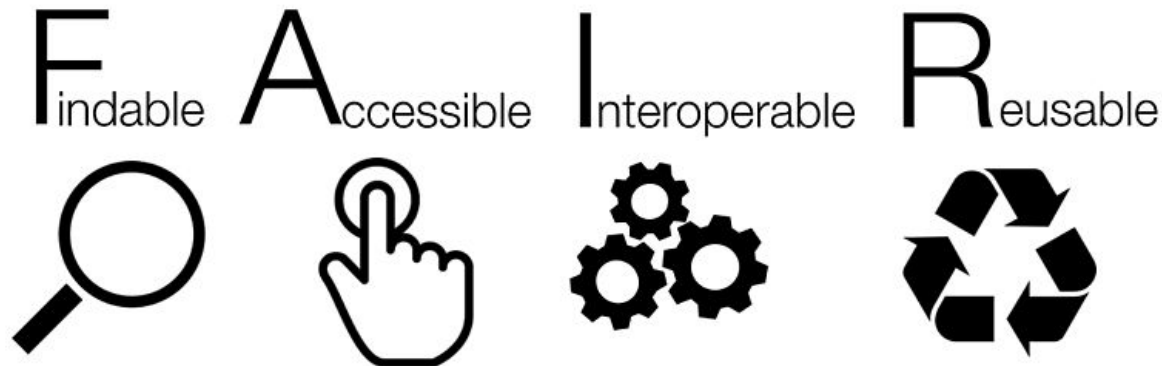


# Accelerating biomedical discovery with an Internet of FAIR data and services



**Michel Dumontier, Ph.D.**  
Distinguished Professor of Data Science  
Director, Institute of Data Science



**Maastricht University**

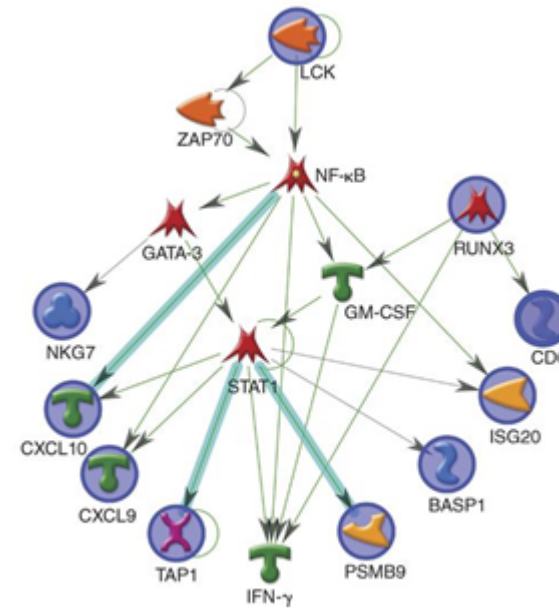
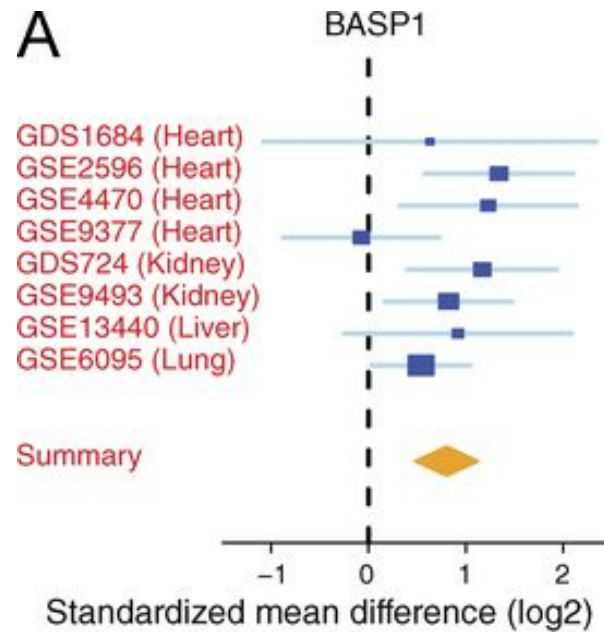


**An increasing number of discoveries  
are made using *already* available data**

# Common rejection module (CRM) for acute rejection across multiple organs identifies novel therapeutics for organ transplantation

Maatri et al. JEM. 210 (11): 2205

DOI: 10.1084/jem.20122709



## Main Findings:

CRM genes **predicted future injury** to a graft

Mice treated with **drugs against the CRM genes extended graft survival**

Retrospective **EHR analysis supports treatment prediction**

## Key Observations:

**Meta-analysis** offers a **more reliable estimate** of the magnitude of the effect

Data can be used to **generate and support/dispute new hypotheses**

However, *significant effort* is still needed to find the right dataset(s), make sense of them, and use for a new purpose

# How do you find your data?

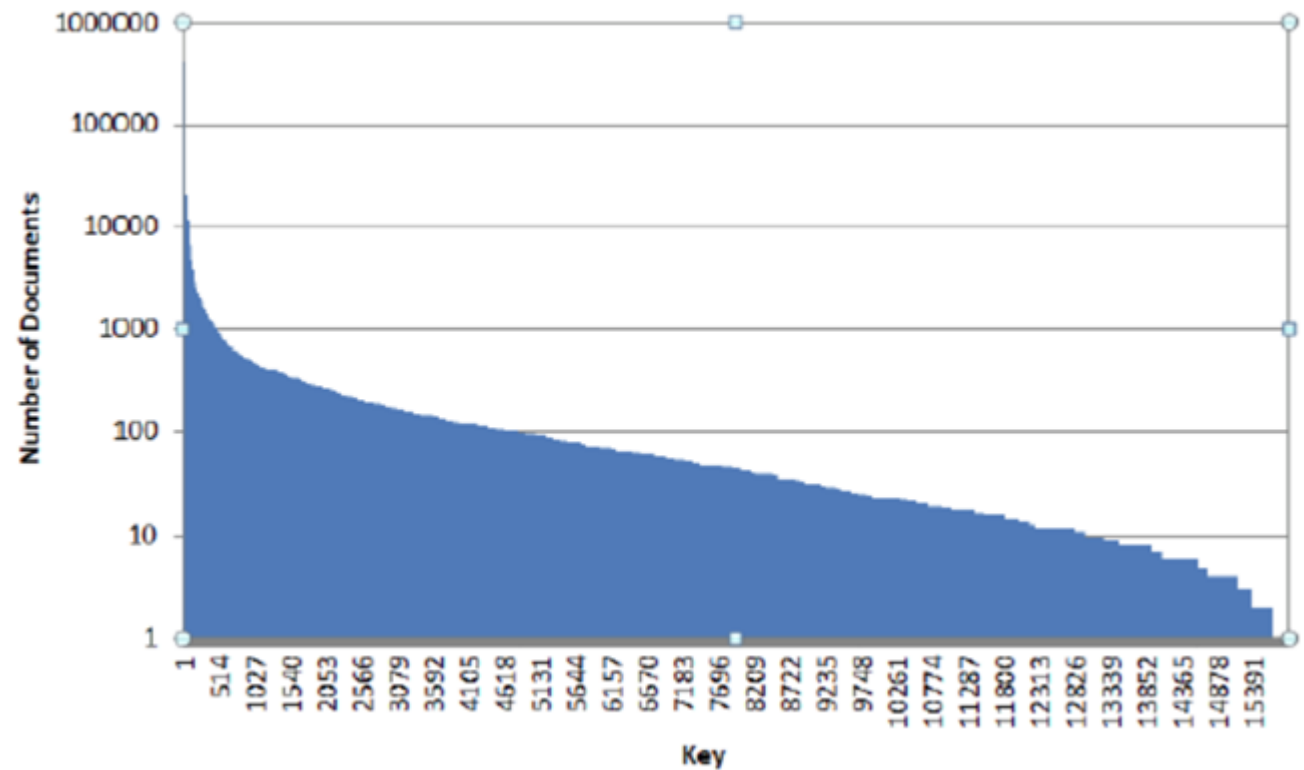
- Datasets you learned about in yesterday's tutorials and workshops
- Datasets used in the lab or organization
- Datasets associated with a paper you read or were told about
- A search in a specific repository
- A search on the internet

# Poor quality metadata frustrates reuse

age	207147
Age	18089
age (yrs)	9891
age (years)	9272
age (y)	6226
age in years	1387
age_years	607
AGE	588
age(years)	558
age (year)	433
age (yr)	373
Age (years)	318
age (in years)	310
Age(years)	267
age [year]	97
age [y]	84
age [years]	83
Age(yrs.)	81
age.year	70
age (yr-old)	64
age(yrs)	59
age of patient	40
Age, year	39
Age (yrs)	36
Age of patient	33
age, years	24
'Age	21
Age (Years)	20
age (after birth)	18
age, yrs	12
age of subjects	4



*Vast number of lexically unique keys (and values)*



# Our ability to reproduce landmark studies is surprisingly low:

**39%** (39/100) in psychology<sup>1</sup>

**21%** (14/67) in pharmacology<sup>2</sup>

**11%** (6/53) in cancer<sup>3</sup>

**unsatisfactory** in machine learning<sup>4</sup>

<sup>1</sup>[doi:10.1038/nature.2015.17433](https://doi.org/10.1038/nature.2015.17433) <sup>2</sup>[doi:10.1038/nrd3439-c1](https://doi.org/10.1038/nrd3439-c1) <sup>3</sup>[doi:10.1038/483531a](https://doi.org/10.1038/483531a) <sup>4</sup><https://openreview.net/pdf?id=By4l2PbQ->

## Most published research findings are false.

- John Ioannidis, Stanford University

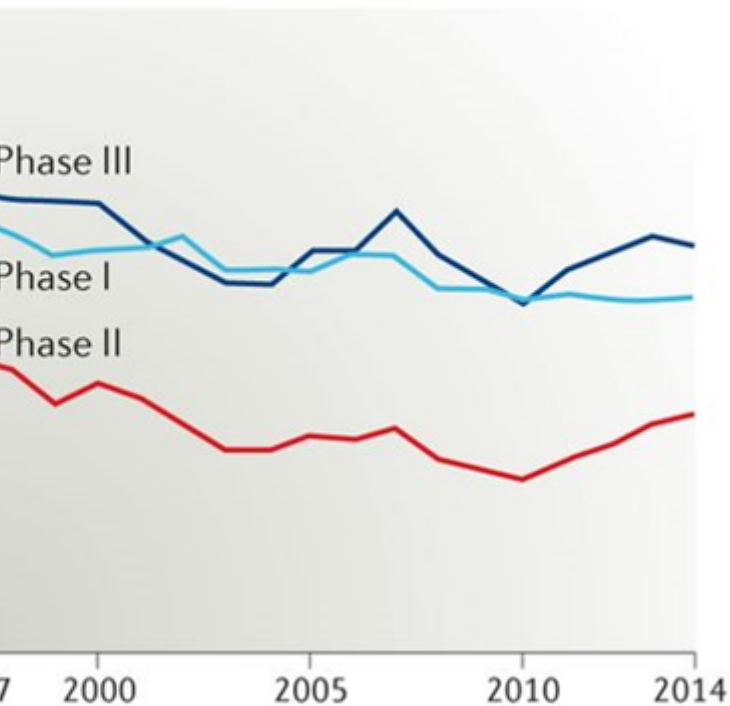
PLoS Med 2005;2(8): e124.

# CLINICAL-TRIAL CLIFF

Companies are removing more compounds from the pipeline at all levels of testing than ever before.

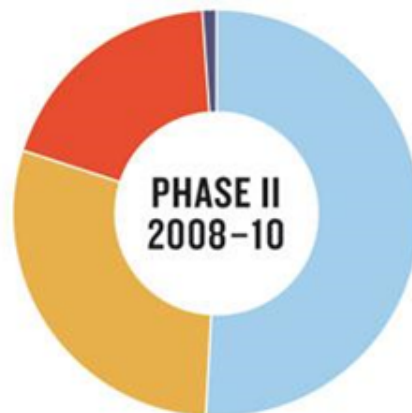
## Success rates by phase

Percentage likelihood of moving to next phase, 3-year rolling average\*



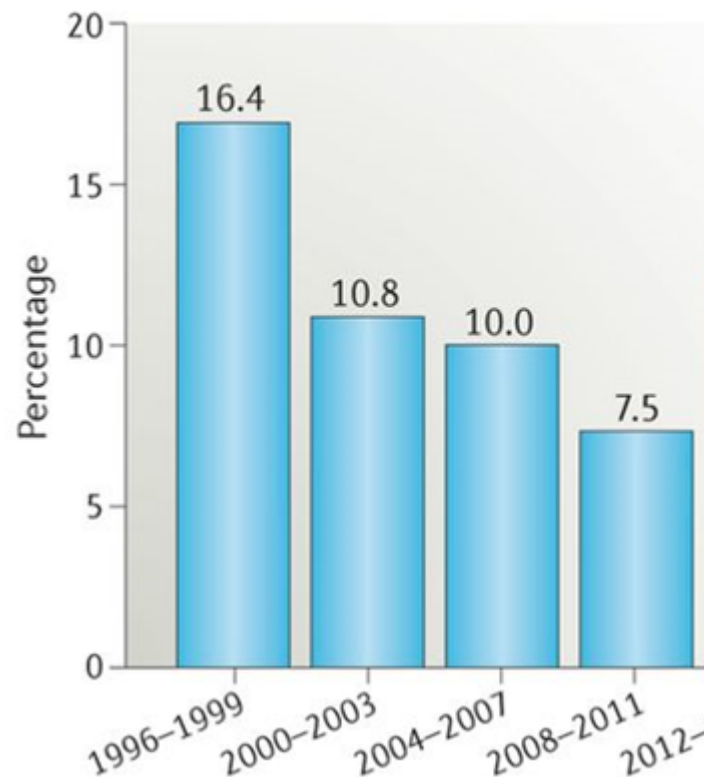
Most of the product failures in phase II and III trials are because researchers are unable to demonstrate efficacy or sufficient safety.

- Efficacy
- Safety
- Strategic
- Pharmacokinetics/ bioavailability
- Commercial/ financial
- Not disclosed

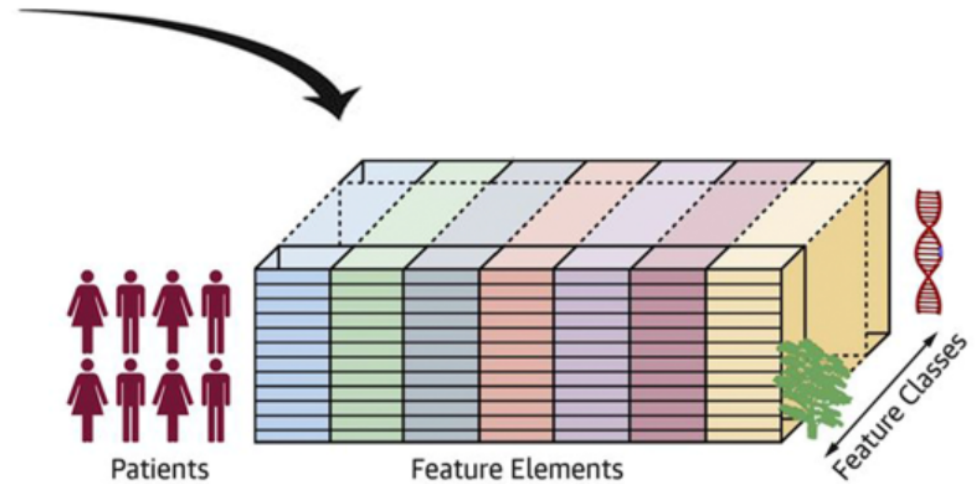
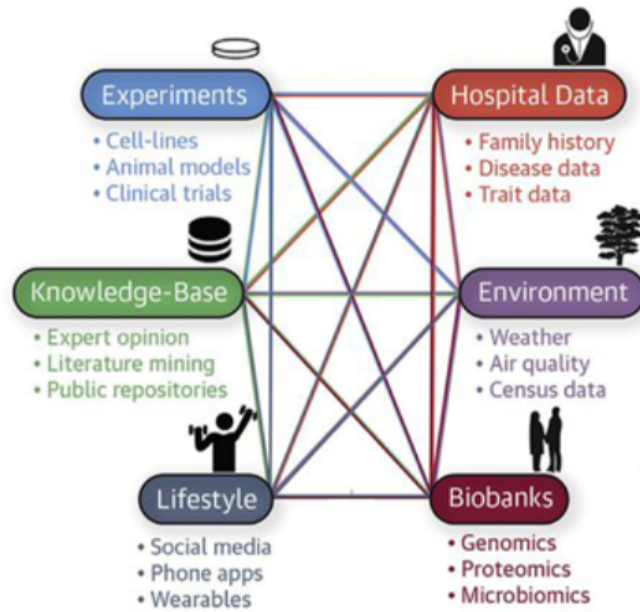


## Cumulative success rate Phase I to launch

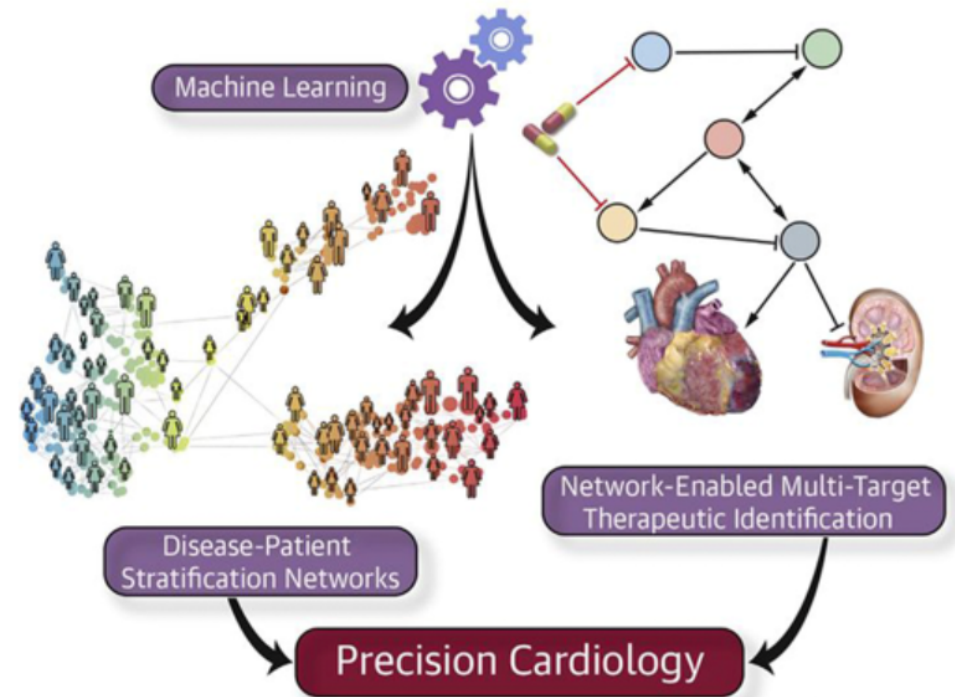
Percentage likelihood of moving from Phase I to launch







What do we really have to realize  
**precision Medicine?**



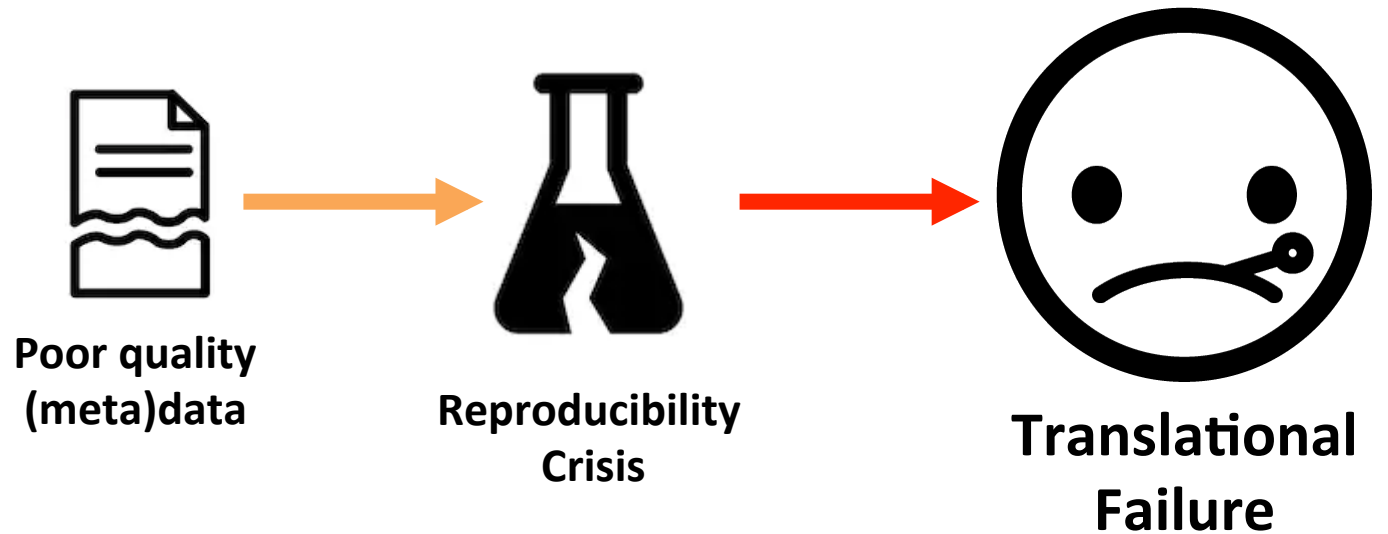
### Broken windows theory

Visible signs of crime, anti-social behavior, and civil disorder create an urban environment that encourages further crime and disorder, including serious crimes

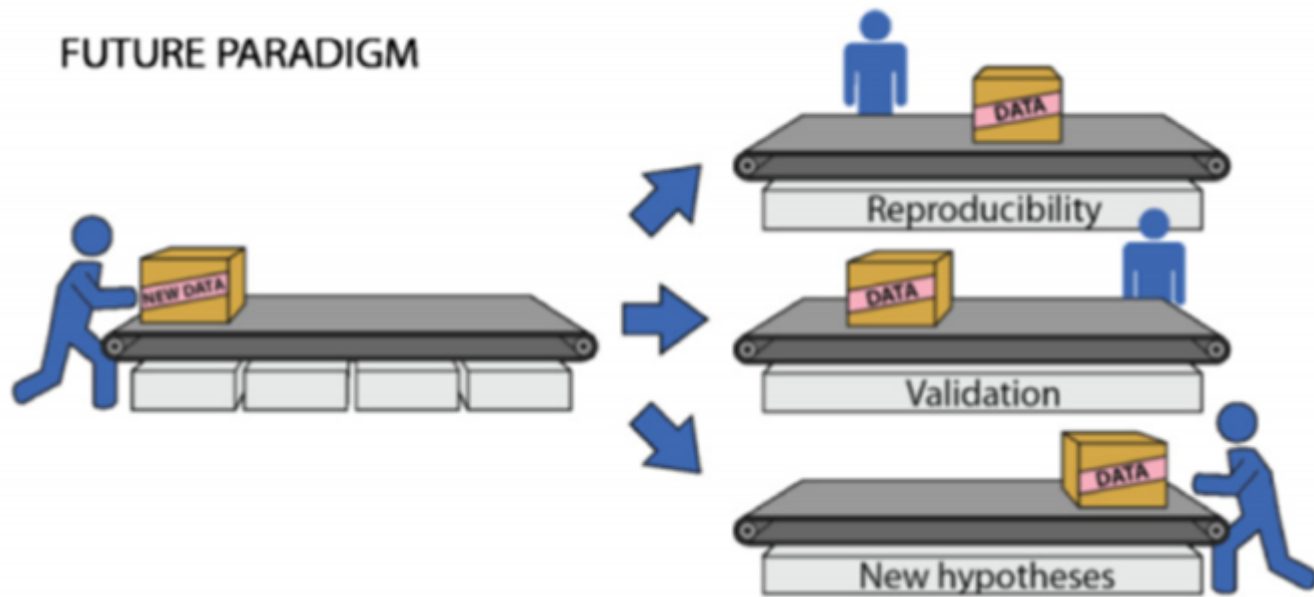
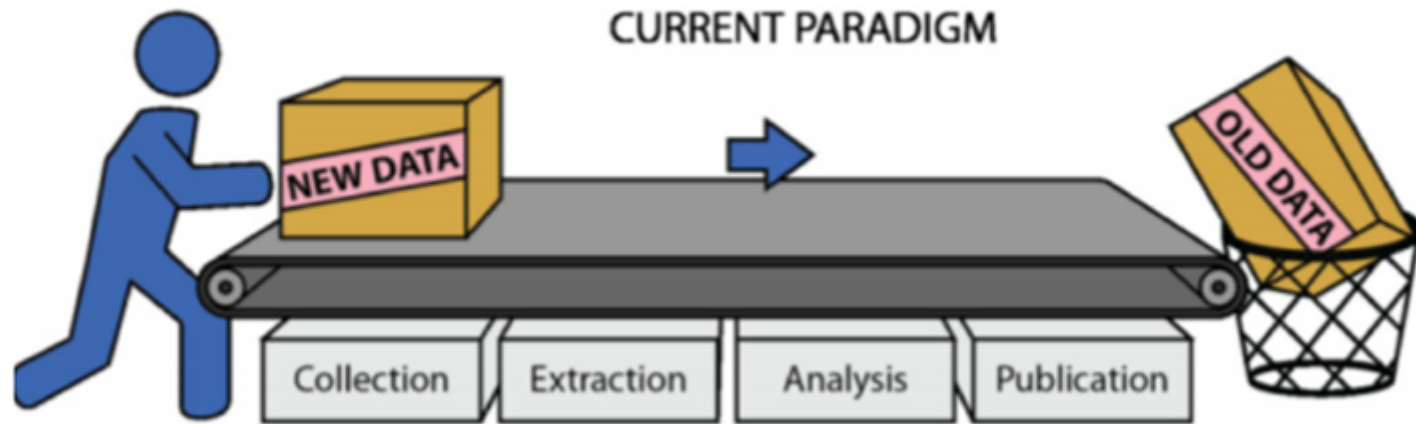


### Inadequate reusability theory

Poor quality metadata and the inaccessibility of original research results make it less likely to produce original work, resulting in an ineffective translation of research into useful applications



**It's time to completely rethink  
how we perform **research****

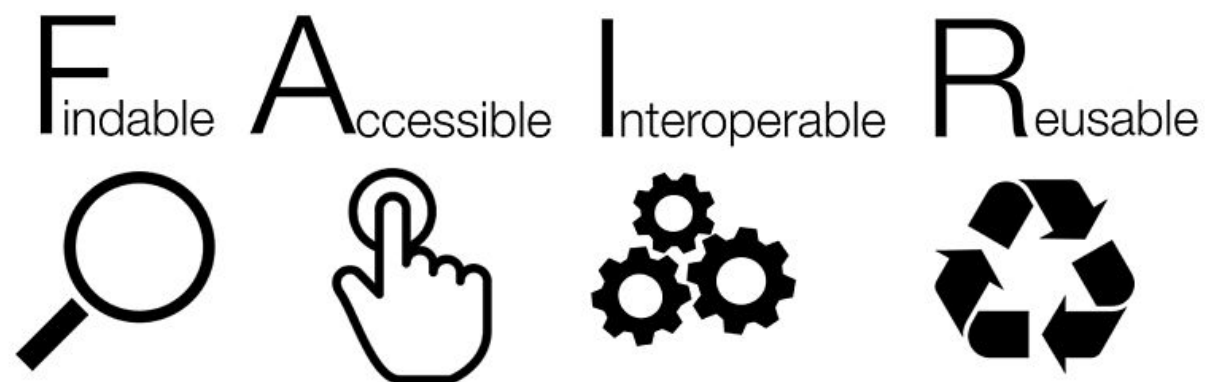


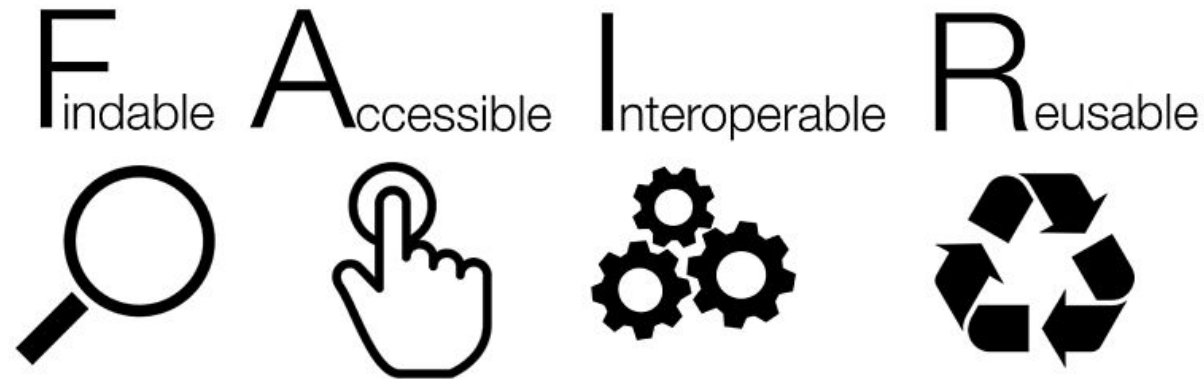
Lambin et al. Radiother Oncol. 2013. 109(1):159-64. doi: 10.1016/j.radonc.2013.07.007

# Human Machine collaboration will be crucial to our future success



We need a new *social contract*, supported by *legal* and *technological* infrastructure to make digital resources available in a responsible manner





**An international, bottom-up paradigm for the discovery and reuse of digital content  
*for the machines that people use***





# The FAIR Guiding Principles for scientific data management and stewardship

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

[Citations](#) | [Contributions](#) | [Corresponding author](#)

*Scientific Data* **3**, Article number: 160018 (2016) | doi:10.1038/sdata.2016.18

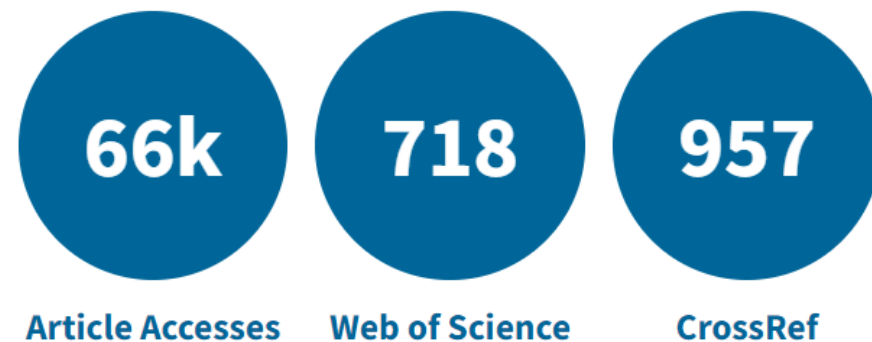
Received 10 December 2015 | Accepted 12 February 2016 | Published online 15

March 2016

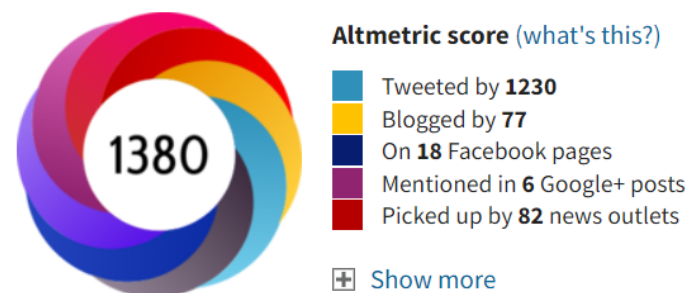
<https://www.nature.com/articles/sdata201618>

Last updated: Mon, 9 Sep 2019 14:18:07 GMT

## Total citations



## Online attention



### This Altmetric score means that the article is:

- in the 99<sup>th</sup> percentile (ranked 76<sup>th</sup>) of the 264,573 tracked articles of a similar age in all journals
- in the 1<sup>st</sup> percentile (ranked 1<sup>st</sup>) of the 1 tracked articles of a similar age in *Scientific Data*

@micheldumontier::Semantics:2

# R: Impact



EUROPEAN COMMISSION  
Press Release Database

European Commission > Press releases database > Press Release details

## Commission - Statement

### Leaders' Communique Hangzhou Summit

5 September 2016

Leaders of the G20, met in Hangzhou, China on 4-5 September 2016.

To drive innovation-driven growth and the creation of innovative ecosystems, we support dialogue and cooperation on innovation, which spans a wide range of domains with science and technology innovation at its core. We deliver the G20 2016 Innovation Action Plan. We pursue pro-innovation strategies and policies, support investment in science, technology and innovation (STI), and support skills development in STI - including support for the entry of more women into these fields - and mobility of STI human resources. We support effort to promote voluntary knowledge diffusion and technology transfer on mutually agreed terms and conditions. **Consistent with this approach, we support appropriate efforts to promote open science and facilitate appropriate access to publicly funded research results on findable, interoperable and reusable (FAIR) principles.** In furtherance of the above, we emphasize the importance of open trade and investment regimes to facilitate innovation through intellectual property rights (IPR) protection, and improving public communication in science and technology. We are committed to foster exchange of knowledge and experience by supporting an online G20 Community of Leaders within the existing Innovation Policy Platform and the release of the 2016 G20 Innovation Report.



Realising the European Open Science Cloud

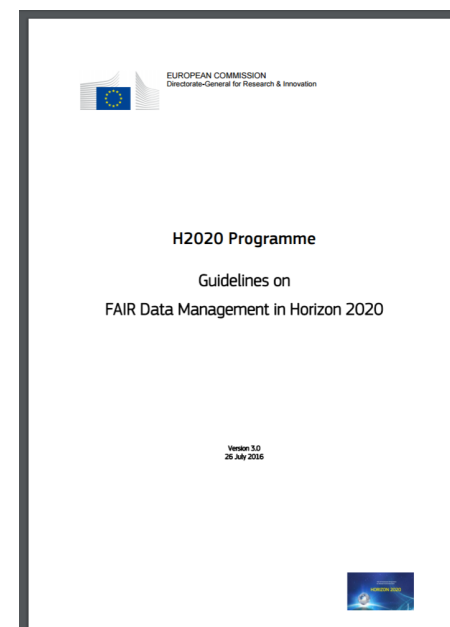
First report and recommendations of the Commission High Level Expert Group on the European Open Science Cloud

DATA SHARING  
LINKING DATA  
CONNECTING SCIENTISTS  
SUSTAINABLE  
OPEN SERVICES  
CONNECTING DISCIPLINES  
BETTER SCIENCE  
RESEARCH AND INNOVATION



European Commission

Final Report  
Com  
F  
RESEARCH AND INNOVATION



EUROPEAN COMMISSION  
Directorate-General for Research & Innovation

H2020 Programme  
Guidelines on  
FAIR Data Management in Horizon 2020

Version 3.0  
20 July 2016



NIH  
Big Data Knowledge

DCPF  
DATA COMMONS PILOT PHASE C

# FAIR in a nutshell

FAIR aims to create **social** and **economic impact** by facilitating the discovery and reuse of **digital resources** through a set of requirements:

- **unique identifiers** to retrieve all forms of digital content and knowledge
- **high quality meta(data)** to enhance discovery of digital resources
- **use of common vocabularies** to share terms and facilitate query
- **establishment of community standards** for more facile knowledge utilisation
- **detailed provenance** to provide context and reproducibility
- **registered in appropriate repositories** with high quality metadata for future content seekers
- **social and technological commitments** to realize reliable access
- **simpler terms of use** to clarify expectations and intensify innovation

## G8 science ministers statement: London, 12 June 2013

G8 science ministers written statement from their UK meeting on international issues that need global cooperation.

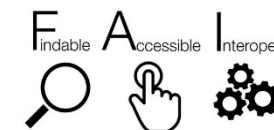
---

Published 13 June 2013

- i. To the greatest extent and with the fewest constraints possible publicly funded scientific research data should be open, while at the same time respecting concerns in relation to privacy, safety, security and commercial interests, whilst acknowledging the legitimate concerns of private partners.
- ii. Open scientific research data should be easily discoverable, accessible, assessable, intelligible, useable, and wherever possible interoperable to specific quality standards.

## FAIR != Open

*Open as possible  
closed as is necessary*



COMMENT · 04 JUNE 2019 · CORRECTION 05 JUNE 2019

# Make scientific data FAIR

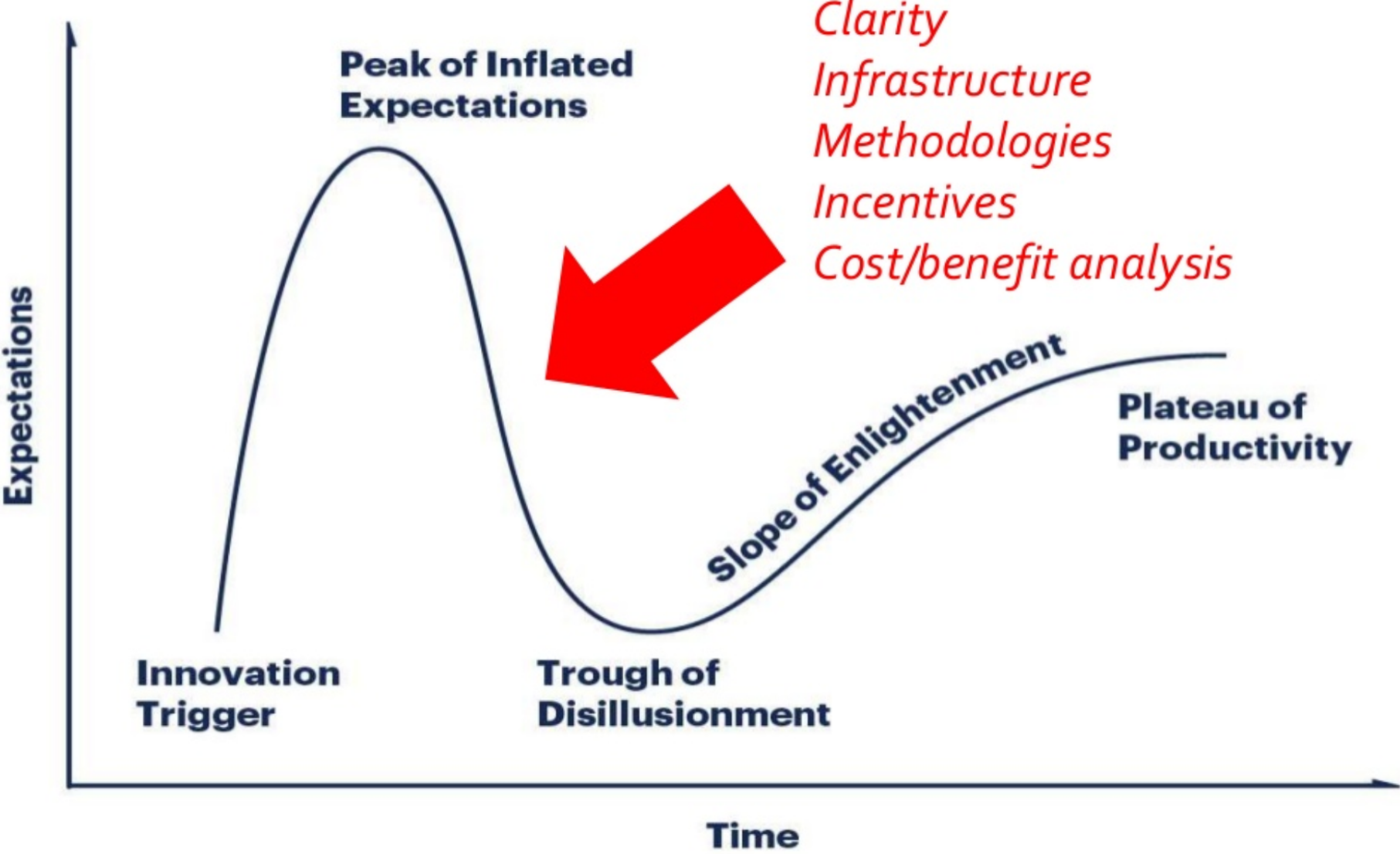
*All disciplines should follow the geosciences and demand best practice for publishing and sharing data, argue Shelley Stall and colleagues.*

---

Shelley Stall , Lynn Yarmey, Joel Cutcher-Gershenfeld, Brooks Hanson, Kerstin Lehnert, Brian Nosek, Mark Parsons, Erin Robinson & Lesley Wyborn

That's why more than 100 repositories, communities, societies, institutions, infrastructures, individuals and publishers (including the Springer Nature journals *Nature* and *Scientific Data*) have signed up since last November to the Enabling FAIR Data Project's Commitment Statement in the Earth, Space, and Environmental Sciences for depositing and sharing data (see [go.nature.com/2wv2jxd](https://go.nature.com/2wv2jxd)). The principles state that research data should be 'findable, accessible, interoperable and reusable' (FAIR)<sup>2</sup>. The idea is not new, but aligning this broad community around common data guidelines is a radical step.

# FAIR Hype Curve



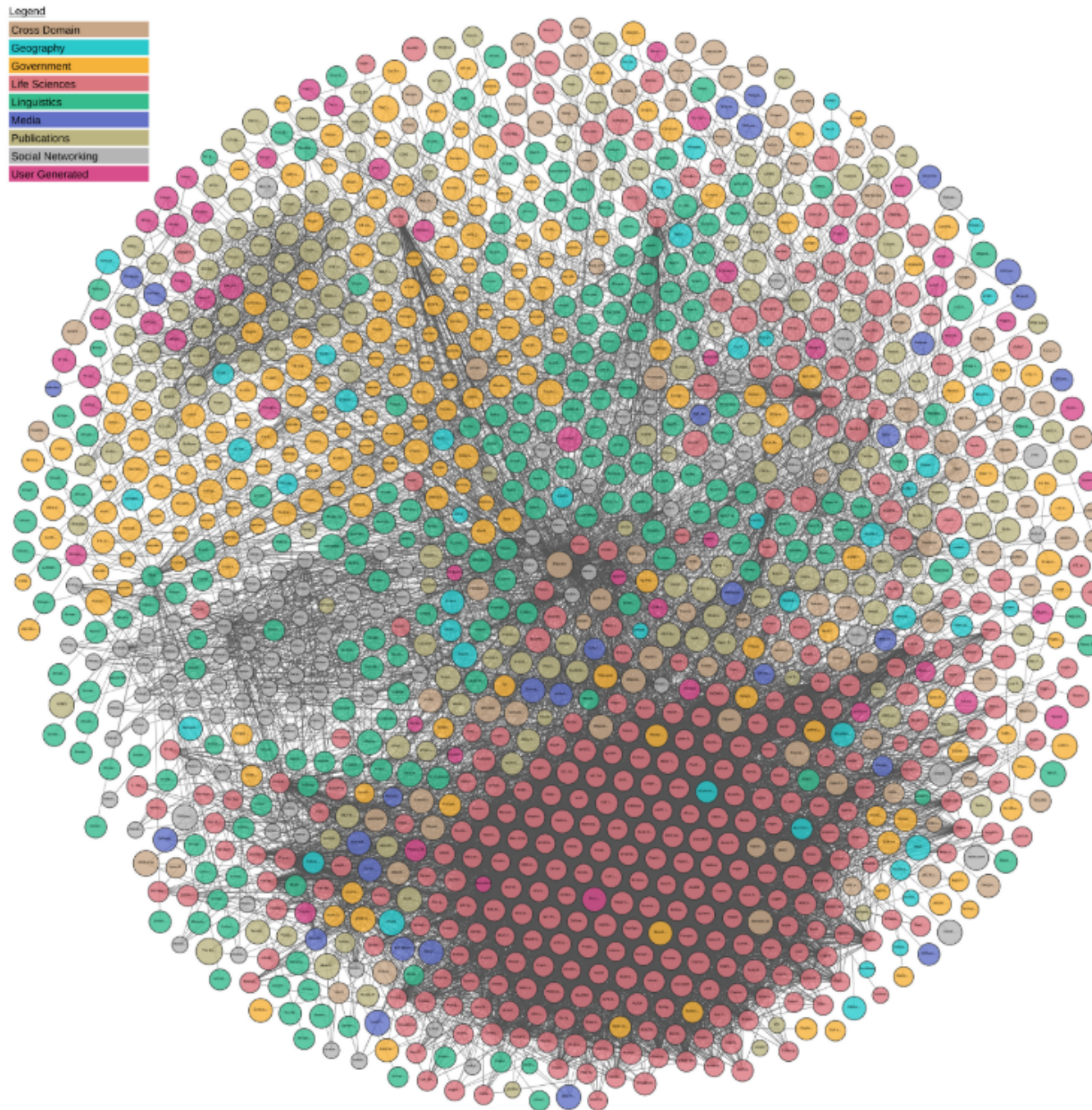
# Why Should *\*I\** Go FAIR?

- Makes it easier for **me to use my own data for a new purpose**
- Makes it easier for **other people to find, use and *cite* my data**, and for them to understand what I expect in return
- Makes it easier/possible for people to **verify my work**
- Ensure that the **data are available in the future**, especially as I may not want the responsibility
- **Satisfy expectations** around data management from institution, funding agency, journal, my peers

**Let's build the Internet of FAIR data and services**



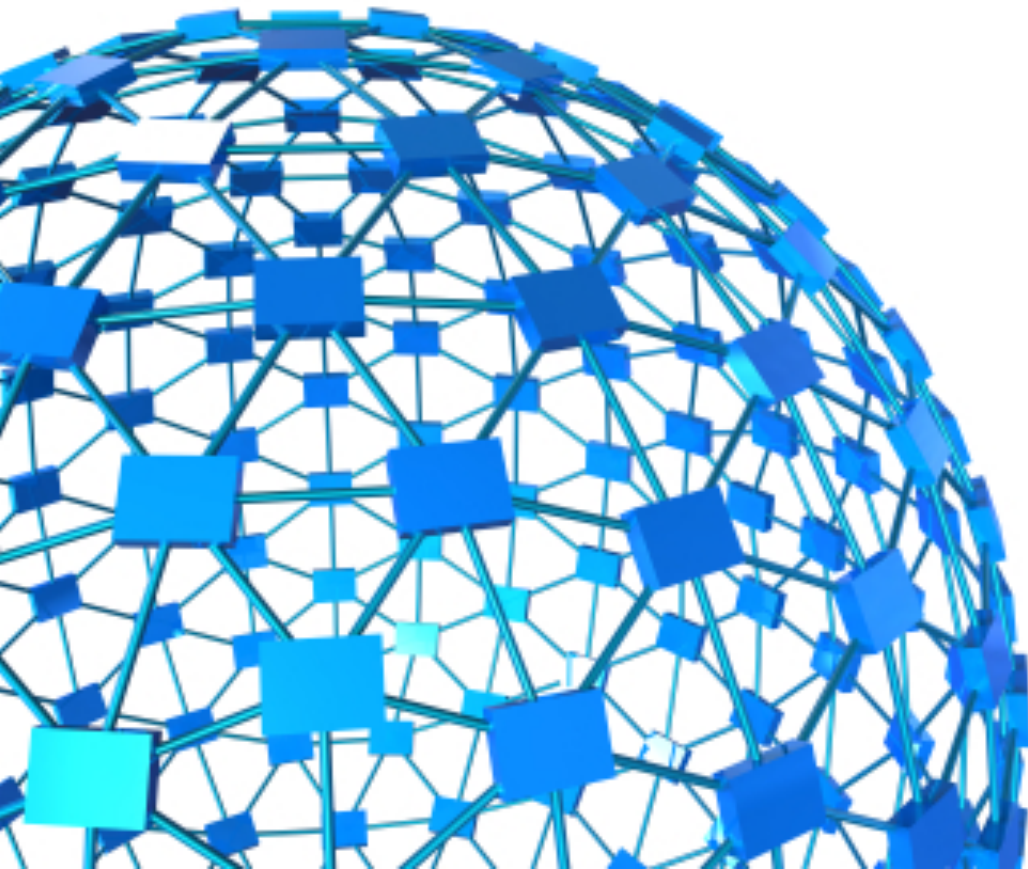
# The Linked Open Data Cloud



# The Semantic Web is a portal to the **web of knowledge**

**standards** for publishing, sharing and querying  
**facts, expert knowledge and services**

scalable approach for the discovery  
of *independently constructed,*  
*collaboratively described,*  
*distributed knowledge*  
(*in principle*)

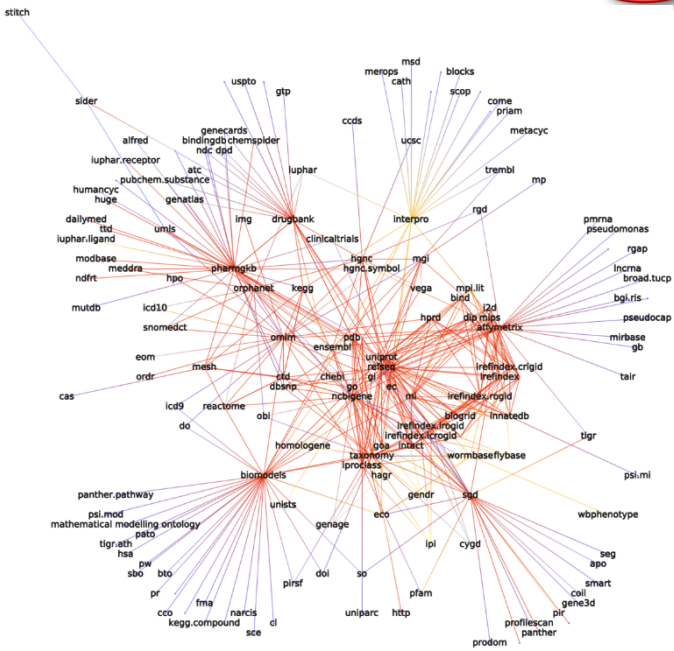
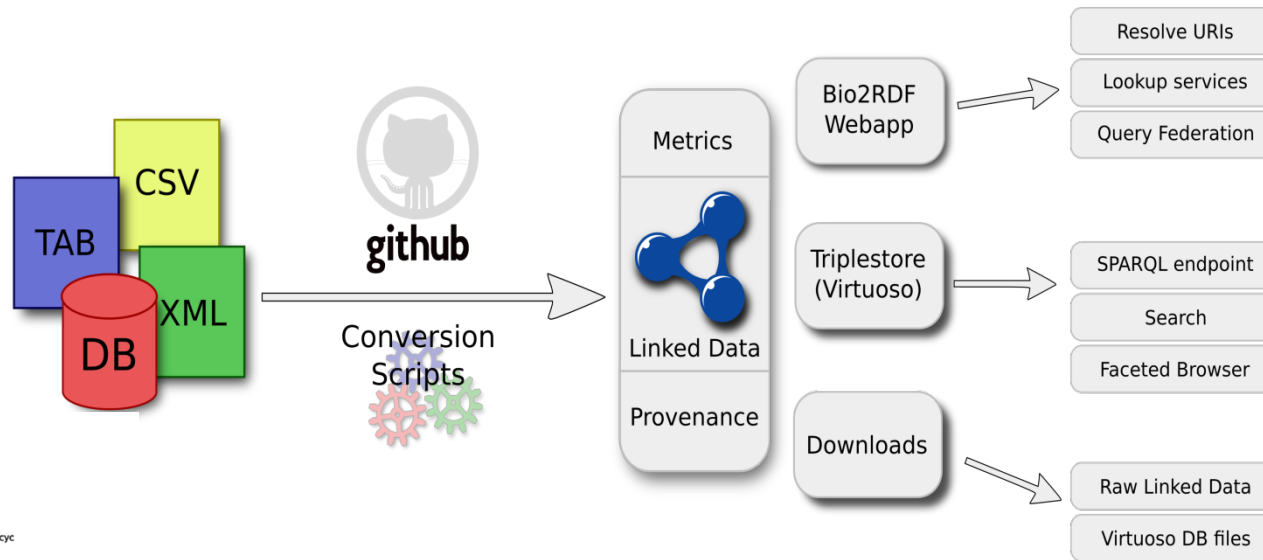


# BIO2RDF

Linked Data for the Life Sciences

Bio2RDF is an open source project that uses semantic web technologies to make it easier to reuse biomedical data

chemicals/drugs/formulations, genomes/genes/  
proteins, domains  
Interactions, complexes & pathways  
animal models and phenotypes  
Disease, genetic markers, treatments  
Terminologies & publications



- **30+** biomedical data sources
- **10B+** interlinked statements
- EBI, SIB, NCBI, DBCLS, NCBO, and many others produce this content

Alison Callahan, Jose Cruz-Toledo, Peter Ansell, Michel Dumontier: Bio2RDF Release 2: Improved Coverage, Interoperability and Provenance of Life Science Linked Data. ESWC 2013: 200-212

# Federated query over the biological web of data

otypes of  
k-out  
se models  
ne targets  
selected  
(Imatinib)

Endpoint :  Output :

```
1 PREFIX dct: <http://purl.org/dc/terms/>
2 SELECT DISTINCT ?phenotype_label
3 WHERE {
4   SERVICE <http://drugbank.bio2rdf.org/sparql> {
5     ?drug <http://bio2rdf.org/drugbank_vocabulary:target> ?target .
6     FILTER(?drug = <http://bio2rdf.org/drugbank:DB00619>)
7     ?target <http://bio2rdf.org/drugbank_vocabulary:x-hgnc> ?hgnc .
8   }
9   SERVICE <http://hgnc.bio2rdf.org/sparql> {
10    ?hgnc <http://bio2rdf.org/hgnc_vocabulary:x-mgi> ?marker .
11  }
12  SERVICE <http://mgi.bio2rdf.org/sparql> {
13    ?model <http://bio2rdf.org/mgi_vocabulary:marker> ?marker .
14    ?model <http://bio2rdf.org/mgi_vocabulary:allele> ?all .
15    ?all <http://bio2rdf.org/mgi_vocabulary:allele-attribute> ?allele_type .
16    ?model <http://bio2rdf.org/mgi_vocabulary:phenotype> ?phenotypes .
17    FILTER (str(?allele_type) = "Null/knockout")
18  }
19  SERVICE <http://bioportal.bio2rdf.org/sparql> {
20    ?phenotypes rdfs:label ?phenotype_label .
21  }
22 }
```

	phenotype_label
1	"hemorrhage [mp:0001914]"@en
2	"intracranial hemorrhage [mp:0001915]"@en
3	"perinatal lethality [mp:0002081]"@en

# Reproduce original research

Mol Syst Biol. 2011; 7: 496.

PMCID: PMC3159979

Published online 2011 Jun 7. doi: [10.1038/msb.2011.26](https://doi.org/10.1038/msb.2011.26)

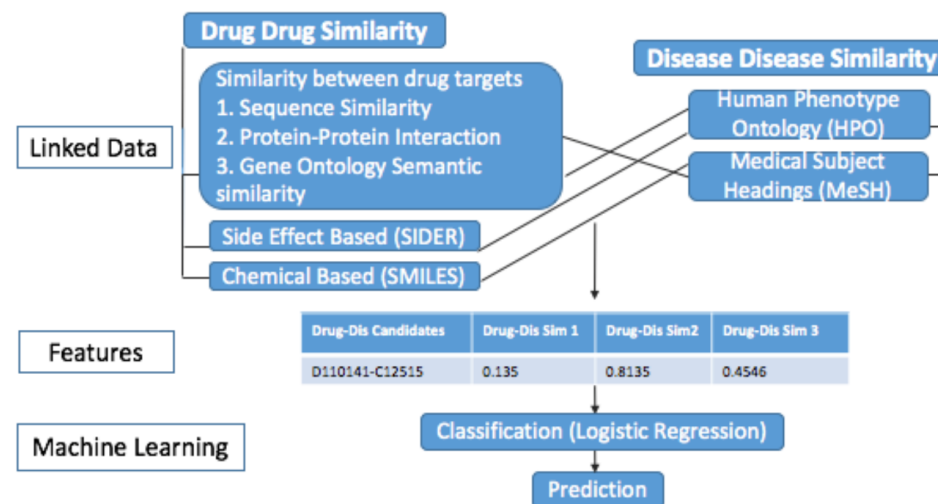
## **PREDICT: a method for inferring novel drug indications with application to personalized medicine**

[Assaf Gottlieb](#),<sup>1</sup> [Gideon Y Stein](#),<sup>2,3</sup> [Eytan Ruppin](#),<sup>1,2</sup> and [Roded Sharan](#)<sup>a,1</sup>

AUC 0.91 across all therapeutic indications

Scripts not available. Feature tables available.

BIO  RDF



**Result: ROCAUC 0.83 ... doesn't quite match**

# Efficiently explore the web of data

hide sidebar

SEARCH DRUGS WITH

0\_0000756 (23 connections)

Search

Disease Protein

Legend

Edge Details

Help

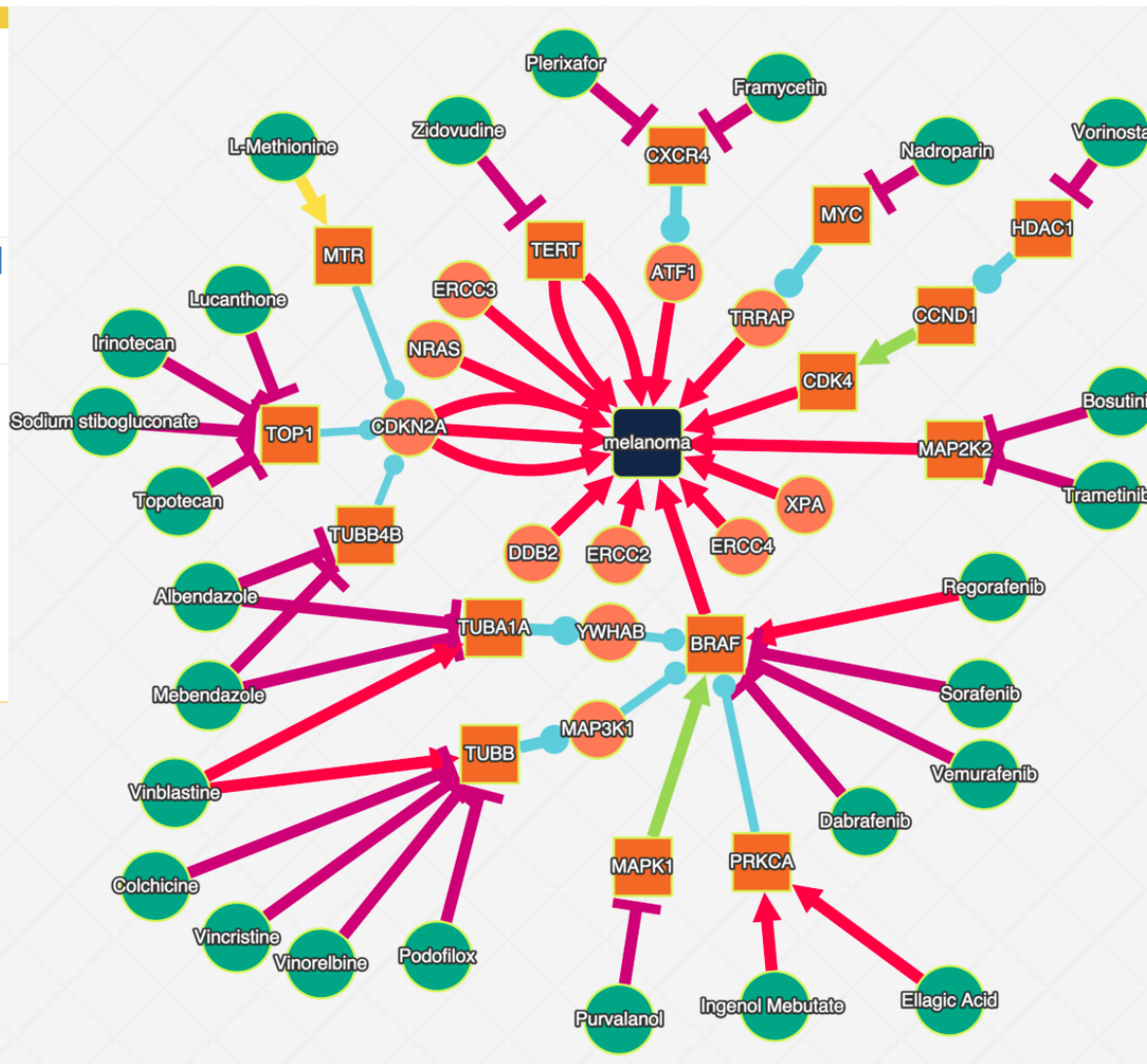
Disease

Undefined

Inhibition

Phosphorylation Reaction

Interaction



by exploring a probabilistic semantic knowledge graph

And validate them against pipelines for drug discovery

Status	Drug	Pathway	Steps
Approved	Vemurafenib <sup>2</sup>	BRAF	2
Phase III	Dabrafenib <sup>13</sup>	BRAF	2
	Sorafenib <sup>14</sup>	BRAF	2
	Vinblastine <sup>18</sup>	MAP kinase	3
	Trametinib <sup>19</sup>	MAP kinase	2
Phase II	Zidovudine <sup>29</sup>	TERT	2
	Regorafenib <sup>15</sup>	BRAF	2
	Nadroparin <sup>30</sup>	MYC	3
	Vinorelbine <sup>20</sup>	MAP kinase	3
	Irinotecan <sup>43</sup>	CDKN2A	3
	Topotecan <sup>44</sup>	CDKN2A	3
Phase I	Sodium stibogluconate <sup>45</sup>	CDKN2A	3
Case Study	Ingenol Mebutate <sup>46</sup>	PRKCA/BRAF	3
In Vitro	Bosutinib <sup>17</sup>	MAP kinase	2
	Purvalanol <sup>21</sup>	MAP kinase/TP53	3
	Ellagic Acid <sup>47</sup>	PRKCA/BRAF	3
	Albendazole <sup>48</sup>	CDKN2A	3
	Colchicine <sup>22</sup>	MAP kinase	3
	Ellagic Acid <sup>48</sup>	CDKN2A	3
In Vivo	Plerixafor <sup>27</sup>	CXCR4	3
	Vincristine <sup>23</sup>	MAP kinase	3
	L-Methionine <sup>49</sup>	CDKN2A	3
	Mebendazole <sup>50</sup>	CDKN2A	3

Finding melanoma drugs through a probabilistic knowledge graph.  
PeerJ Computer Science. 2017. 3:e106 <https://doi.org/10.7717/peerj-cs.106>

# Search registries for relevant datasets

Open Data Cloud Browse Submit a dataset Diagram Subclouds About

## Datasets

Search

	Identifier	View	Edit
Open Data Web	data.odw.tw		
2000 U.S. Census in RDF (rdfabout.com)	2000-us-census-rdf		
2001 Spanish Census to RDF	2001-spanish-census-to-rdf		
BibSonomy - The blue social bookmark and publication sharing system.	BibSonomy		
Linked open EP data	European Patent Information		
Face Link	Face Link		
HealthCare Ontology	HealthCare Ontology		
Inztrow	Inztrow		
Location Ontology	Location Ontology		
ontology terrorist attack	ROSHANI		

First Previous 1 2 3 4 5 6 7 8 9 10 Next Last

# Success depends on quality of metadata

## Wikidata (Edit)

### About this dataset

free knowledge database project hosted by Wikimedia and edited by volunteers

License: <https://creativecommons.org/publicdomain/zero/1.0/>

cross\_domain wikimedia wikipedia

### Contact Details

Contact Point: [Lucas Werkmeister](#)

Website: <https://www.wikidata.org/>

### Download Links

#### Full Downloads

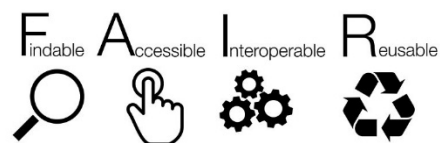
- [Latest full dump, bzip2 compressed](#) (The latest full dump of all Wikidata entity data, in Turtle format, compressed with bzip2. Dumps are generated on a weekly basis.)
- [Latest full dump, gzip compressed / Mirror 1](#) (The latest full dump of all Wikidata entity data, in Turtle format, compressed with gzip. Dumps are generated on a weekly basis.)
- [Latest truthy dump, bzip2 compressed](#) (The latest "truthy" dump of all Wikidata entity data, in N-Triples format, compressed with bzip2. Contains only "truthy" or "best" statements, without qualifiers or references. Dumps are generated on a weekly basis.)
- [Latest truthy dump, gzip compressed](#) (The latest "truthy" dump of all Wikidata entity data, in N-Triples format, compressed with gzip. Contains only "truthy" or "best" statements, without qualifiers or references. Dumps are generated on a weekly basis.)

#### SPARQL Endpoints

- [Wikidata Query Service](#) (The Wikidata Query Service generally contains the full data of Wikidata, modulo a slight delay in updating (usually less than a few seconds), a handful of spurious synchronization errors (requests are routed to a randomly selected server from the pool, and servers may be slightly out of sync), and a few differences between the RDF dump format and the WDQS version (see [here](#)).

Load metadata as:

N  
/XML  
le  
iples



- Metadata identifier
- Resource identifier
- Standardized, machine readable format
- Use of community vocabularies

```
ix void: <http://rdfs.org/ns/void#> .
ix xsd: <http://www.w3.org/2001/XMLSchema#> .
ix dcterms: <http://purl.org/dc/terms/> .
ix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
ix dcat: <http://www.w3.org/ns/dcat#> .
ix foaf: <http://xmlns.com/foaf/0.1/> .
```

- License?
- Provenance?

```
p://lod-cloud.net/dataset/wikidata>
  a void:Dataset ;
  dcterms:description "free knowledge database project hosted by Wikimedia and edited by volunteers"@en ;
  dcterms:publisher [ rdfs:label "Lucas Werkmeister" ;
                    foaf:mbox "wikidata@wikimedia.de"
                  ] ;
  dcterms:subject "cross_domain" , "wikimedia" , "wikipedia" ;
  dcterms:title "Wikidata"@en ;
  void:dataDump <https://dumps.wikimedia.org/wikidatawiki/entities/latest-all.ttl.gz> ,
ps://dumps.wikimedia.org/wikidatawiki/entities/latest-all.ttl.bz2> , <https://dumps.wikimedia.org/wikidatawiki/entities/latest-
hy.nt.gz> , <https://dumps.wikimedia.org/wikidatawiki/entities/latest-truthy.nt.bz2> ;
  void:exampleResource <https://www.wikidata.org/wiki/Q8023> , <http://www.wikidata.org/entity/Q8023> ,
ps://www.wikidata.org/wiki/Special:EntityData/Q8023> , <https://www.wikidata.org/wiki/Special:EntityData/Q8023.nt> ,
ps://www.wikidata.org/wiki/Special:EntityData/Q8023.ttl> , <https://www.wikidata.org/wiki/Special:EntityData/Q8023.rdf> ,
ps://www.wikidata.org/wiki/Special:EntityData/Q8023.json> ;
  void:sparqlEndpoint <https://query.wikidata.org/sparql> ;
  void:triples 5800000000 ;
  dcat:distribution [ dcat:accessURL <https://query.wikidata.org/> ] ;
  foaf:homepage <https://www.wikidata.org/> .
```





<http://www.w3.org/TR/hcls-dataset/>

## Dataset Descriptions: HCLS Community Profile

### Editors working draft.

#### Editors:

Alasdair J.G. Gray, Heriott-Watt University, UK <[A.J.G.Gray@hw.ac.uk](mailto:A.J.G.Gray@hw.ac.uk)>  
Joachim Baran, Stanford University, USA <[joachim.baran@stanford.edu](mailto:joachim.baran@stanford.edu)>  
M. Scott Marshall, MAASTRO Clinic, The Netherlands <[m.scott\\_marshall@maastro.nl](mailto:m.scott_marshall@maastro.nl)>  
Michel Dumontier, Stanford University, USA <[michel.dumontier@stanford.edu](mailto:michel.dumontier@stanford.edu)>

#### Contributors:

Vladimir Alexiev, Ontotext Corp, Bulgaria <[vladimir.alexiev@ontotext.com](mailto:vladimir.alexiev@ontotext.com)>  
Peter Ansell, CSIRO, Australia <[peter.ansell@csiro.au](mailto:peter.ansell@csiro.au)>  
Gary D. Bader, The Donnelly Centre, University of Toronto, Canada <[gary.bader@utoronto.ca](mailto:gary.bader@utoronto.ca)>  
Asuka Bando, NBDC, Japan <[bando@biosciencedbc.jp](mailto:bando@biosciencedbc.jp)>  
Jerven Bolleman, SIB Swiss Institute of Bioinformatics, Switzerland <[jerven.bolleman@isb-sib.ch](mailto:jerven.bolleman@isb-sib.ch)>  
Alison Callahan, Carleton University, Canada <[alison.callahan@carleton.ca](mailto:alison.callahan@carleton.ca)>  
José Cruz-Toledo, Carleton University, Canada <[josecruztoledo@email.carleton.ca](mailto:josecruztoledo@email.carleton.ca)>  
Pascale Gaudet, SIB Swiss Institute of Bioinformatics, Switzerland <[pascale.gaudet@isb-sib.ch](mailto:pascale.gaudet@isb-sib.ch)>  
Erich Gombocz, IO Informatics, USA <[egombocz@io-informatics.com](mailto:egombocz@io-informatics.com)>  
Alejandra Gonzalez-Beltran, University of Oxford, UK <[alejandra.gonzalez.beltran@gmail.com](mailto:alejandra.gonzalez.beltran@gmail.com)>  
Paul Groth, VU University Amsterdam, The Netherlands <[p.t.groth@vu.nl](mailto:p.t.groth@vu.nl)>  
Melissa Haendel, Oregon Health and Science University, USA <[haendel@ohsu.edu](mailto:haendel@ohsu.edu)>  
Maori Ito, NIBIO, Japan <[maori@nibio.go.jp](mailto:maori@nibio.go.jp)>  
Simon Jupp, EMBL-EBI, UK <[jupp@ebi.ac.uk](mailto:jupp@ebi.ac.uk)>  
Nick Juty, EMBL-EBI, UK <[juty@ebi.ac.uk](mailto:juty@ebi.ac.uk)>  
Toshiaki Katayama, Database Center for Life Sciences, Japan <[ktym@dbcls.jp](mailto:ktym@dbcls.jp)>  
Norio Kobayashi, RIKEN, Japan <[norio.kobayashi@riken.jp](mailto:norio.kobayashi@riken.jp)>  
Kalpana Krishnaswami, Metaome, USA <[kalpana@metaome.com](mailto:kalpana@metaome.com)>  
Camille Laibe, EMBL-EBI, UK <[laibe@ebi.ac.uk](mailto:laibe@ebi.ac.uk)>  
Nicolas Le Novère, Babraham Institute, UK <[n.lenovere@gmail.com](mailto:n.lenovere@gmail.com)>  
Simon Lin, Marshfield Clinic Research Foundation, USA <[lin.simon@mcrf.mfldclin.edu](mailto:lin.simon@mcrf.mfldclin.edu)>  
James Malone, EMBL-EBI, UK <[malone@ebi.ac.uk](mailto:malone@ebi.ac.uk)>  
Michael Miller, Institute for Systems Biology, USA <[mmiller@systemsbiology.org](mailto:mmiller@systemsbiology.org)>  
Chris Mungall, Lawrence Berkeley National Laboratory, USA <[cjm@berkeleybop.org](mailto:cjm@berkeleybop.org)>  
Laurens Rietveld, VU University Amsterdam, The Netherlands <[laurens.rietveld@vu.nl](mailto:laurens.rietveld@vu.nl)>  
Sarala M. Wimalaratne, EMBL-EBI, UK <[sarala@ebi.ac.uk](mailto:sarala@ebi.ac.uk)>  
Atsuko Yamaguchi, Database Center for Life Sciences, Japan <[atsuko@dbcls.jp](mailto:atsuko@dbcls.jp)>

standard is  
registered in  
FAIRsharing



The screenshot shows the FAIRsharing.org interface. At the top, the logo for FAIRsharing.org is displayed with the tagline 'standards, databases, policies'. Below the logo, a breadcrumb trail reads 'standards > model/format > doi:10.25504/fairsharing.s248mf'. A blue button labeled 'Ask Question' is visible. The main heading is 'W3C HCLS Dataset Description'. Underneath, there is a 'General Information' section with a paragraph of text: 'Access to consistent, high-quality metadata is critical to finding, understanding, and reusing scientific data. This document describes a consensus among participating stakeholders in the Health Care and the Life Sciences domain on the description of datasets using the Resource Description Framework (RDF). This specification meets key functional requirements, reuses existing vocabularies to the extent that it is possible, and addresses elements of data description, versioning, provenance, discovery, exchange, query, and retrieval.' Below this, there is a 'How to cite this record' section with the following text: 'FAIRsharing.org: W3C HCLS Dataset Description; W3C HCLS Dataset Description; DOI: https://doi.org/10.25504/FAIRsharing.s248mf; Last edited: Feb. 22, 2018, 2:09 p.m.; Last accessed: Mar 18 2018 10:11 p.m.' There is also a 'Homepage' link to 'http://www.w3.org/TR/hcls-dataset/' and a list of countries where it was developed: 'Australia, Bulgaria, Canada, Japan, Netherlands, Switzerland, United Kingdom, United States'. It notes it was 'Created in 2011' and has a 'Taxonomic range' of 'All'. At the bottom, there are tags for 'Annotation', 'Biomedical Science', 'Life Science', and 'Medicine'. The footer features logos for the 'RESEARCH DATA ALLIANCE' (RDA) and 'FORCE11 The Future of Research Communications and e-Scholarship'.

# Conformance to the (meta)data standard should be machine actionable

<https://hw-swel.github.io/Validata/>

**Validata: RDF Validator using Shape Expressions**  
Intuitive, standalone web-based tool to help building valid RDF documents by validating against preset schemas written in the [Shape Expressions \(ShEx\)](#) language.  
Licensed under the permissive MIT license by various authors.

Schema  
Data  
Configure Options  
Validation Results

Summary  
 Summary

IBL 2015 Demo

### Select Schema

Select which schema your data should be validated against:

HCLS Community Profile [Show Source](#)

<b>Description</b>	ShEx schema capturing the HCLS Community Profile published on 14 May 2015
<b>Creation Date</b>	27/07/2015 13:40

### Input Data (Turtle, TriG, N-Triples or N-Quads)

**Upload Data File**  
[Choose File](#) No file chosen  
Select a data file from your computer.

**Directly Input Data**

1	<a href="#">+</a> <a href="#">-</a>
---	-------------------------------------

<https://github.com/micheldumontier/hcls-shex>

```
<HCLSSummaryShape> IRI
  // dct:creator "Michel Dumontier"
  // dct:created "2019-09-05 03:26:38"
  // prov:used <https://github.com/micheldumontier/hcls-shex>
  EXTRA rdf:type
{
  rdf:type [dctypes:Dataset]
  // rdfs:comment "summary level MUST use rdf:type with dctypes:Dataset"
  // :metadata-element "Type declaration"
  // :requirement-level :MUST;
```



IOHACKATH  
in FUKU

constraint validation tool that is configurable  
any profile

declarative reusable schema description




Shape Expression (ShEx) based, but with extension

Now compliant with ShEx  
Convertible to SHACL

Working on conversion to JSON-Schema

# A design framework and exemplar metrics for FAIRness

<http://fairmetrics.org>

Mark D. Wilkinson , Susanna-Assunta Sansone , Erik Schultes, Peter Doorn, Luiz Olavo Bonino da Silva Santos & Michel Dumontier 

*Scientific Data* **5**, Article number: 180118 (2018) | [Download Citation](#) ↓

- **14 universal metrics** covering each of the FAIR sub-principles. The **metrics** don't dictate any particular standards. They **simply demand evidence (using protocols of the Web)** that the resource has met community expectations.
- Digital resource providers must **provide** at least one web-accessible document with **machine-readable metadata** (FM-F2, FM-F3), **resource management plan** (FM-A2), and any **additional authorization** procedures (FM-A1.2).
- They must use **publically registered: identifier schemes** (FM-F1A), (secure) **access protocols** (FM-A1.1), **knowledge representation languages** (FM-I1), **licenses** (FM-R1.1), **provenance** specifications (FM-R1.2), and **community standards** (FM-R1.3)
- They must **evidence that their resource** can be located in **search results** (FM-F4), that it provides **links** to other (FAIR) resources (FM-I3; FM-I2), and **it validates against community standards** (FM-R1.3)

**Table 2. Summary of FAIR metrics self-scoring.**

Green = passes FAIR Metric

Red = fails FAIR Metric

Yellow = problematic (for example,

Gray = Can not be evaluated







IRI = Respondent gives an IRI

none = Respondent answered "none"

NRP = No Response Provided

FM	Question	Dataverse	Dryad	Nano-pub	Zenodo	Yale ISPS	Figshare	Broad's SCP	<del>SeaData</del> Net's CDI	Wikidata
IRI Exists	1	IRI	IRI	IRI	IRI	IRI	IRI	IRI	IRI	IRI
F1A	2	IRI	IRI	IRI	IRI	IRI	IRI	IRI	IRI	IRI
F1B	3	IRI	IRI	IRI	NRP	none	IRI	IRI	IRI	IRI
F2A	4A	IRI	IRI	IRI	IRI	none	none	IRI	IRI	IRI
F2A	4B	IRI	none	IRI	IRI	"Multiple"	none	IRI	IRI	IRI
F3	5A	IRI	IRI	IRI	IRI	none	NRP	IRI	IRI	IRI
F3	5B	IRI	IRI	IRI	IRI	IRI	IRI	IRI	none	IRI
F4	6A	IRI	IRI	IRI	IRI	IRI	IRI	IRI	IRI	IRI
F4	6B	IRI	IRI	IRI	IRI	IRI	IRI	IRI	IRI	IRI
A1.1	7A	IRI	IRI	IRI	IRI	IRI	IRI	IRI	IRI	IRI
A1.1	7B	true	true	true	true	true	true	true	true	true
A1.1	7C	true	true	true	true	true	true	true	true	true
A1.2	8A	false	false	false	false	false	false	false	true	false
A1.2	8B	N/A	N/A	N/A	N/A	NRP	NRP	NRP	link	N/A
A2	9	IRI	IRI	none	IRI	none	IRI	none	IRI	NRP
I1	10	IRI	IRI	IRI	IRI	none	none	NRP	IRI	IRI
I2	11	IRI	IRI	IRI	none	none	none	IRI	IRI	IRI
I3	12	NRP	IRI	IRI	none	none	none	NRP	NRP	IRI
R1.1	13	IRI	IRI	IRI	IRI	IRI	IRI	NRP	IRI	IRI
R1.2	14A	IRI	IRI	IRI	IRI	none	none		NRP	NRP
R1.2	14B		none		none	none	none			
R1.3	15	NRP			none	none	none	NRP		

# Automating FAIR Maturity Through a Scalable, Automated, Community-Governed Framework


Mark D Wilkinson,  Michel Dumontier, Anna-Assunta Sansone, Luiz Olavo Bonino da Silva Santos,  Mario Prieto, Dominique Batista,  Peter McQuilton,  Tobias Kuhn, Philippe Rocca-Serra,  Mercè Crosas,  Erik Schultes


<https://doi.org/10.1101/649202>

May 28, 2019.


to appear in *Nature Scientific Data*











































<http://w3id.org/AmIFAIR>

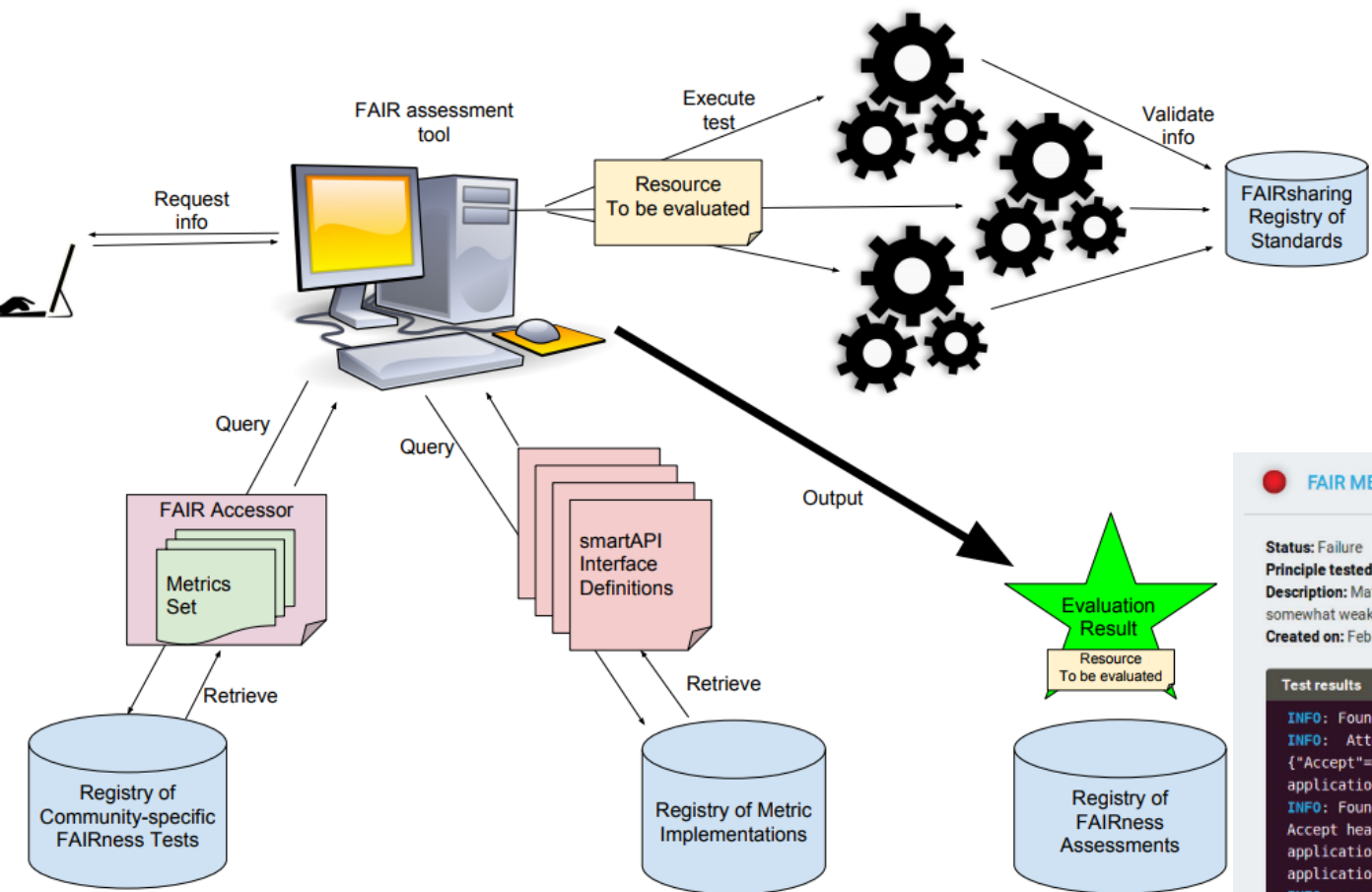
FAIR Assessment of the FAIR Evaluation Service 

Summary:   
Description: FAIR Metrics Evaluation: FAIR Assessment of the FAIR Evaluation Service. Tested Identifier: <http://w3id.org/AmIFAIR>, generated by <https://orcid.org/0000-0003-4727-9438>  
Resource: <http://w3id.org/AmIFAIR>  
Collection: 6  
Observations: Ran 22 tests (14 succeeded, 8 failed).

Tests passing and failing



 FAIR METRICS GEN2 - UNIQUE IDENTIFIER 
 FAIR METRICS GEN2 - IDENTIFIER PERSISTENCE 
 FAIR METRICS GEN2 - DATA IDENTIFIER PERSISTENCE 
 FAIR METRICS GEN2 - STRUCTURED METADATA 
 FAIR METRICS GEN2 - GROUNDED METADATA 
 FAIR METRICS GEN2 - DATA IDENTIFIER EXPLICITLY IN METADATA 
 FAIR METRICS GEN2 - METADATA IDENTIFIER EXPLICITLY IN METADATA 
 FAIR METRICS GEN2 - SEARCHABLE IN MAJOR SEARCH ENGINE 
 FAIR METRICS GEN2 - USES OPEN FREE PROTOCOL FOR DATA RETRIEVAL 
 FAIR METRICS GEN2 - USES OPEN FREE PROTOCOL FOR METADATA RETRIEVAL 
 FAIR METRICS GEN2 - DATA AUTHENTICATION AND AUTHORIZATION 
 FAIR METRICS GEN2 - METADATA AUTHENTICATION AND AUTHORIZATION 
 FAIR METRICS GEN2 - METADATA PERSISTENCE 
 FAIR METRICS GEN2 - METADATA KNOWLEDGE REPRESENTATION LANGUAGE (WEAK) 
 FAIR METRICS GEN2 - METADATA KNOWLEDGE REPRESENTATION LANGUAGE (STRONG) 
 FAIR METRICS GEN2 - DATA KNOWLEDGE REPRESENTATION LANGUAGE (WEAK) 
 FAIR METRICS GEN2 - DATA KNOWLEDGE REPRESENTATION LANGUAGE (STRONG) 
 FAIR METRICS GEN2 - METADATA USES FAIR VOCABULARIES (WEAK) 
 FAIR METRICS GEN2 - METADATA USES FAIR VOCABULARIES (STRONG) 
 FAIR METRICS GEN2 - METADATA CONTAINS QUALIFIED OUTWARD REFERENCES 
 FAIR METRICS GEN2 - METADATA INCLUDES LICENSE (STRONG) 
 FAIR METRICS GEN2 - METADATA INCLUDES LICENSE (WEAK) 



**FAIR METRICS GEN2 - METADATA USES FAIR VOCABULARIES (WEAK)**

**Status:** Failure  
**Principle tested:** I2  
**Description:** Maturity Indicator to test if the linked data metadata uses terms that resolve. This tests only if they resolve, not if they resolve to FAIR somewhat weak test.  
**Created on:** Feb 21, 2019 by [Mark D Wilkinson](#) (updated on Feb 21, 2019).

**Test results**

```

INFO: Found a URI.
INFO: Attempting to resolve https://fairsharing.github.io/FAIR-Maturity-FrontEnd/ using HTTP Headers {"Accept"=>"text/turtle, application/ld+json, application/rdf+xml, text/xhtml+xml, application/n3, application/turtle, application/x-turtle, text/n3, text/turtle, text/rdf+n3, text/rdf+turtle, application/n-triples"}.
INFO: Found html text/html type of content when resolving https://fairsharing.github.io/FAIR-Maturity-FrontEnd/ using HTTP Headers {"Accept"=>"text/turtle, application/ld+json, application/rdf+xml, text/xhtml+xml, application/n3, application/turtle, application/x-turtle, text/n3, text/turtle, text/rdf+n3, text/rdf+turtle, application/n-triples"}.
INFO: parsing as HTML.
INFO: Using 'extract' to try to extract metadata from return value (message body) of https://fairsharing.github.io/FAIR-Maturity-FrontEnd/.
INFO: the extract tool found parseable data at https://fairsharing.github.io/FAIR-Maturity-FrontEnd/.
INFO: The response message body component appears to contain JSON:LD:Format.
INFO: Attempting to https://fairsharing.github.io/FAIR-Maturity-FrontEnd/ using HTTP Headers {"Accept"=>"text/turtle, application/ld+json, application/rdf+xml, text/xhtml+xml, application/n3, application/rdf+n3, application/turtle, application/x-turtle, text/n3, text/turtle, text/rdf+n3, text/rdf+turtle, application/n-triples"}.
INFO: Found html text/html type of content when resolving https://fairsharing.github.io/FAIR-Maturity-FrontEnd/ using HTTP Headers {"Accept"=>"**/*"}.
INFO: parsing as HTML.
INFO: Using 'extract' to try to extract metadata from return value (message body) of https://fairsharing.github.io/FAIR-Maturity-FrontEnd/.
INFO: the extract tool found parseable data at https://fairsharing.github.io/FAIR-Maturity-FrontEnd/.
INFO: The response message body component appears to contain JSON:LD:Format.
INFO: Linked data was found.
FAILURE: 0 of the first 4 predicates discovered in the linked data could be resolved. The minimum to pass this test is 50%.
  
```



# FAIR Data Maturity Model Working Group

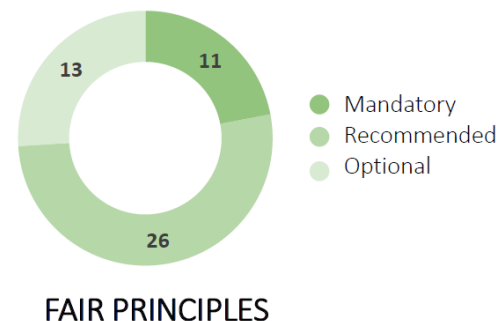
Bring together stakeholders to build on existing approaches and expertise

Establish core assessment criteria for FAIRness

Explore a FAIR data maturity model & toolset

Produce an RDA Recommendation

Develop FAIR data checklist



**Weighting** the indicators, developed as part of the WG, following the [key words for use](#) in RFC2119

- > **Mandatory / Essential**: indicator **MUST** be satisfied for FAIRness
- > **Recommended / Important**: indicator **SHOULD** be satisfied, if at all possible, to increase FAIRness
- > **Optional / Useful**: indicator **MAY** be satisfied, but not necessarily so

PRINCIPLE	INDICATOR_ID	INDICATORS	PRIORITY	
F	F1	F1-01M	Metadata is identified by a persistent identifier	Recommended
	F1	F1-01D	Data is identified by a persistent identifier	Mandatory
	F1	F1-02M	Metadata is identified by a universally unique identifier	Recommended
	F1	F1-02D	Data is identified by a universally unique identifier	Mandatory
	F2	F2-01M	Sufficient metadata is provided to allow discovery, following domain/discipline-specific metadata standard	Recommended
	F2	F2-02M	Metadata is provided for the discovery-related elements defined by the RDA Metadata IG, as much as possible and relevant, if no domain/discipline-specific metadata standard is available	Recommended
	F3	F3-01M	Metadata includes the identifier for the data	Mandatory
	F4	F4-01M	Metadata is offered/published/exposed in such a way that it can be harvested and indexed	Recommended



*Your invitation to participate*

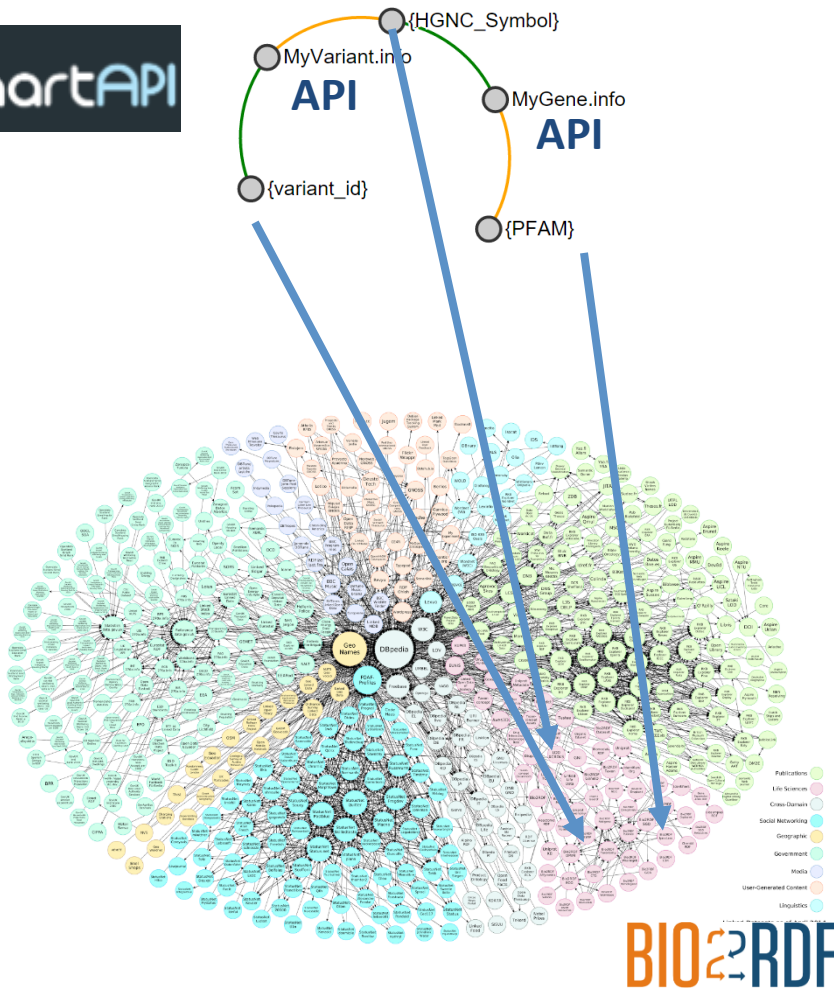


<https://osf.io/n7uwp/>  
[erik.schultes@go-fair.org](mailto:erik.schultes@go-fair.org)



# The Internet of FAIR data and services

Must enable seamless traversal of heterogeneous digital resource



a) Simplified MyGene.info object with JSON-LD context

```
1. {
2.   "@type": "http://identifiers.org/ncbigene/",
3.   "@context": {
4.     "_id": "@id",
5.     "name": "http://schema.org/name",
6.     "interpro": {
7.       "@id": "http://identifiers.org/interpro/",
8.       "@type": "@id"
9.     },
10.    "description": "http://schema.org/description"
11.  },
12.  "_id": "1017",
13.  "symbol": "CDK2",
14.  "name": "cyclin-dependent kinase 2",
15.  "interpro": {
16.    "_id": "IPR000719",
17.    "description": "Protein kinase-like domain"
18.  }
19. }
```



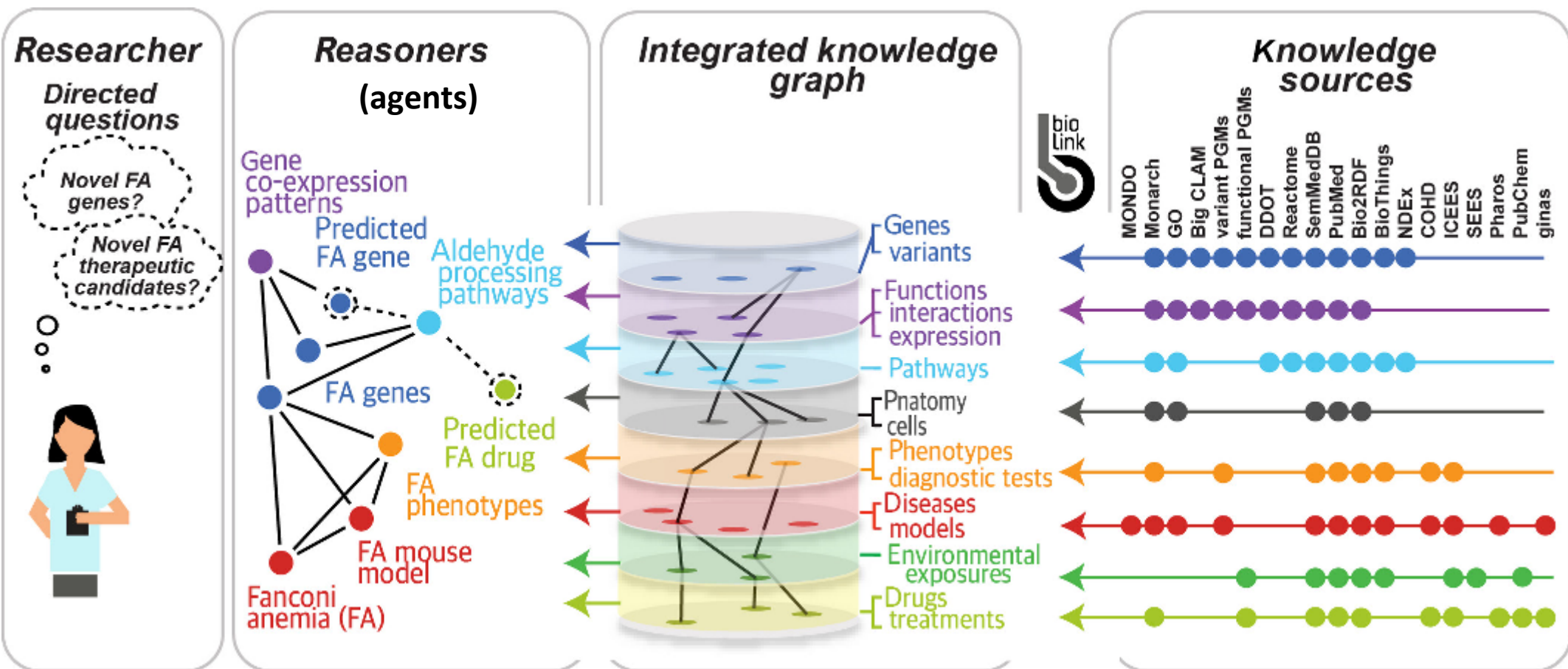
b) Transformed JSON-LD object with semantic URIs included

```
1. {
2.   "@id": "1017",
3.   "@type": "http://identifiers.org/ncbigene/",
4.   "http://schema.org/name": "cyclin-dependent kinase 2",
5.   "http://identifiers.org/interpro/": {
6.     "@id": "IPR000719",
7.     "http://schema.org/description": "Protein kinase-like
8.     domain"
9.   }
10. }
```



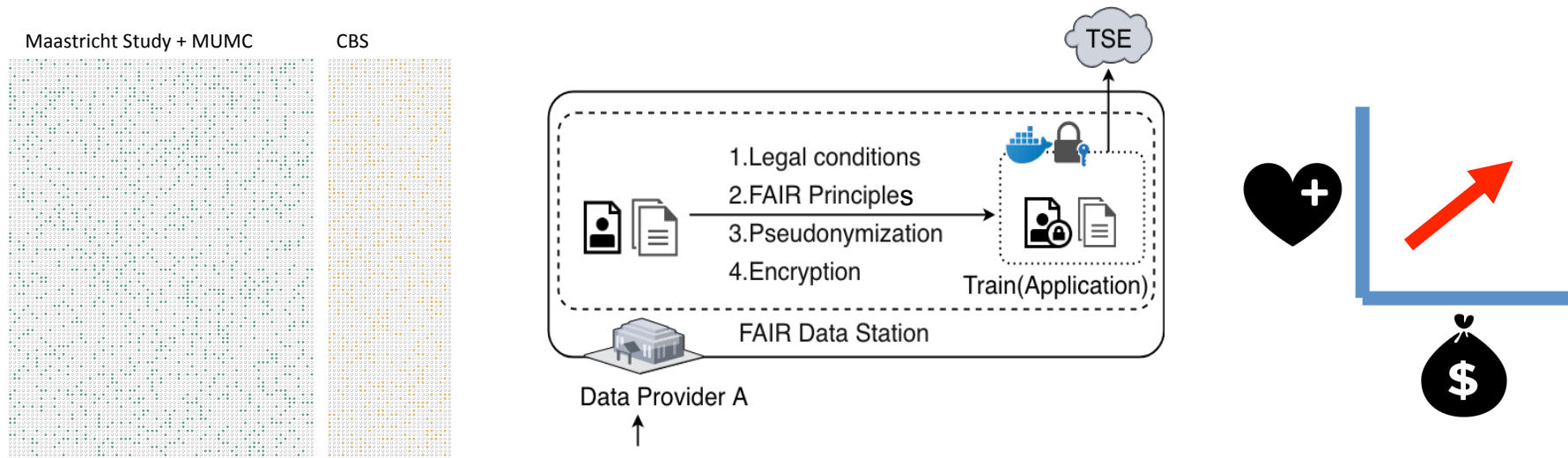
National Center  
for Advancing  
Translational Sciences

# Biomedical Data Translator



A community building a shared infrastructure...

# Mine distributed, access restricted FAIR datasets in a privacy preserving manner



Goal is to learn **high confidence** determinants of health in a **privacy preserving** manner over **vertically partitioned data** from the Maastricht Study and Statistics Netherlands. The data are made available through **FAIR data stations** that provide access to *allowable* subsets of data to *authorized* users of *approved* algorithms.

Establish a **new social, legal, ethical and technological infrastructure** for discovery science in and across health and non-health settings, including scalable **governance** and flexible **consent** to underpin the responsible use of Big Data.

# Summary

**FAIR** represents a global initiative to enhance the discovery and reuse of all kinds of digital resources. *It is a work in progress!*

It demands a **new social, legal, ethical, scientific and technological infrastructure** that currently doesn't exist *in whole*, but has to be built for and adopted by digital savvy communities! It must answer the questions:

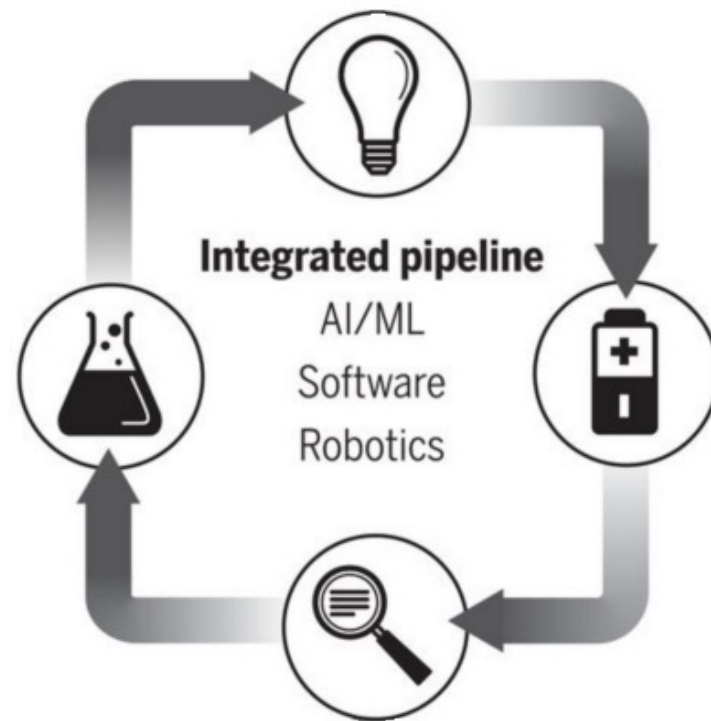
- How can we share data and perform analyses in a responsible manner?
- What incentives, rewards and penalties are needed to maximize trust, participation, legality, and utility?

Semantics, coupled with AI technologies, may enable **humans, aided by intelligent machine agents, to exploit the Internet of FAIR data and services**, and hence to accelerate discovery in biomedicine and in other disciplines.

# FAIR is a part of the solution that will enable arbitrary machines to work with each other



erners-Lee



Ross Kin

Semantic Web



Robot Science

Large Scale, Autonomous Scientific Discovery

# Acknowledgements

## Dumontier Lab (Maastricht University, Stanford University, Carleton University)

U: Seun Adekunle, Remzi Celebi, Dorina Claessens, Ricardo De Miranda Azevedo, Pedro Hernandez Serrano, Massimiliano Grassi, Andine Havelange, Anne Ippel, Alexander Malic, Kody Moodley, Stuti Nayak, Nadine Rouleaux, Claudia van open, Chang Sun, Amrapali Zaveri






J: Sandeep Ayyar, Remzi Celebi, Shima Dastgheib, Maulik Kamdar, David Odgers, Maryam Panahiazar, Amrapali Zaveri

U: Alison Callahan, Jose Toledo-Cruz, Natalia Villaneuva-Rosales

## FAIR

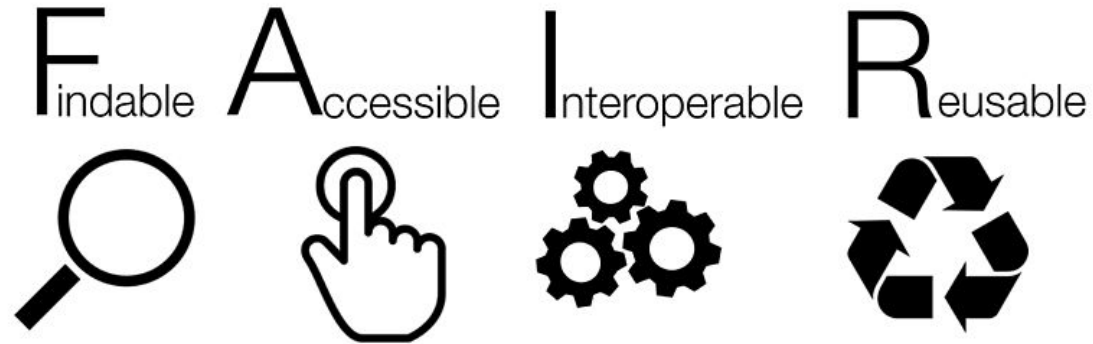
Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao & Barend Mons

## FAIR metrics

 Mark D Wilkinson,  Susanna-Assunta Sansone,  Erik Schultes, Peter Door  
 Luiz Olavo Bonino da Silva Santos,  Michel Dumontier



MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD



The mission of the **Institute of Data Science at Maastricht University** is to foster a collaborative environment for multi-disciplinary data science research, interdisciplinary training, and data-driven innovation.

We tackle key **scientific, technical, social, legal, ethical issues** that advance our understanding across a variety of disciplines and strengthen our communities in the face of these developments.

michel.dumontier@maastrichtuniversity.nl