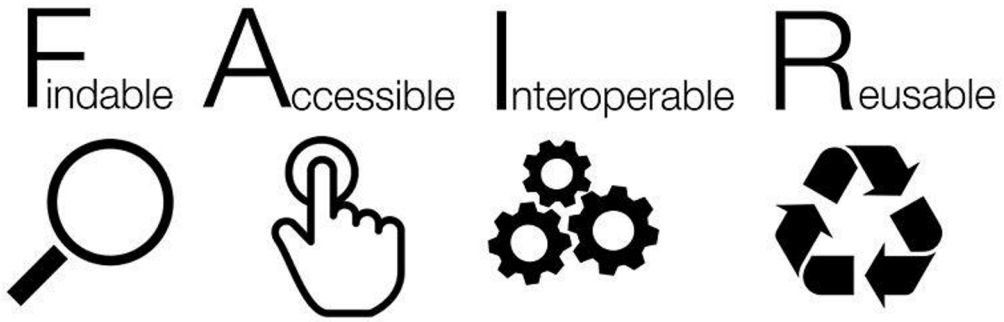
Data-Driven Discovery Science with FAIR Knowledge Graphs



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

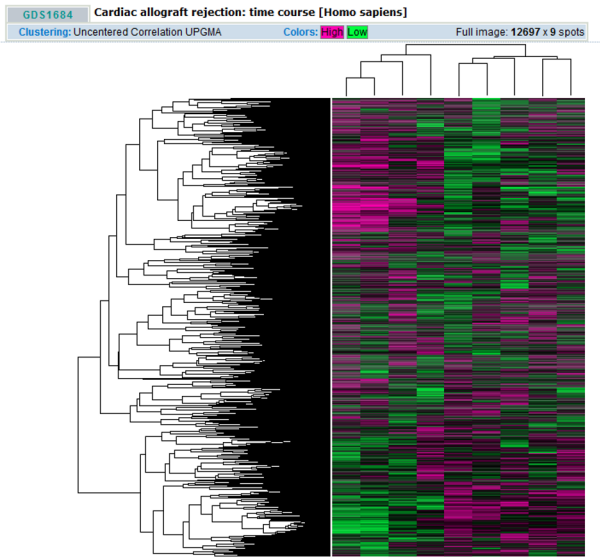
Maastricht University

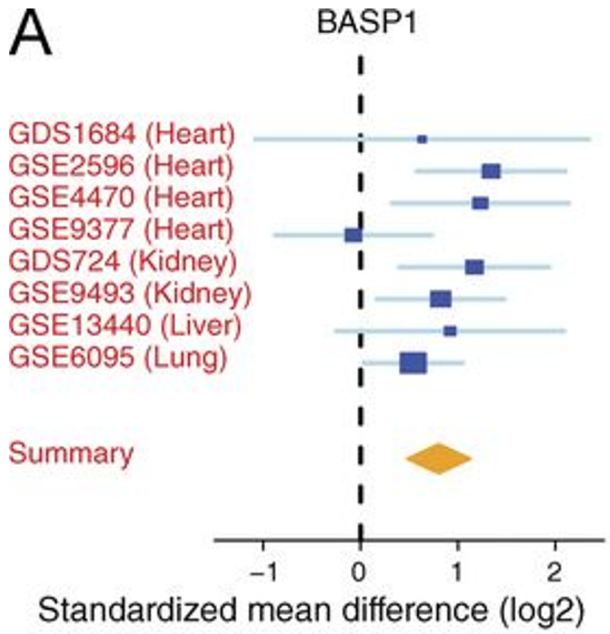
Large Scale Data-Driven Discovery is increasingly possible with growing amounts of open data

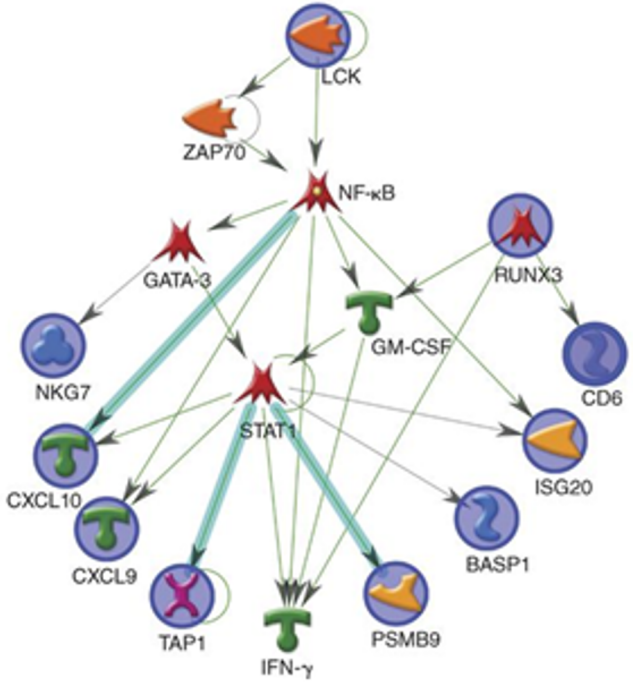

iNSiGHT

A common rejection module (CRM) for acute rejection across multiple organs identifies novel therapeutics for organ transplantation

Khatri et al. JEM. 210 (11): 2205
DOI: 10.1084/jem.20122709







Main Findings:

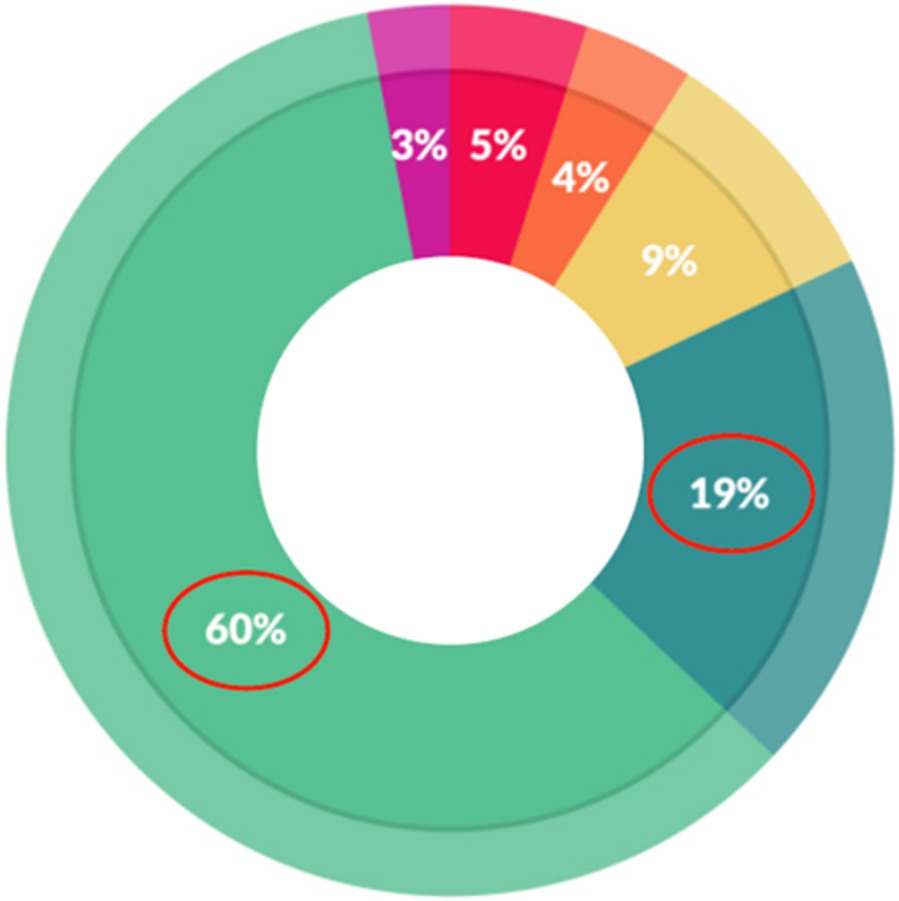
1. CRM of 11 overexpressed genes **predicted future injury** to a graft
2. Mice treated with **existing drugs against specific CRM genes extended graft survival**
3. Retrospective **EHR data analysis supports treatment prediction**

Key Observations:

1. **Meta-analysis** offers a **more reliable estimate** of the direction and magnitude of the effect
2. Existing data can be used to **generate and validate new hypotheses**

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

Data scientists could be more productive



What data scientists spend the most time doing

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

figure eight

CrowdFlower

Low reproducibility of landmark studies

39% (39/100) in psychology[1]
21% (14/67) in pharmacology[2]
11% (6/53) in cancer[3]
unsatisfactory in machine learning[4]

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

Most published research findings are false.

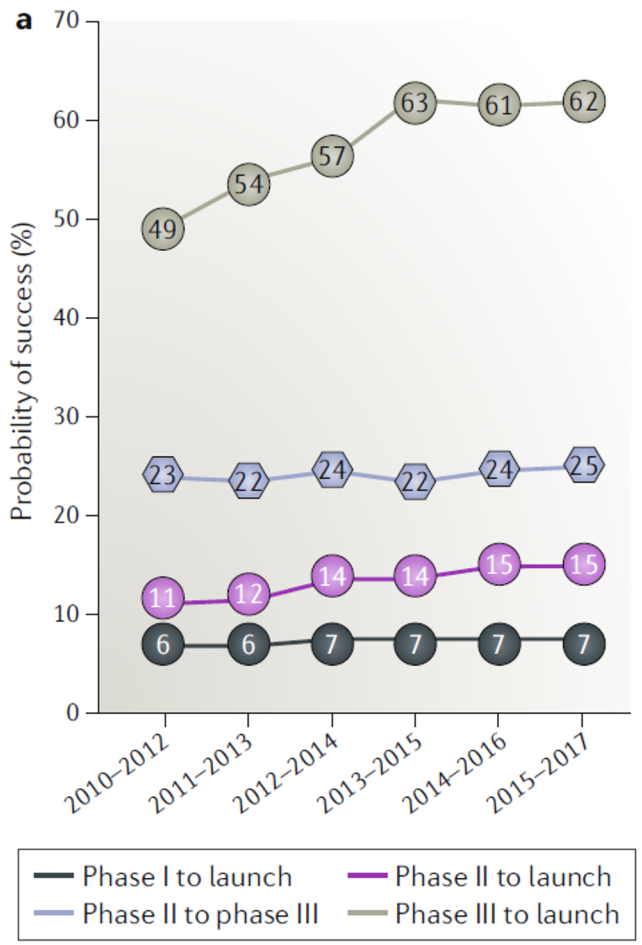
- John Ioannidis, Stanford University

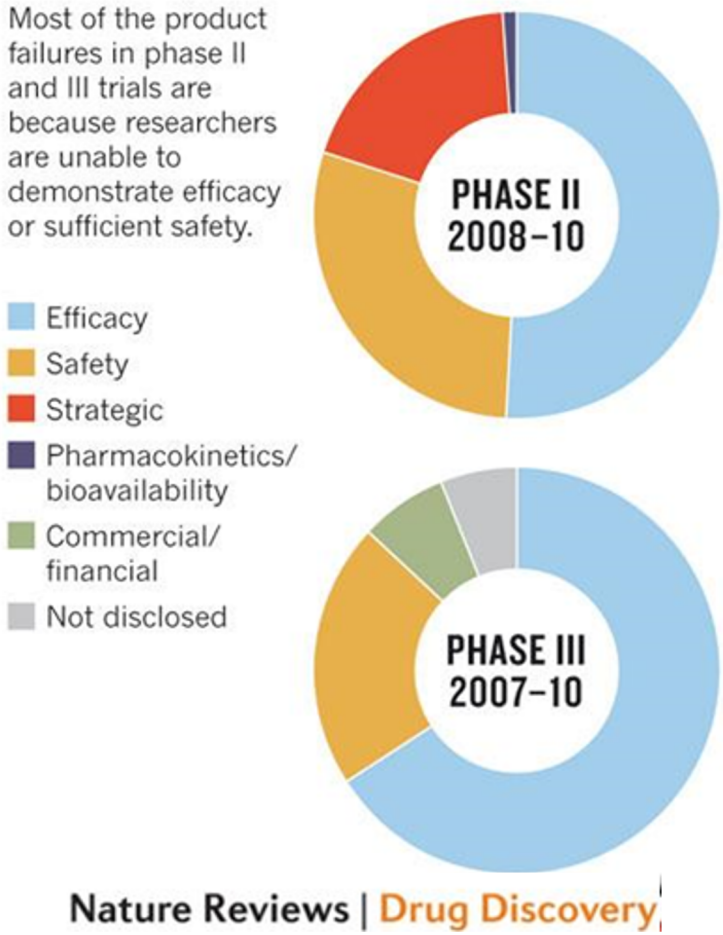
PLoS Med 2005;2(8): e124.

THE CLINICAL-TRIAL CLIFF

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

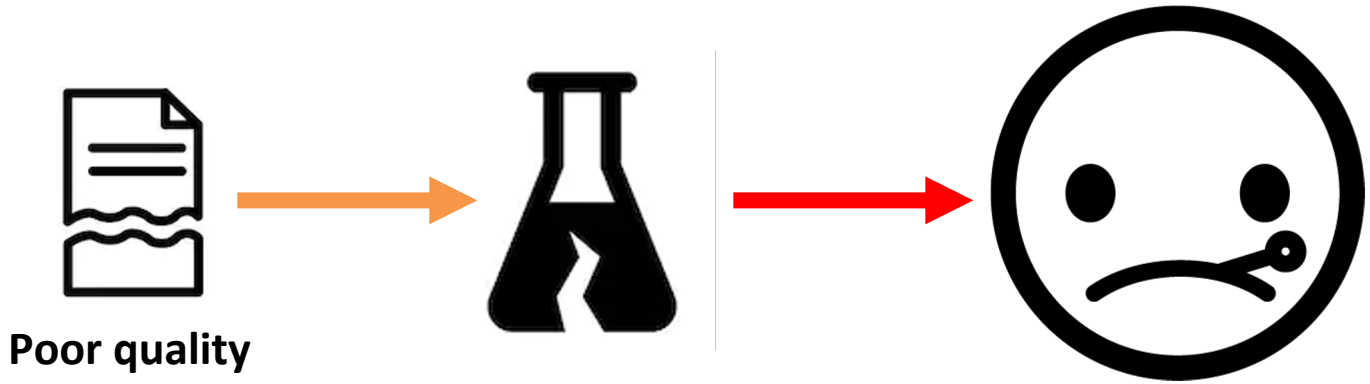
Probability to launch





Nature Reviews Drug Discovery **18**, 495-496 (2019)
https://doi.org/10.1038/d41573-019-00074-z

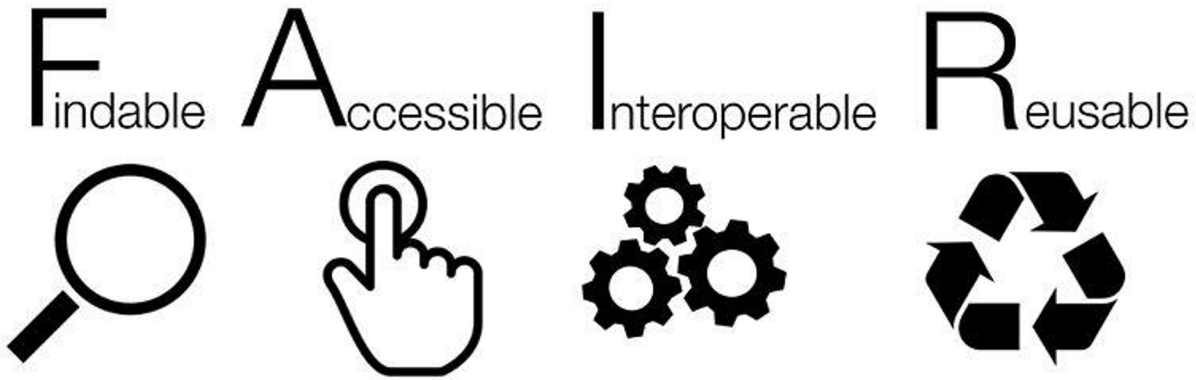
It’s time to completely rethink how we perform and document empirical research

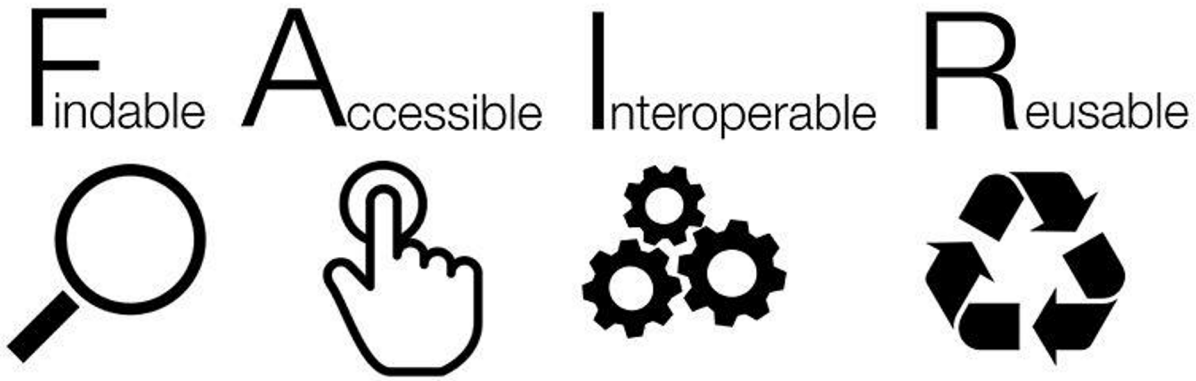

Poor quality
(meta)data
Reproducibility
Crisis
Translational
Failure

Human Machine collaboration is crucial to our future success

Machines, not people,
need to be able to discover and reuse data

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


Findable
Accessible
Interoperable
Reusable



An international, bottom-up paradigm for the discovery and reuse of digital content

for the machines that people use

Open Access | Published: 15 March 2016

The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson, Michel Dumontier, ... Barend Mons + Show authors

Scientific Data **3**, Article number: 160018 (2016) | Cite this article

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

EUROPEAN COMMISSION
Press Release Database
European Commission > Press releases database > Press Release details

European Commission - Statement

G20 Leaders' Communique Hangzhou Summit

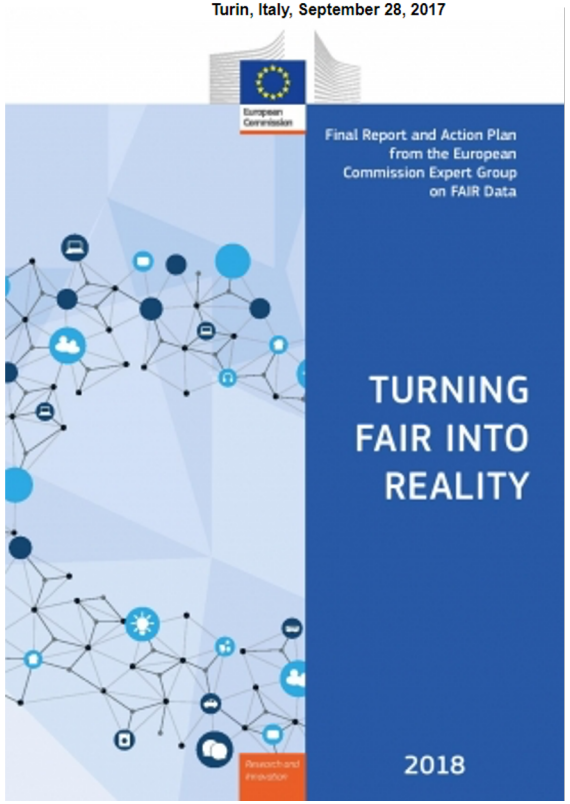
Hangzhou, 5 September 2016

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

G7 2017 ITALIA

Annex 4:
G7 Expert Group on Open Science

Turin, Italy, September 28, 2017



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

FAIR in a nutshell

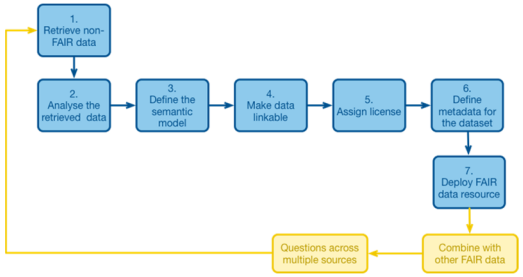
FAIR aims to create **social** and **economic impact** by facilitating the discovery and reuse of **digital resources** through a set of requirements:

- **unique identifiers** to retrieve all forms of digital content and knowledge
- **high quality meta(data)** to enhance discovery of relevant digital resources
- **shared vocabularies** to facilitate query and statistical analysis
- **community standards** to reduce the effort in wrangling different forms of data
- **detailed provenance** to provide adequate context that fosters understanding and reproducibility
- **repositories** to make content available to others in the long term
- **standardized terms of use** to clarify expectations and intensify innovation

A number of guides are now available to make FAIR data

RDMkit

GO FAIR



FAIR cookbook

Article | Open Access | Published: 20 September 2019

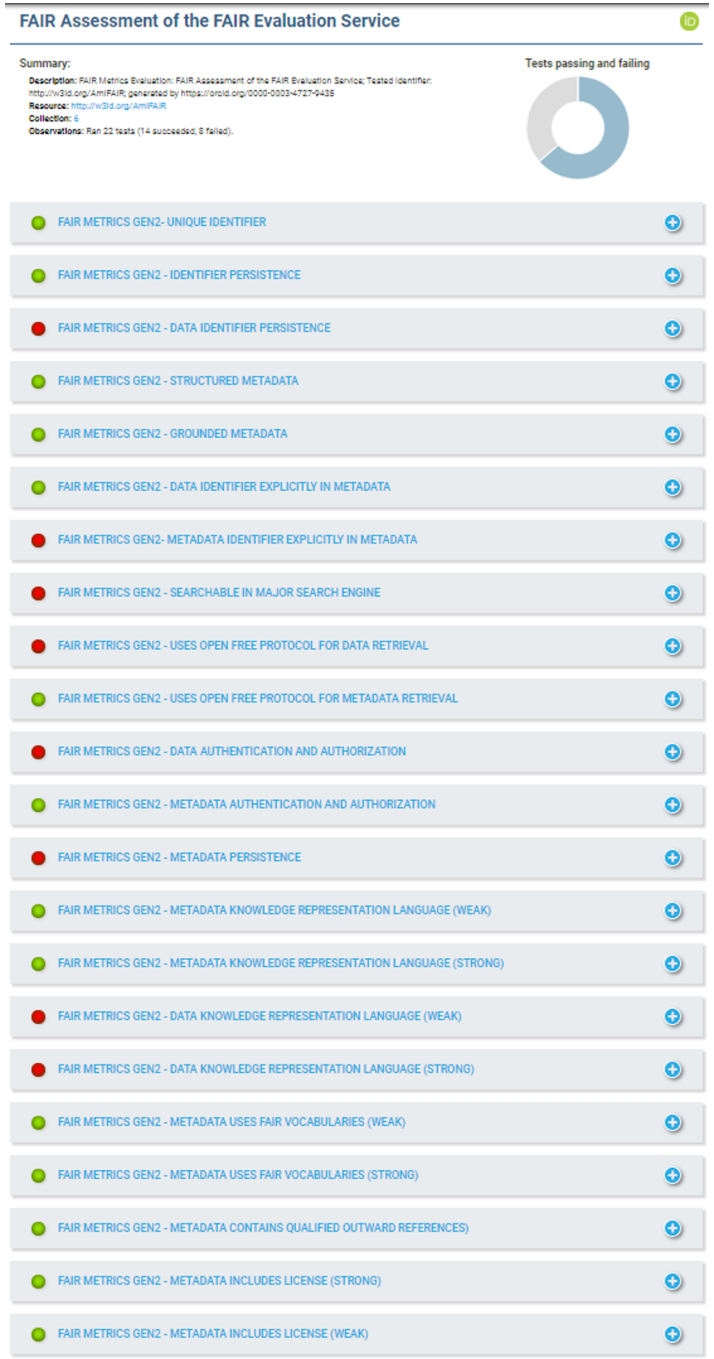
Evaluating FAIR maturity through a scalable, automated, community-governed framework

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

Scientific Data **6**, Article number: 174 (2019) | Cite this article

http://w3id.org/AmIFAIR

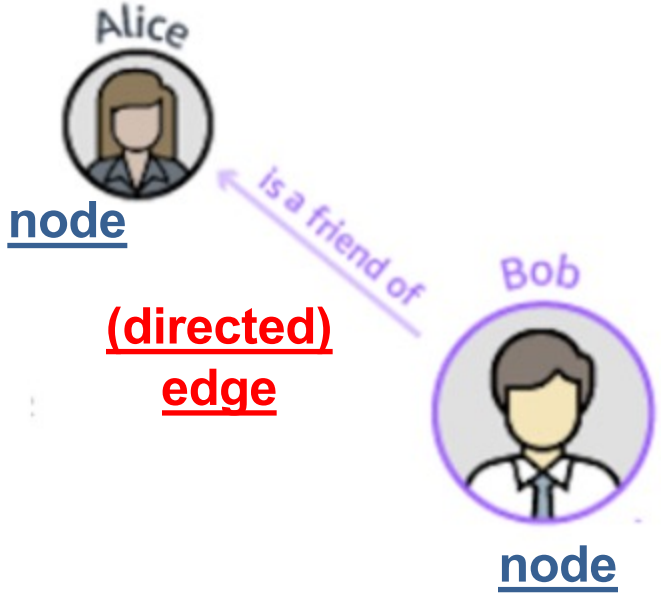
Other schemes: https://fairassist.org



FAIR Knowledge Graphs

What is a Knowledge Graph?

A **knowledge graph** is a **graph** in which the **nodes** (vertices) represent entities and are related to other entities/attributes via **edges** that represent relations.

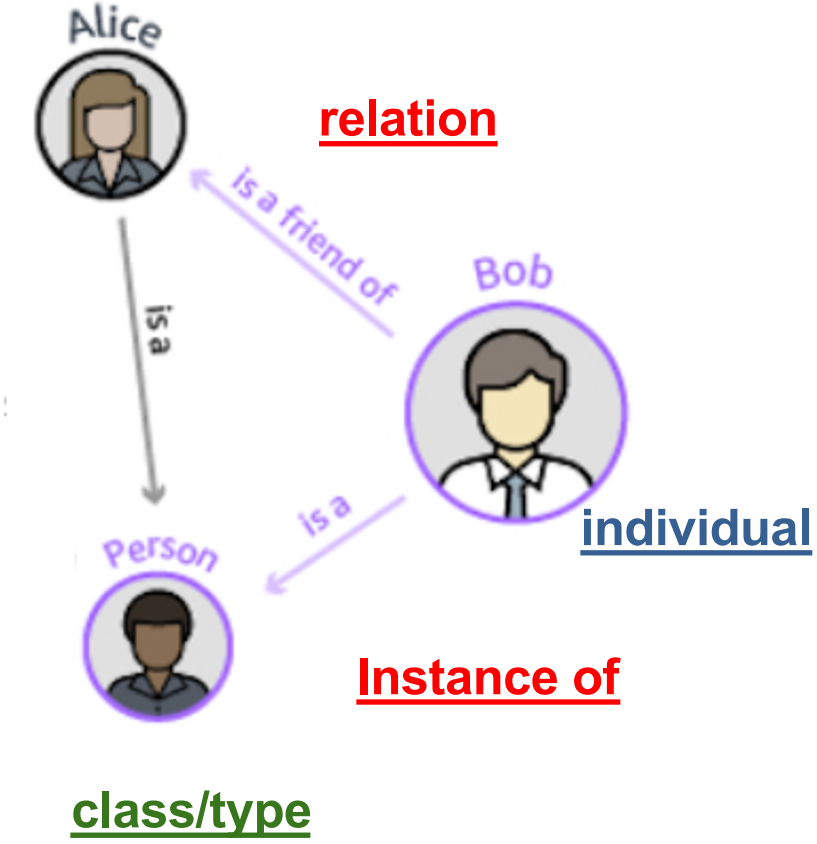


What is a Knowledge Graph?

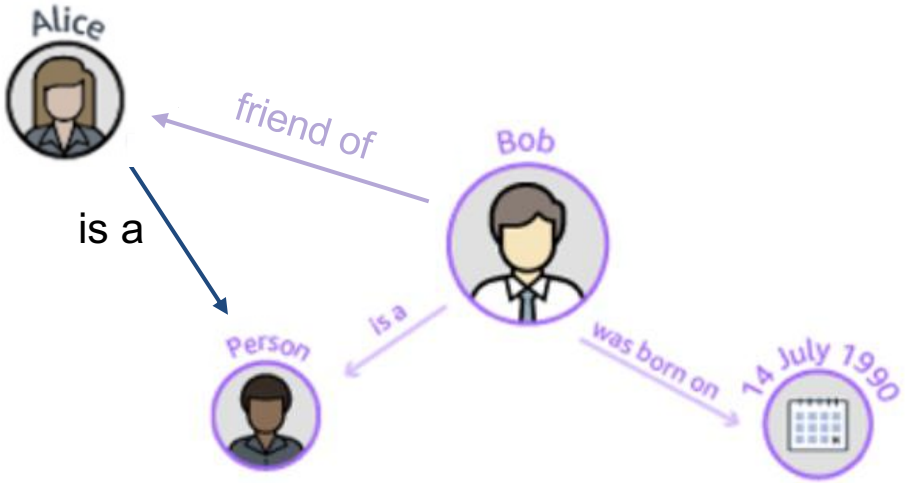
A **knowledge graph** is a **graph** in which the **nodes** (vertices) represent entities and are related to other entities/attributes via **edges** that represent relations.

A knowledge graph mainly consists of **relations between individuals.**

But it can also contain relations between individuals and classes (*type*), where individuals are **instances** of a specified classes, and relations between classes (aka **class expressions**)



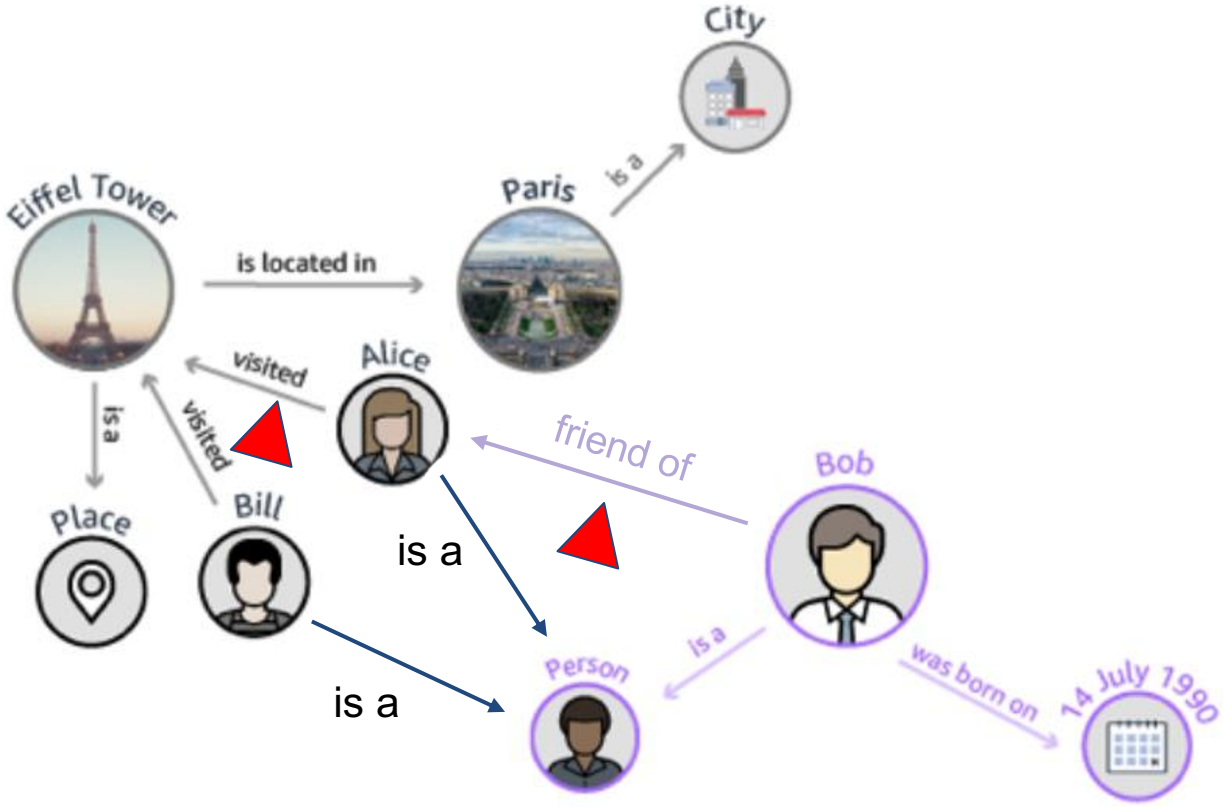
Linking Knowledge Graphs



social network

Linking Knowledge Graphs

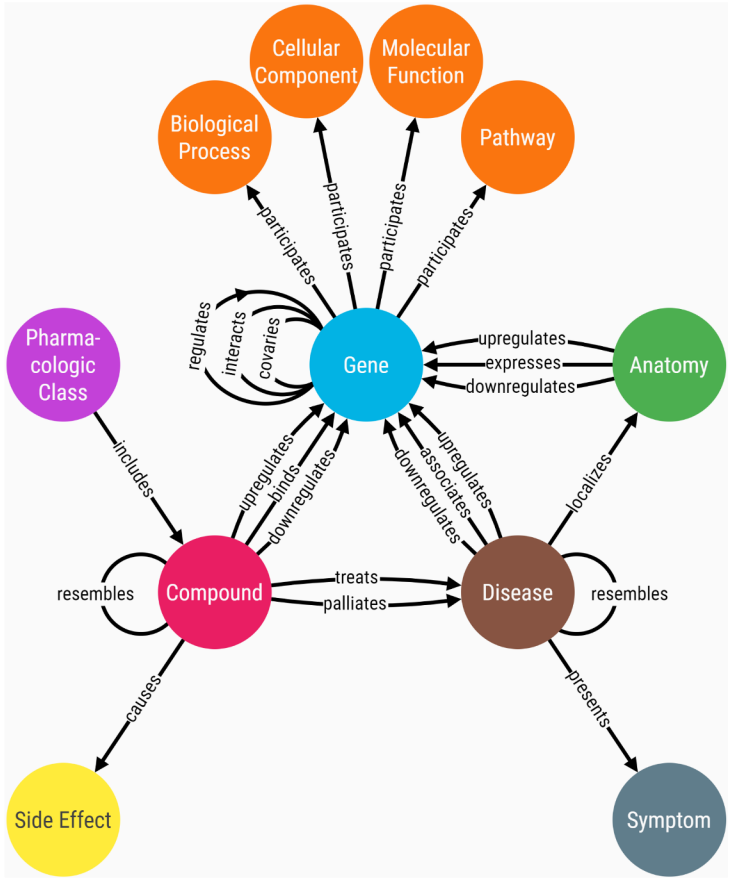
travel network



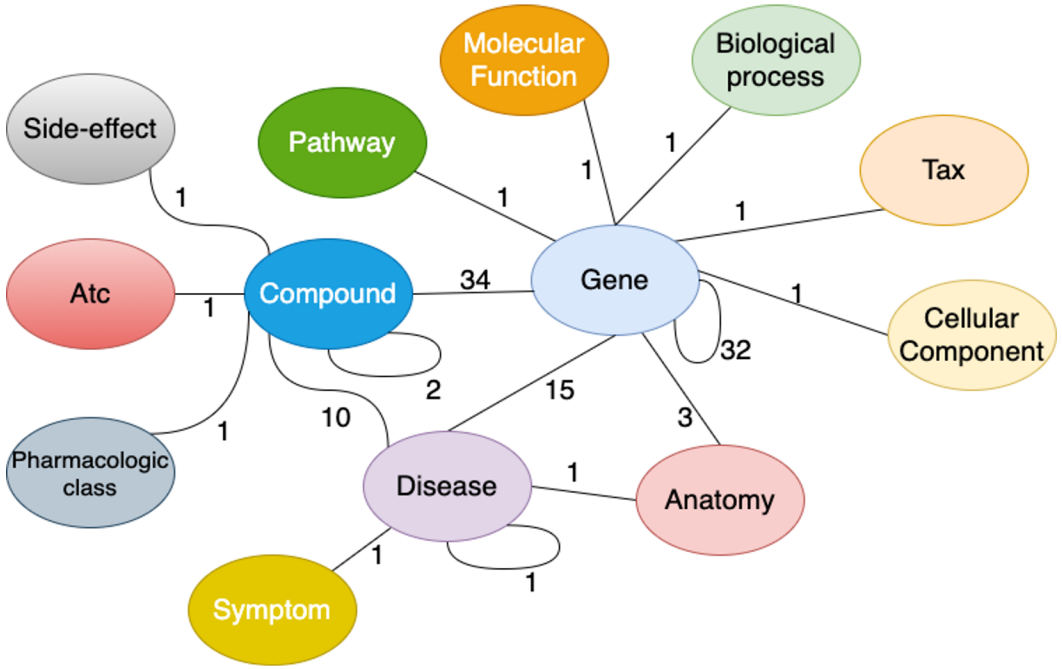
social network

linking relations

A number of (centralized) biomedical KGs do exist, but more work is needed to make them FAIR



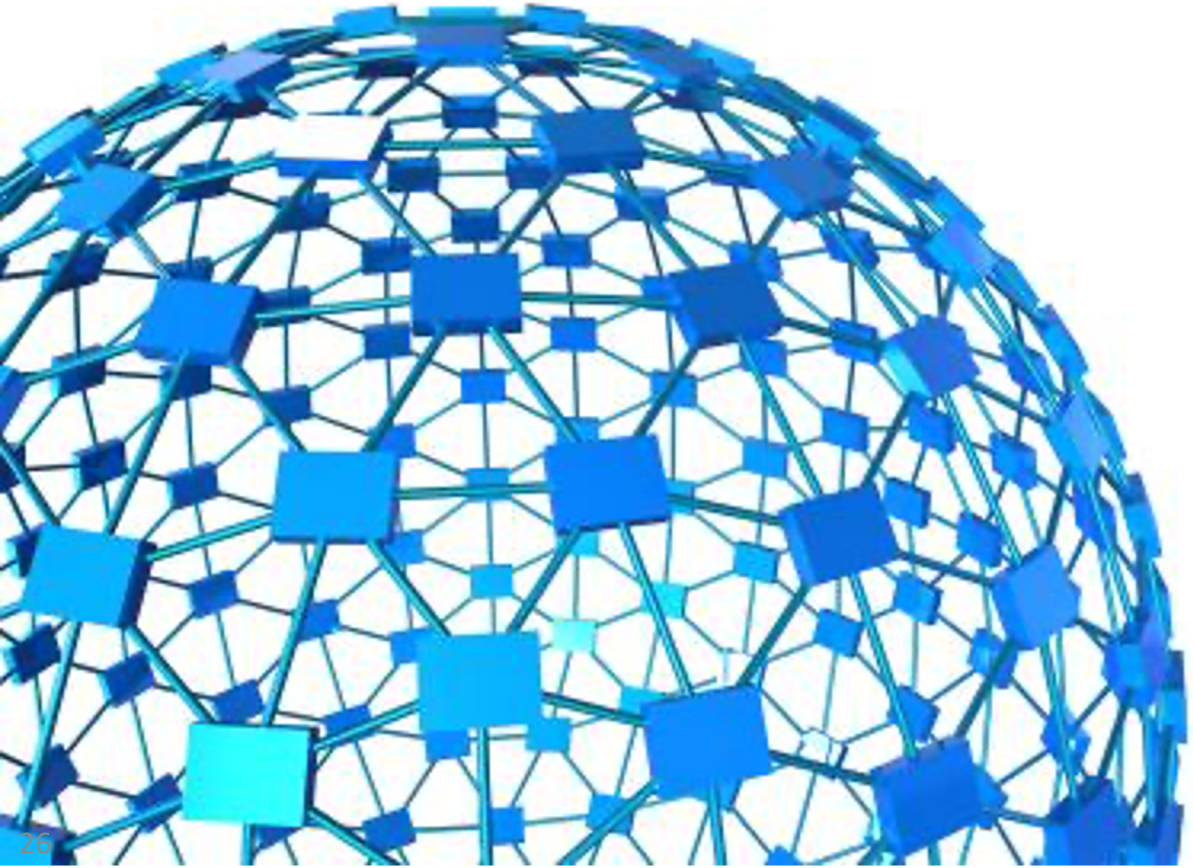
metagraph for [het.io](het.io)



metagraph for DRKG
The number next to an edge indicates the number of relation-types for that entity-pair in DRKG

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

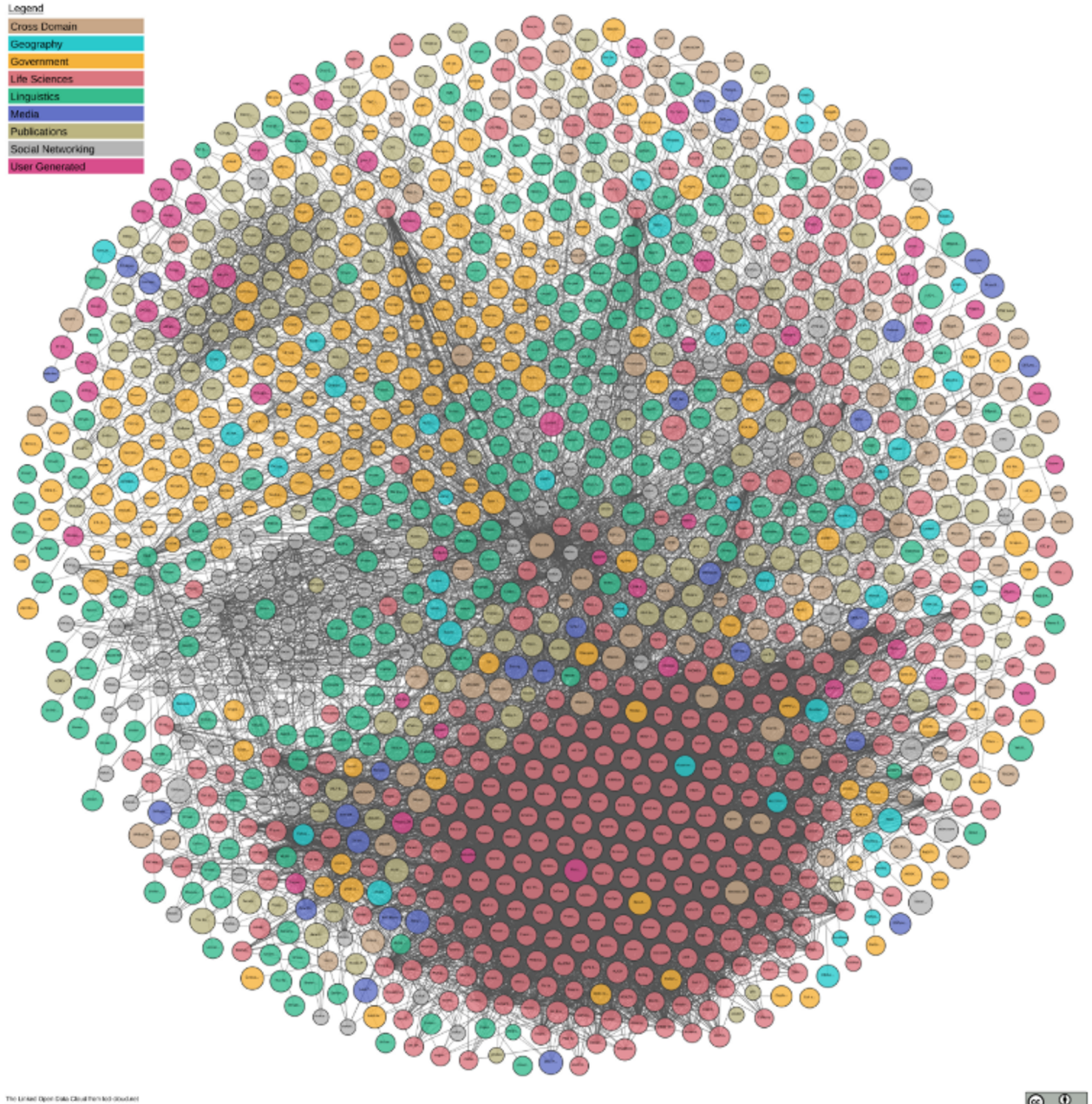
standards for publishing, sharing and querying
facts, expert knowledge and services



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

W3C Semantic Web

The Linked Open Data Cloud



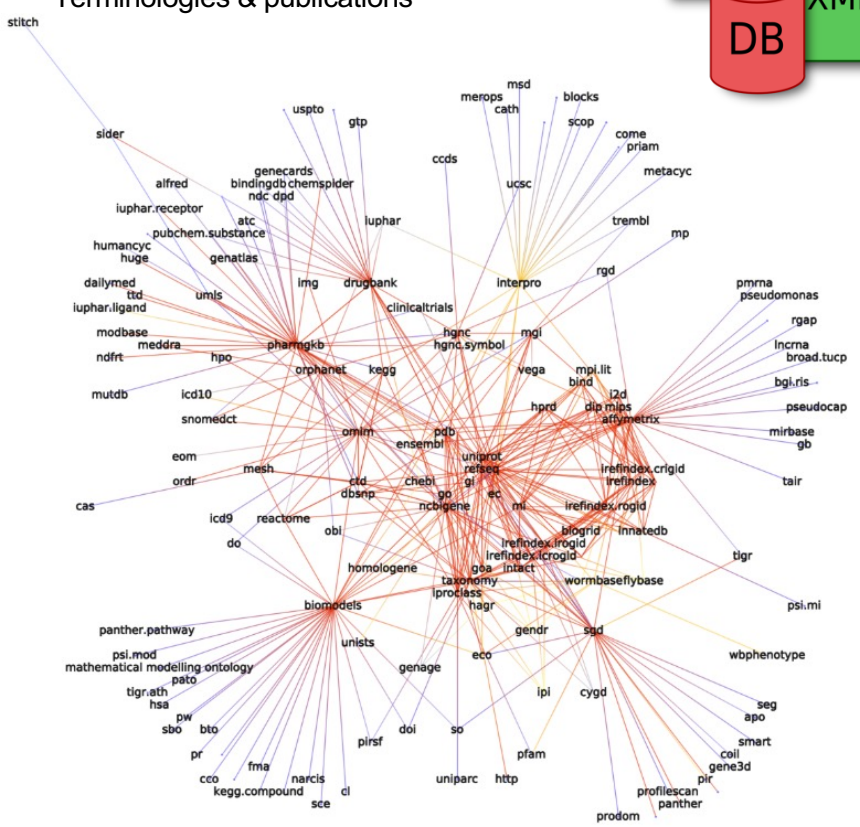
https://lod-cloud.net/

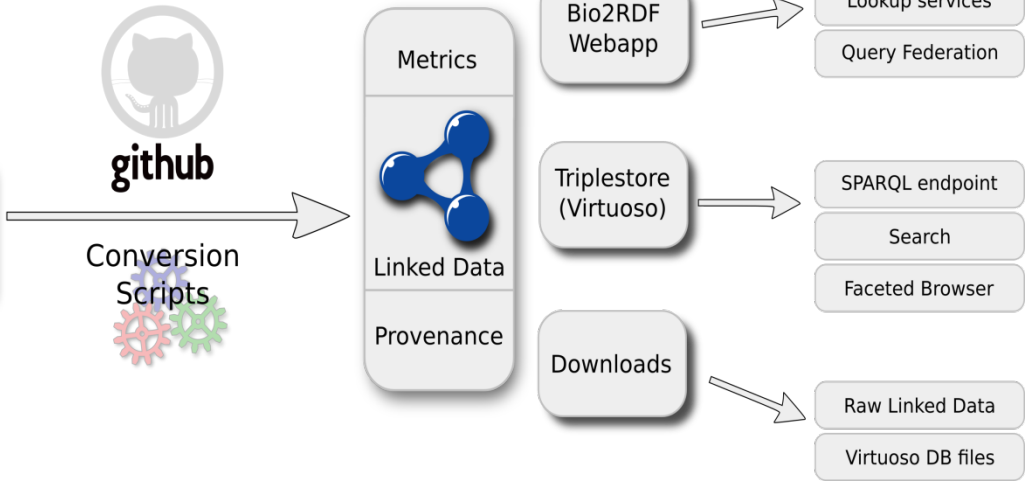
BIO2RDF

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

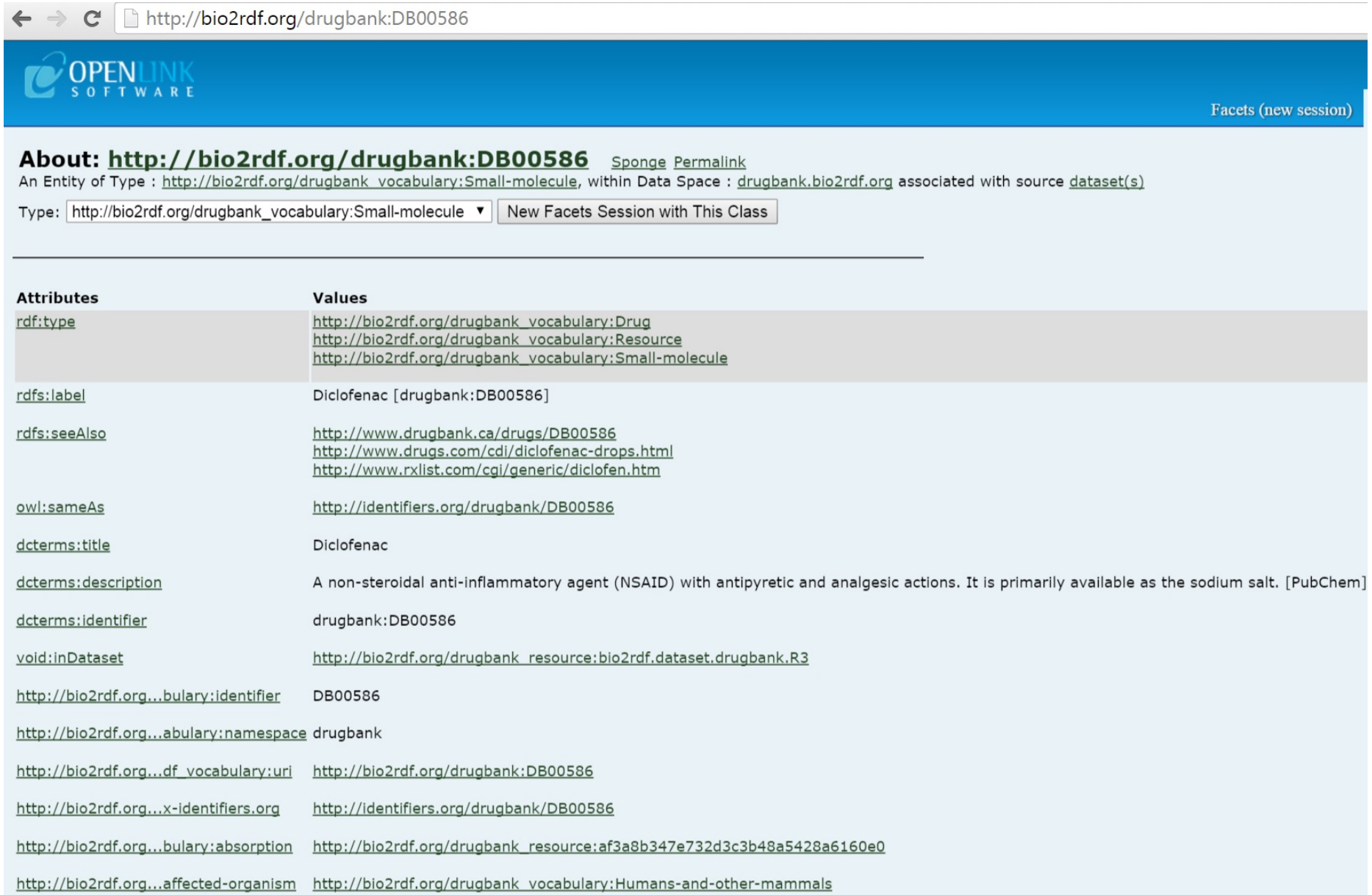




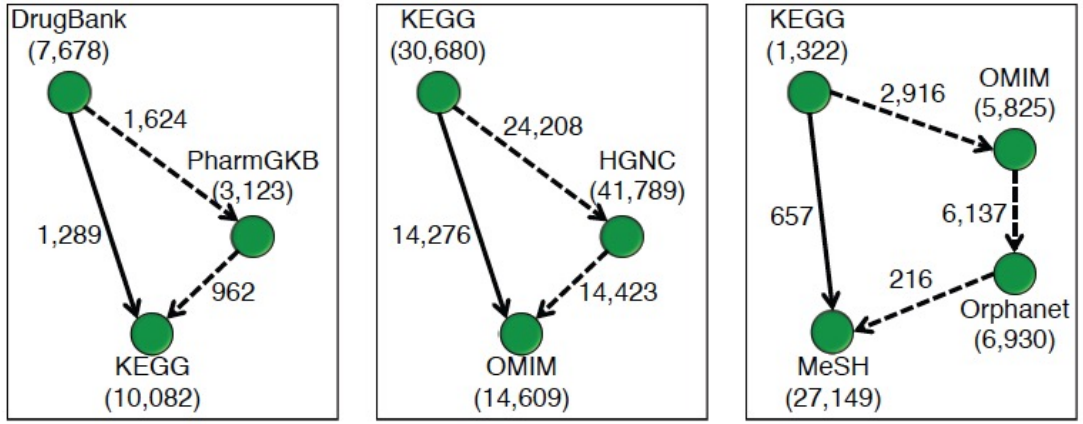
W3C Semantic Web

- **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


http://bio2rdf.org/drugbank:DB00586
OPENLINK SOFTWARE
Facets (new session)
About: http://bio2rdf.org/drugbank:DB00586 Sponge Permalink
An Entity of Type : http://bio2rdf.org/drugbank_vocabulary:Small-molecule, within Data Space : drugbank.bio2rdf.org associated with source dataset(s)
Type: http://bio2rdf.org/drugbank_vocabulary:Small-molecule
New Facets Session with This Class
Attributes
Values
rdf:type
http://bio2rdf.org/drugbank_vocabulary:Drug
http://bio2rdf.org/drugbank_vocabulary:Resource
http://bio2rdf.org/drugbank_vocabulary:Small-molecule
rdfs:label
Diclofenac [drugbank:DB00586]
rdfs:seeAlso
http://www.drugbank.ca/drugs/DB00586
http://www.drugs.com/cdi/diclofenac-drops.html
http://www.rxlist.com/cgi/generic/diclofen.htm
owl:sameAs
http://identifiers.org/drugbank/DB00586
dcterms:title
Diclofenac
dcterms:description
A non-steroidal anti-inflammatory agent (NSAID) with antipyretic and analgesic actions. It is primarily available as the sodium salt. [PubChem]
dcterms:identifier
drugbank:DB00586
void:inDataset
http://bio2rdf.org/drugbank_resource:bio2rdf.dataset.drugbank.R3
http://bio2rdf.org...bulary:identifier
DB00586
http://bio2rdf.org...abulary:namespace
drugbank
http://bio2rdf.org...df_vocabulary:uri
http://bio2rdf.org/drugbank:DB00586
http://bio2rdf.org...x-identifiers.org
http://identifiers.org/drugbank/DB00586
http://bio2rdf.org...bulary:absorption
http://bio2rdf.org/drugbank_resource:af3a8b347e732d3c3b48a5428a6160e0
http://bio2rdf.org...affected-organism
http://bio2rdf.org/drugbank_vocabulary:Humans-and-other-mammals

graph methods for data quality to find mismatches and discover new links



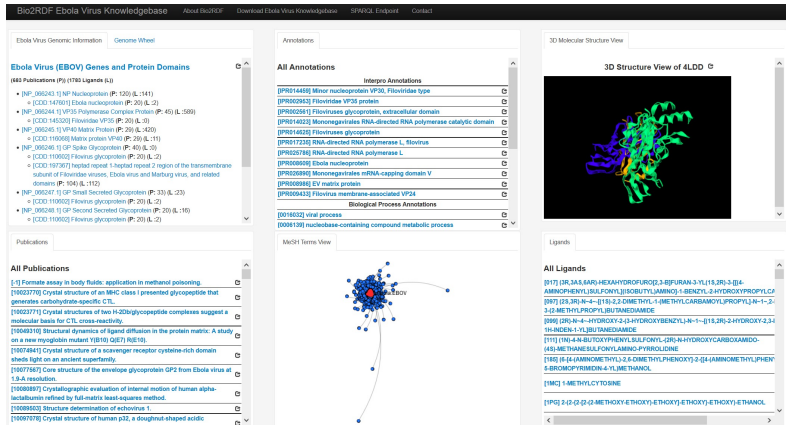
(a) Drugs (b) Human genes (c) Diseases

	Direct links	Transitive paths	Identical ending entities	Different entities	Missing direct	Missing transitive	Total
Drugs	1,289	954	946	6	2	343	1,297
Human genes	14,276	14,250	14,236	5	9	40	14,290
Diseases	657	33	8	18	7	649	682

Fig. 3. Transitivity analysis of entity links: (i) the value in each parenthesis denotes the number of entities given a specified topic; and (ii) the solid arcs represent direct links between entities while the dashed arcs form transitive paths. The value on each arc denotes the number of entity links from one dataset to the other.

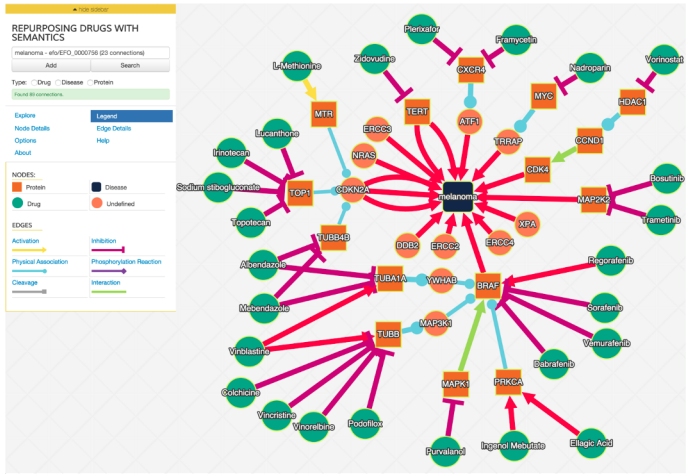
W Hu, H Qiu, **M Dumontier**. Link Analysis of Life Science Linked Data. International Semantic Web Conference (2) 2015: 446-462.

Custom Knowledge Portal: EbolaKB



https://doi.org/10.1093/database/bav049

Exploration: drug-target-disease networks



https://doi.org/10.7717/peerj-cs.106

Information Retrieval: Phenotypes of knock-out mouse models for the targets of a selected drug

Endpoint : http://drugbank.bio2rdf.org/sparql

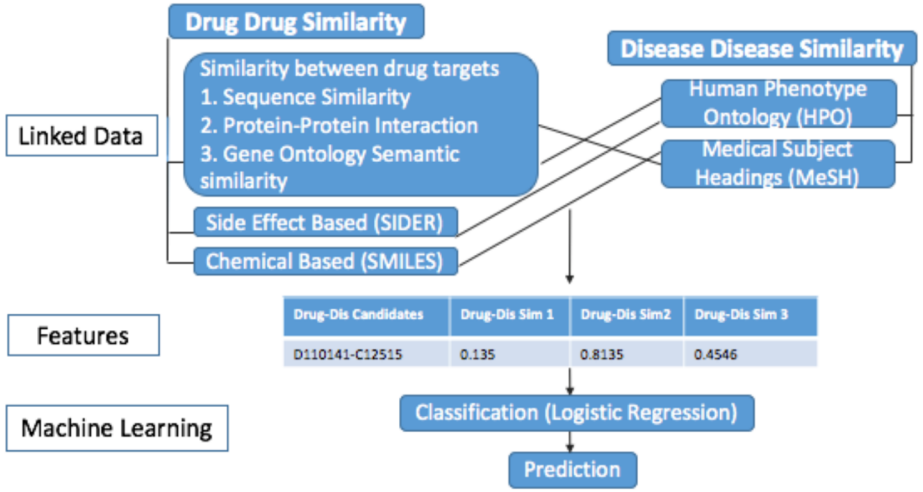
Output : Table

Configure request

```
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

Prediction: new uses for existing drugs

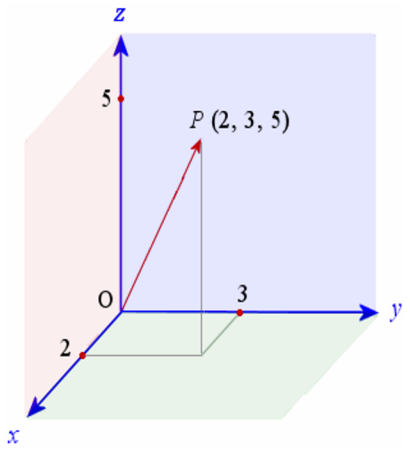


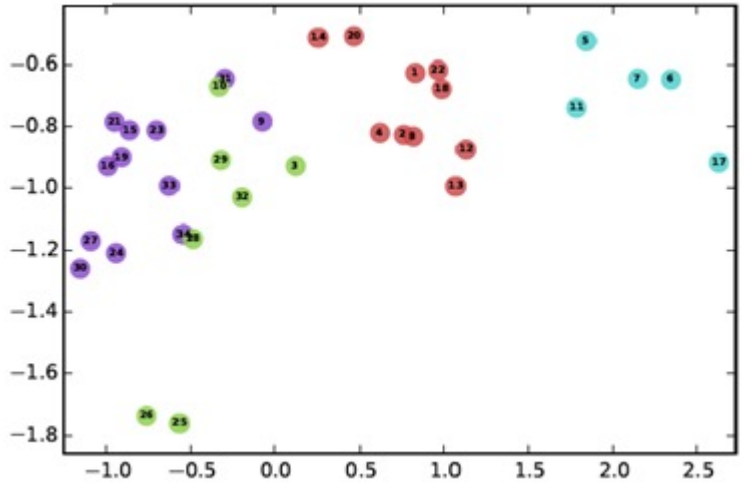
https://doi.org/10.7717/peerj-cs.281

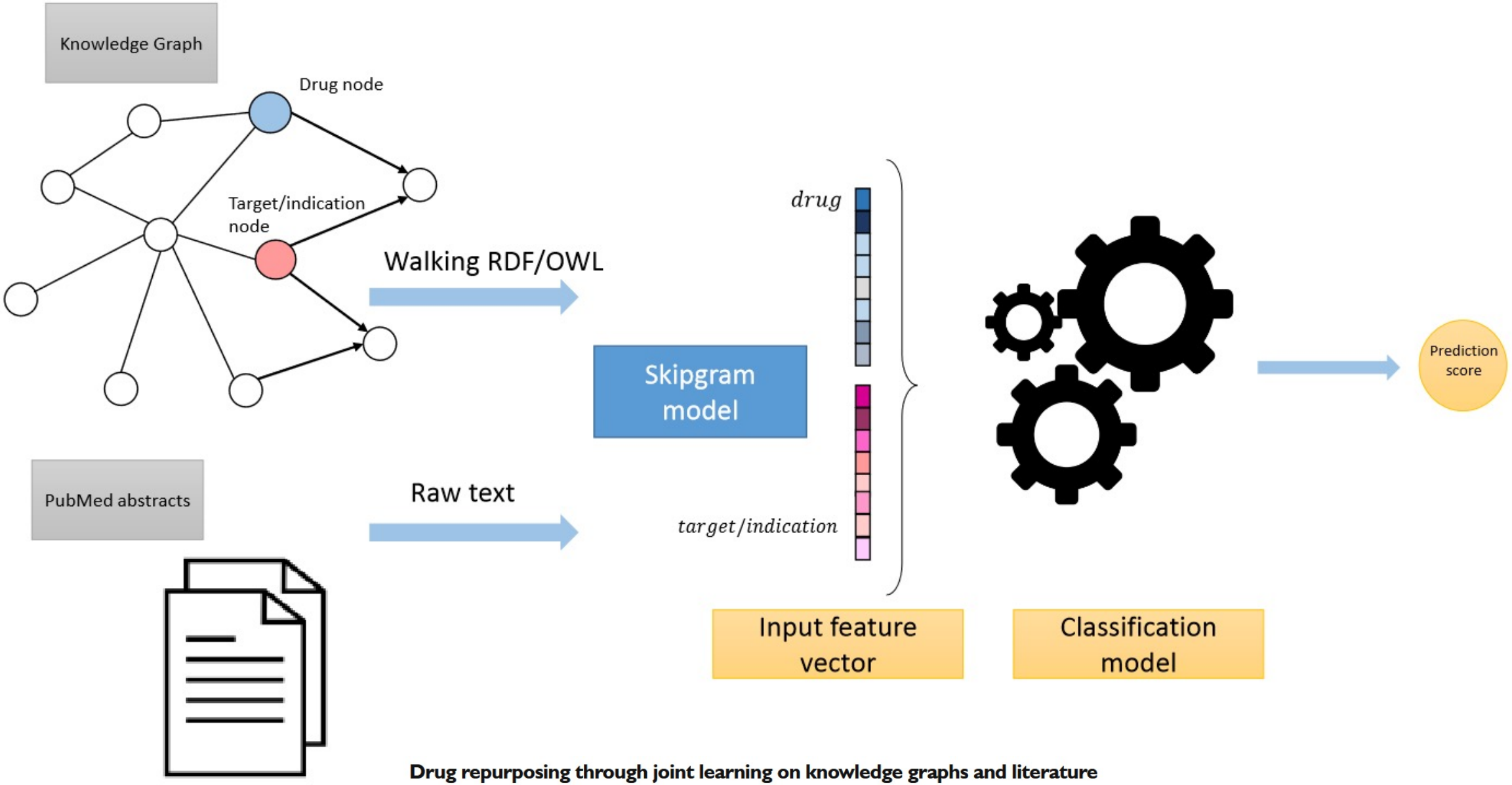
Knowledge Graph Embeddings

Map high-dimensional data into low-dimensional vector space.

- Represent entity/relation/graph with a vector, not just a symbol
- Capture similarity and semantic relationships
- Allow vector operations (addition, subtraction, etc.)







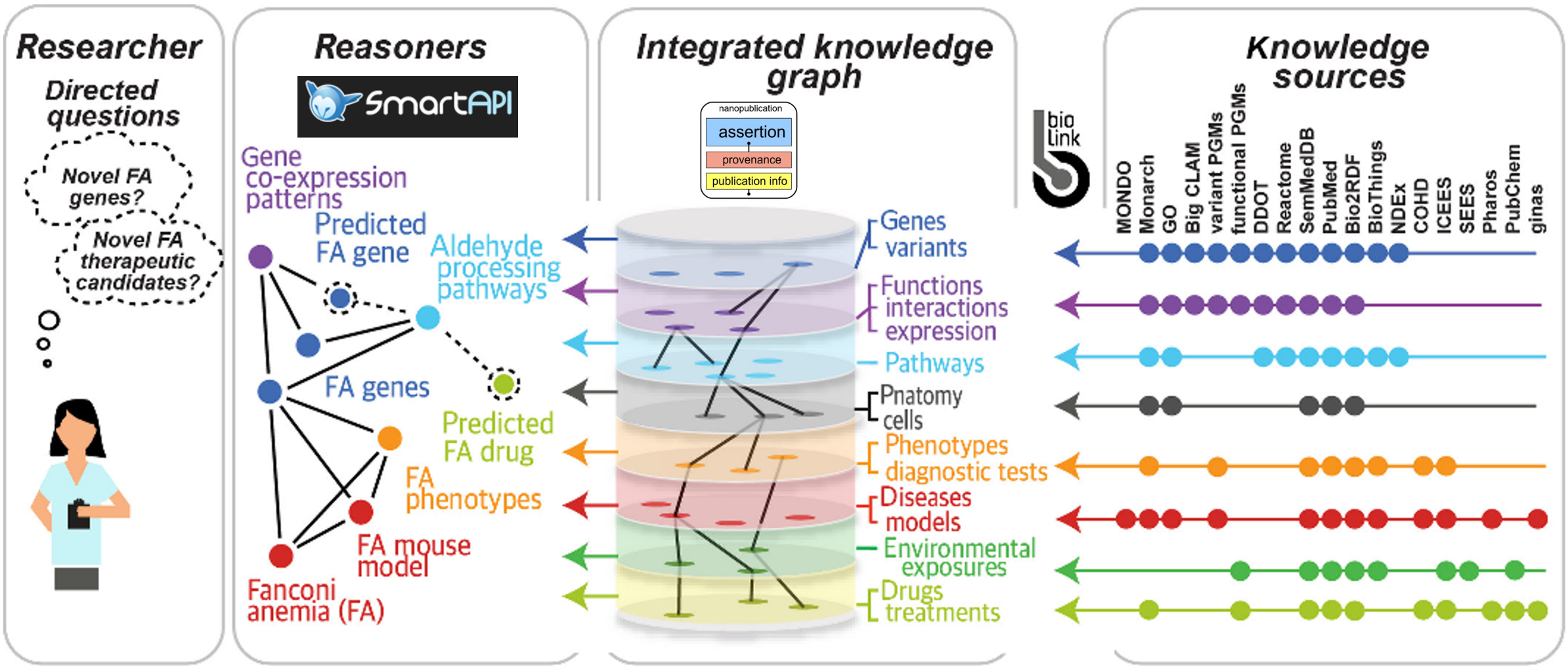
Drug repurposing through joint learning on knowledge graphs and literature

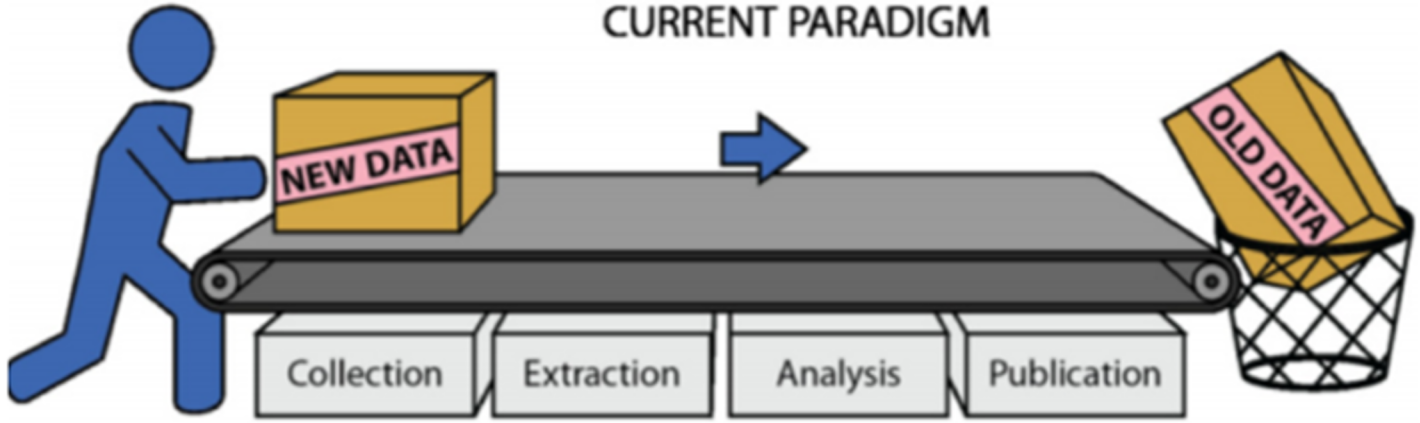
Mona Alshahrani, Robert Hoehndorf

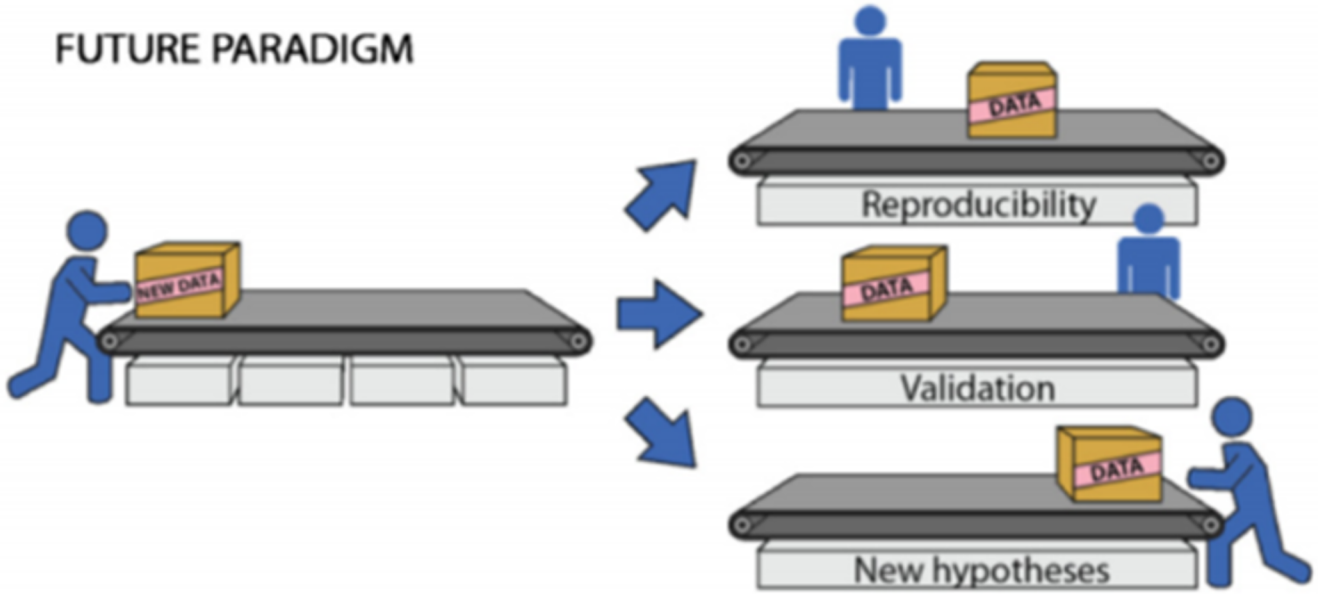
doi: https://doi.org/10.1101/385617

National Center for Advancing Translational Sciences

Biomedical Data Translator

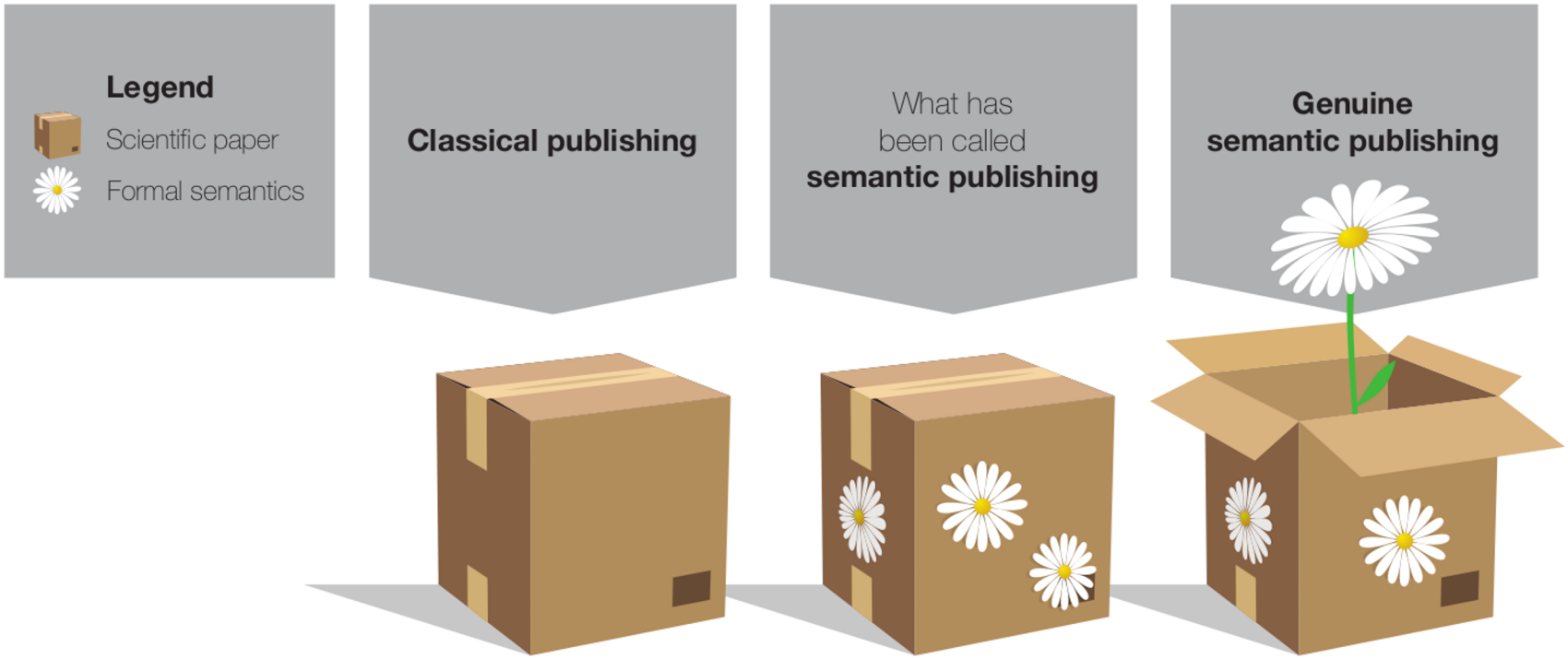






Lambin et al. Radiother Oncol. 2013. 109(1):159-64. doi: 10.1016/j.radonc.2013.07.007

Rethinking Publishing Data-Driven Research



Genuine Semantic Publishing

by Tobias Kuhn and Michel Dumontier

Content:

- as PDF
- as HTML/Dokieli
- as HTML/RASH
- as RDF/Turtle
- as RDF/TriG

Data Science. 2017 1(1-2):139-154. DOI: 10.3233/DS-170010
http://www.tkuhn.org/pub/sempub/

Data Science 5 (2022) 53–56
DOI 10.3233/DS-210047
IOS Press

A formalization of one of the main claims of "The FAIR Guiding Principles for scientific data management and stewardship" by Wilkinson et al. 2016[1]

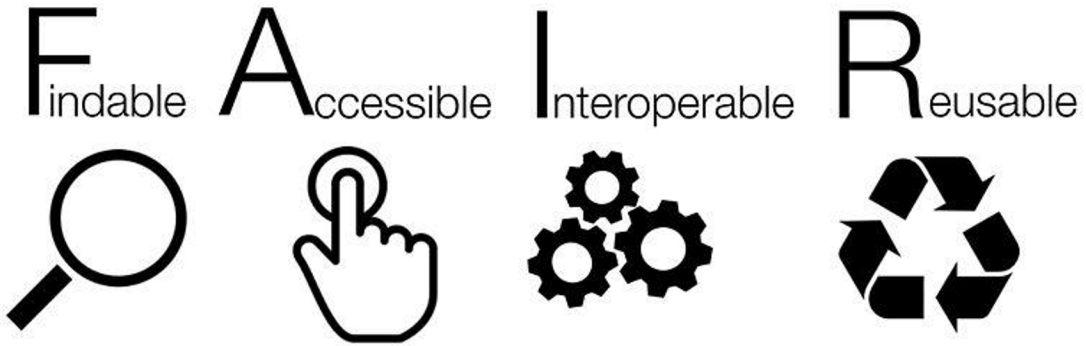
Michel Dumontier
Maastricht University, The Netherlands
E-mails: michel.dumontier@maastrichtuniversity.nl, michel.dumontier@gmail.com; ORCID: https://orcid.org/0000-0003-4727-9435

Summary

FAIR data requires certain features for machine discovery and reuse, and offers new research possibilities that may reduce effort and improve transparency and confidence in published research

FAIR KGs aim to provide semantically annotated, standardized, and AI-ready data

AI technologies, coupled with semantics, will enable **researchers to exploit an emerging Internet of FAIR data and services in a (semi)automated manner**, and hence to accelerate discovery in biomedicine and in other disciplines, and to help realize the unforeseen value of existing data.



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

Website: http://maastrichtuniversity.nl/ids

Maastricht University