <http://www.w3.org/XML/1998/namespace> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

<http://rdf.biosemantics.org/nanopubs/cpm/gene_disease_associations/000001#assertion> = {
    gda:association_000001 a sio:statistical-association ;
        sio:has-measurement-value gda:association_000001_percentile_value ;
        sio:refers-to <http://rdf.biosemantics.org/emco/v1.5/concepts/C1836621>,
            <http://rdf.biosemantics.org/emco/v1.5/concepts/C3065064> .

    gda:association_000001_percentile_value a <http://purl.obolibrary.org/obo/STATO_0000293> ;
        sio:has-value "100.0"^^xsd:float .
}

<http://rdf.biosemantics.org/nanopubs/cpm/gene_disease_associations/000001#provenance> = {
    <#assertion> prov:wasDerivedFrom tm:gene_disease_concept_profiles_1980_2014 ;
        prov:wasGeneratedBy tm:gene_disease_concept_profiles_matching_1980_2014 .
}

<http://rdf.biosemantics.org/nanopubs/cpm/gene_disease_associations/000001> = {
    <> a np:Nanopublication ;
        np:hasAssertion <#assertion> ;
        np:hasProvenance <#provenance> ;
        np:hasPublicationInfo <#publicationInfo> .
}
```



This Photo by Unknown Author is licensed under CC BY-SA-NC

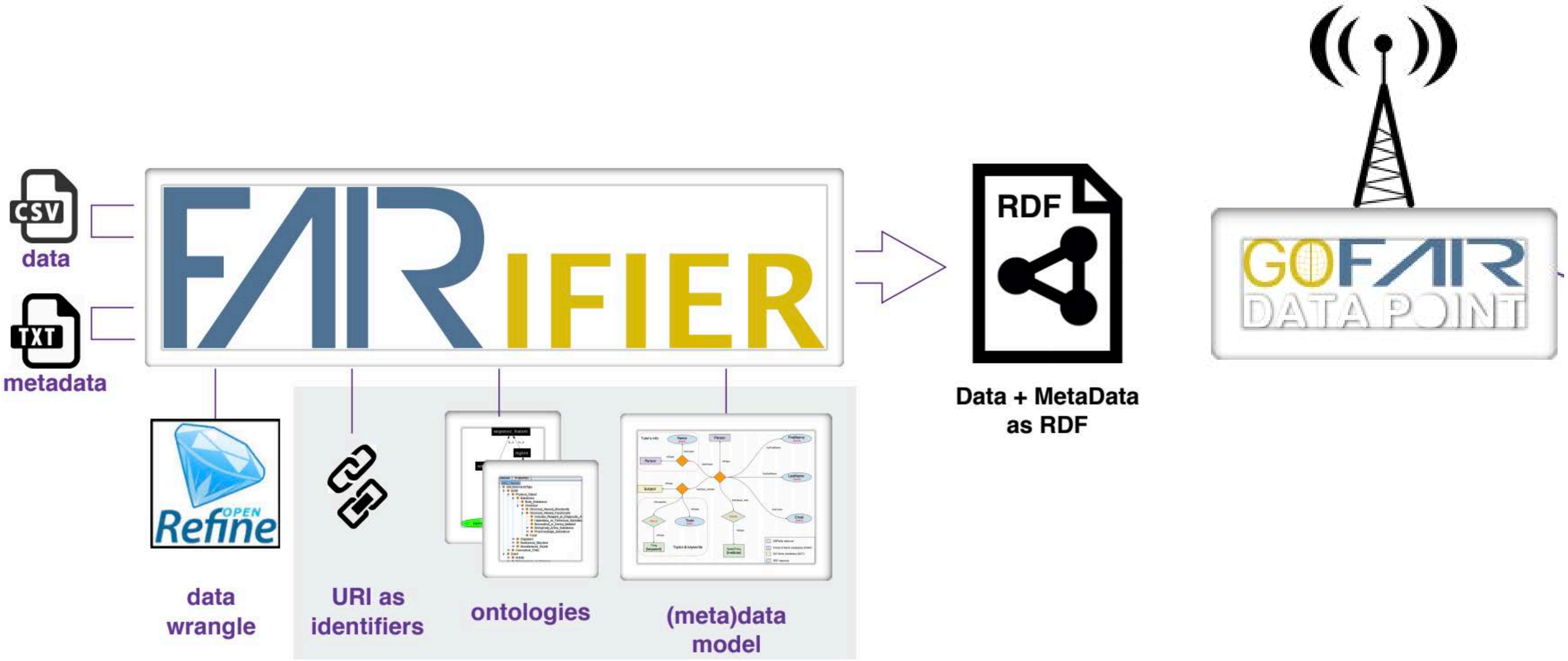
Code and data overview (I, R)

By Mark Thompson

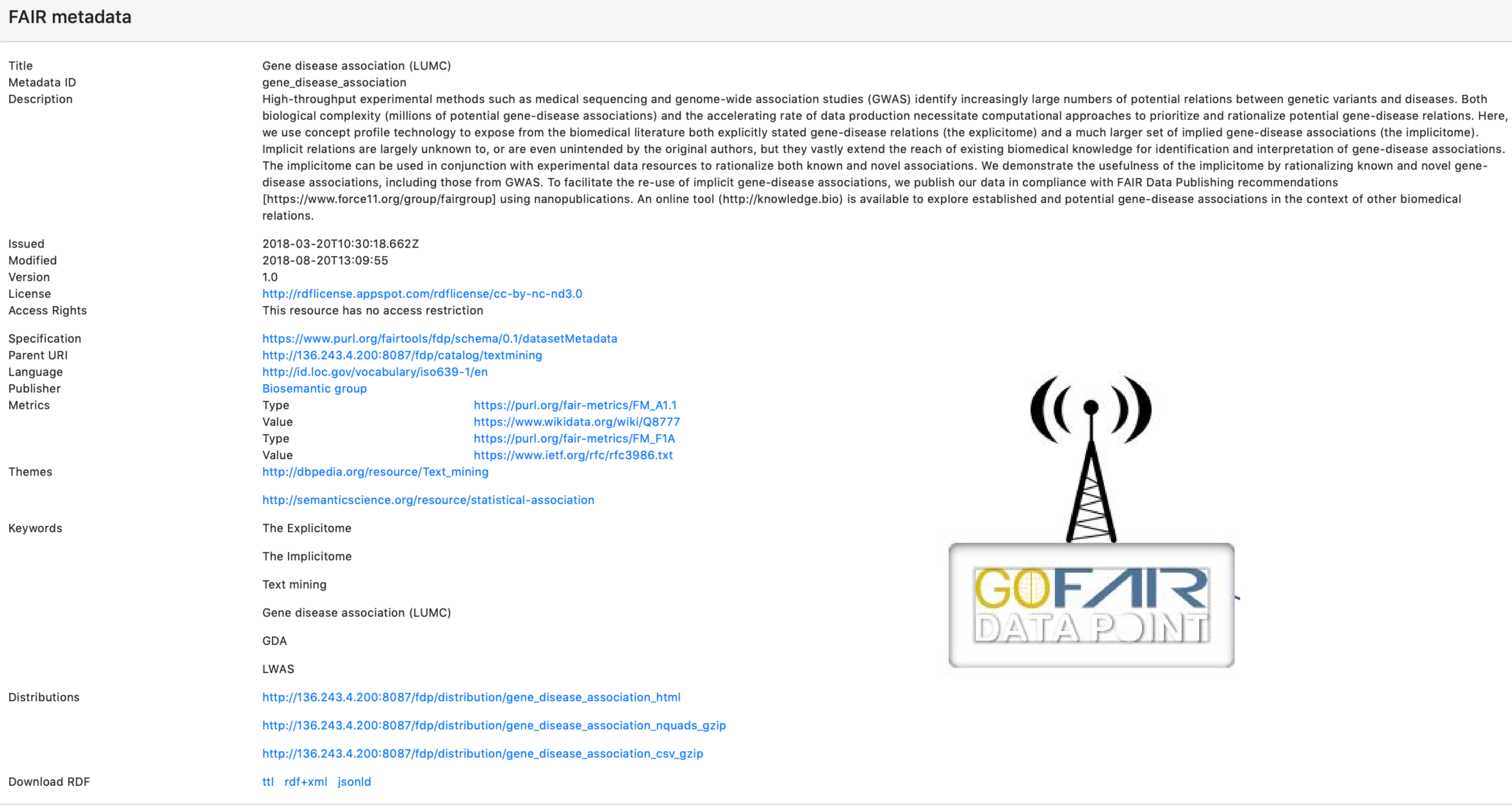

Medline
Thesaurus
NLM data license required
Optimized Medline Indexer using Peregrine
Medline Index
indexed_PMIDs.txt.gz
ConceptFre-quencies.java
conceptid2frequency.txt
BuildConceptProfile AndMatch.java
Fig 1b
Concept Profiles
ConceptProfileS tats.java
overlappingStats
diseasesStats
genesStat
figure-1d-e-f.R
Fig 1f
Fig 1e
Fig 1d
BuildConceptProfile AndMatch.java
Fig 1c
Match Scores
Figure1h.java
Fig1h.tsv
Fig 1h
TopMatchScoresWith ConceptLabels.java
Supplementary Table 3.xlsx
MatchScoreFile Utilities.py
AddCoOccToMa tchScore.java
matchscores_with_c p_length
Match scores with co-occurrences
add_percentile_to_s orted_matchscore_fi le()
Match Scores with percentile score
AddLabel-Columns.java
matchscores. release.csv.gz
figure-2d-boxplot.sh
figure-2d-input-file
figure-2d-boxPlot.R
Fig 2c
Fig 2d
Thesaurus website
Linked Data
Same associations, different data representation
Co-occurence nanopublications
homonym frequency RDF
Biosemantics concepts RDF
Match score association nanopublications
ConceptAnalysis.py
homonym_frequency_cas e_insen_2015-06-24.csv
homonym_fre quency.rb

https://goo.gl/N3GWCi

NOW: FAIRification workflow



Workflow picture from Erik Schultes, GO FAIR



FAIR metadata

Title	Gene disease association (LUMC)
Metadata ID	gene_disease_association
Description	High-throughput experimental methods such as medical sequencing and genome-wide association studies (GWAS) identify increasingly large numbers of potential relations between genetic variants and diseases. Both biological complexity (millions of potential gene-disease associations) and the accelerating rate of data production necessitate computational approaches to prioritize and rationalize potential gene-disease relations. Here, we use concept profile technology to expose from the biomedical literature both explicitly stated gene-disease relations (the explicitome) and a much larger set of implied gene-disease associations (the implicitome). Implicit relations are largely unknown to, or are even unintended by the original authors, but they vastly extend the reach of existing biomedical knowledge for identification and interpretation of gene-disease associations. The implicitome can be used in conjunction with experimental data resources to rationalize both known and novel associations. We demonstrate the usefulness of the implicitome by rationalizing known and novel gene-disease associations, including those from GWAS. To facilitate the re-use of implicit gene-disease associations, we publish our data in compliance with FAIR Data Publishing recommendations [https://www.force11.org/group/fairgroup] using nanopublications. An online tool (http://knowledge.bio) is available to explore established and potential gene-disease associations in the context of other biomedical relations.
Issued	2018-03-20T10:30:18.662Z
Modified	2018-08-20T13:09:55
Version	1.0
License	http://rdflicense.appspot.com/rdflicense/cc-by-nc-nd3.0
Access Rights	This resource has no access restriction
Specification	https://www.purl.org/fairtools/fdp/schema/0.1/datasetMetadata
Parent URI	http://136.243.4.200:8087/fdp/catalog/textmining
Language	http://id.loc.gov/vocabulary/iso639-1/en
Publisher	Biosemantic group
Metrics	Type https://purl.org/fair-metrics/FM_A1.1 Value https://www.wikidata.org/wiki/Q8777 Type https://purl.org/fair-metrics/FM_F1A Value https://www.ietf.org/rfc/rfc3986.txt
Themes	http://dbpedia.org/resource/Text_mining http://semanticscience.org/resource/statistical-association
Keywords	The Explicitome The Implicitome Text mining Gene disease association (LUMC) GDA LWAS
Distributions	http://136.243.4.200:8087/fdp/distribution/gene_disease_association_html http://136.243.4.200:8087/fdp/distribution/gene_disease_association_nquads_gzip http://136.243.4.200:8087/fdp/distribution/gene_disease_association_csv_gzip
Download RDF	ttl rdf+xml jsonld

http://136.243.4.200:8087/fdp/dataset/gene_disease_association

Take home messages

- FAIRification takes time and effort (~20% of research time first time, ~5% second time), thus plan enough time for it and start early!
- Quick wins:
  - F: Put your data and code in a trusted repository
  - A: Make sure there is a data access policy for the repository
  - I: Describe your data using data and metadata standards
  - R: Choose a license for your data and code
- For the pioneers:
  - I, R: Create a data model
  - F, I, R: Describe your data in triple (RDF) format with persistent identifiers

Tip for discipline-specific guides: Top 10 FAIR Data and Software Things
https://doi.org/10.5281/zenodo.2555498

Acknowledgements

EMC Biosemantics group + alumni

LUMC Biosemantics group + alumni

GO FAIR

Leiden University Libraries

Universiteit
Leiden
The Netherlands