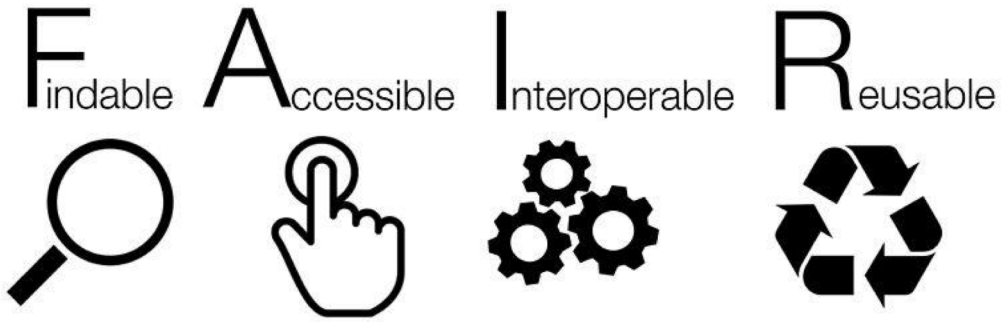
Data-Driven Discovery Science with FAIR Knowledge Graphs



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

Maastricht University

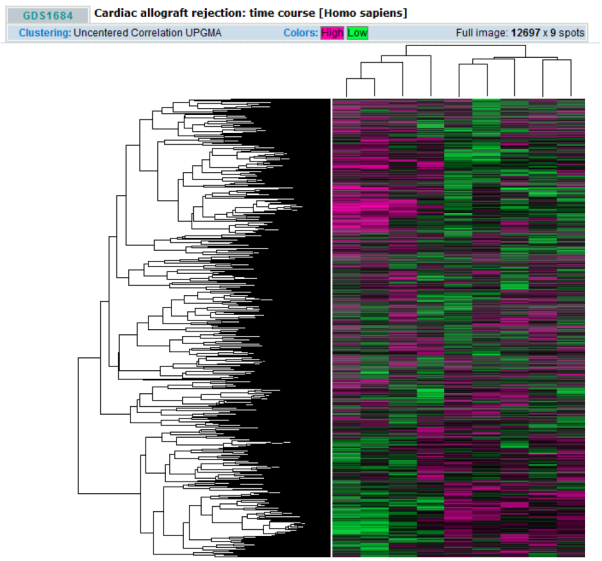
CC BY

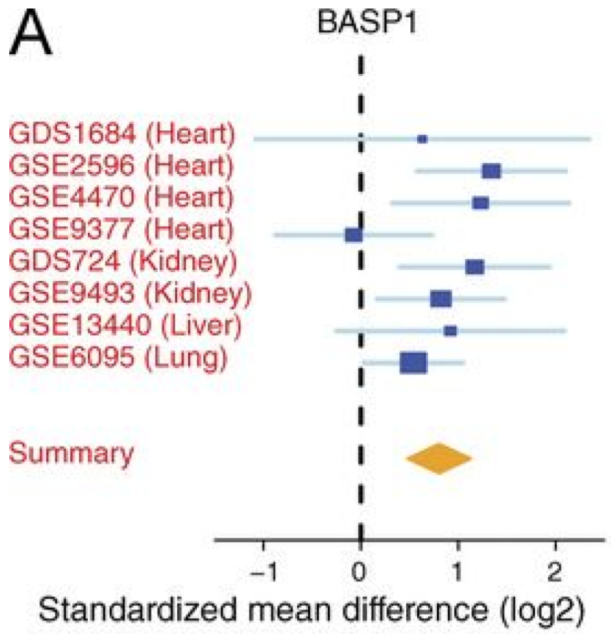
Vast amounts of (open) data are enabling and supporting Discovery Science

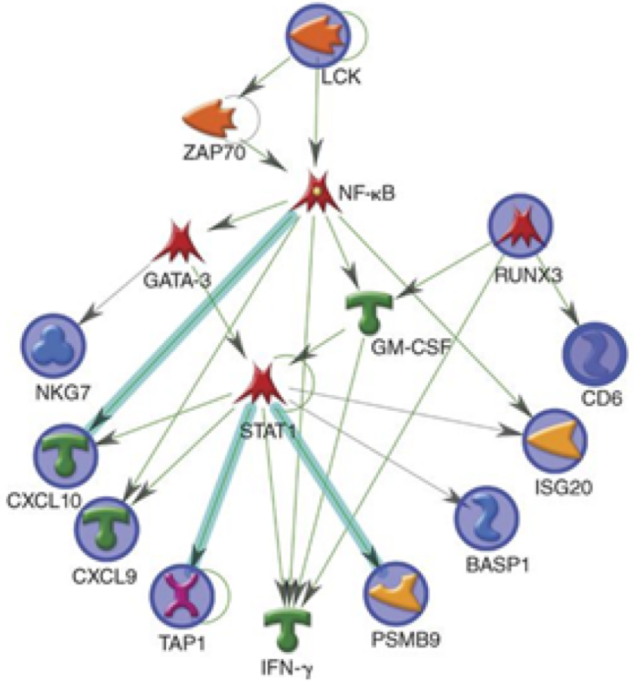
A common rejection module (CRM) for acute rejection across multiple organs identifies novel therapeutics for organ transplantation

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**

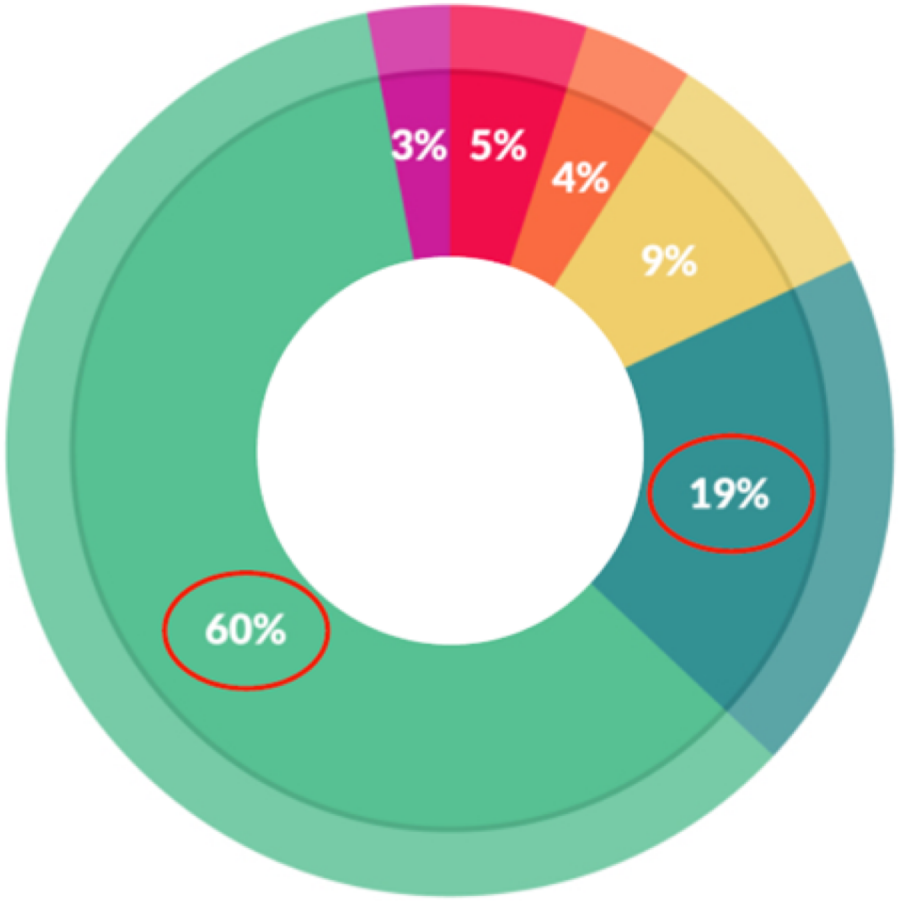
Key Observations:

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**

iNSiGHT

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.

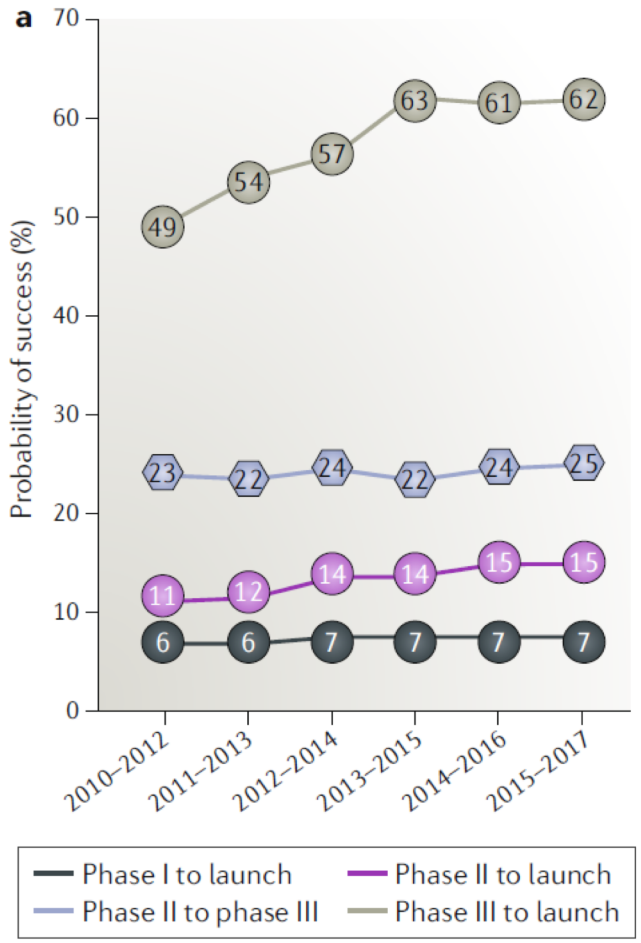
- John Ioannidis, Stanford University

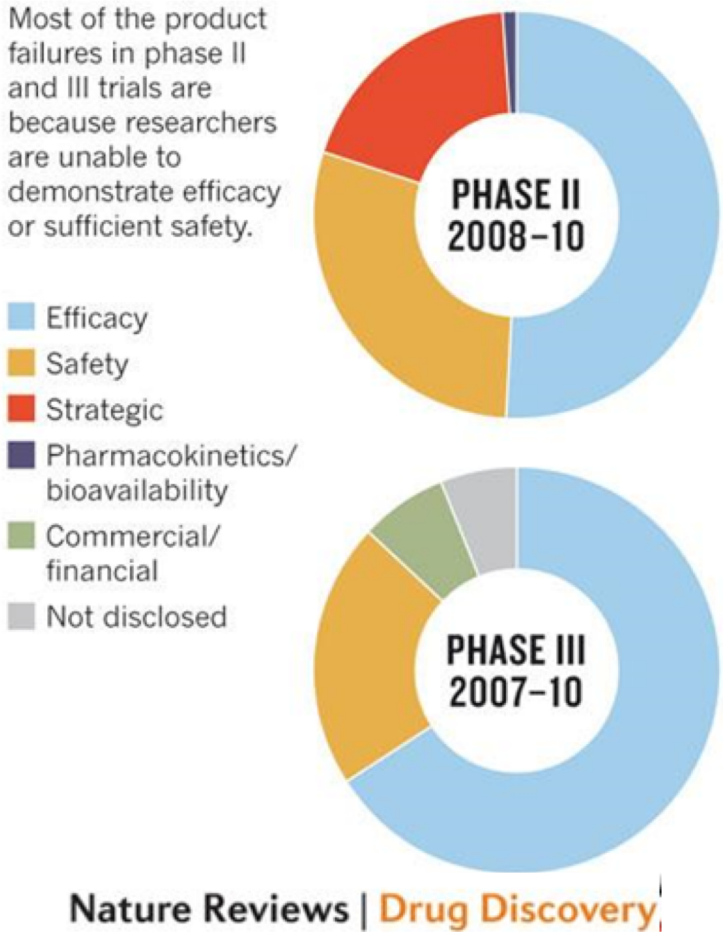
PLoS Med 2005;2(8): e124.

THE CLINICAL-TRIAL CLIFF

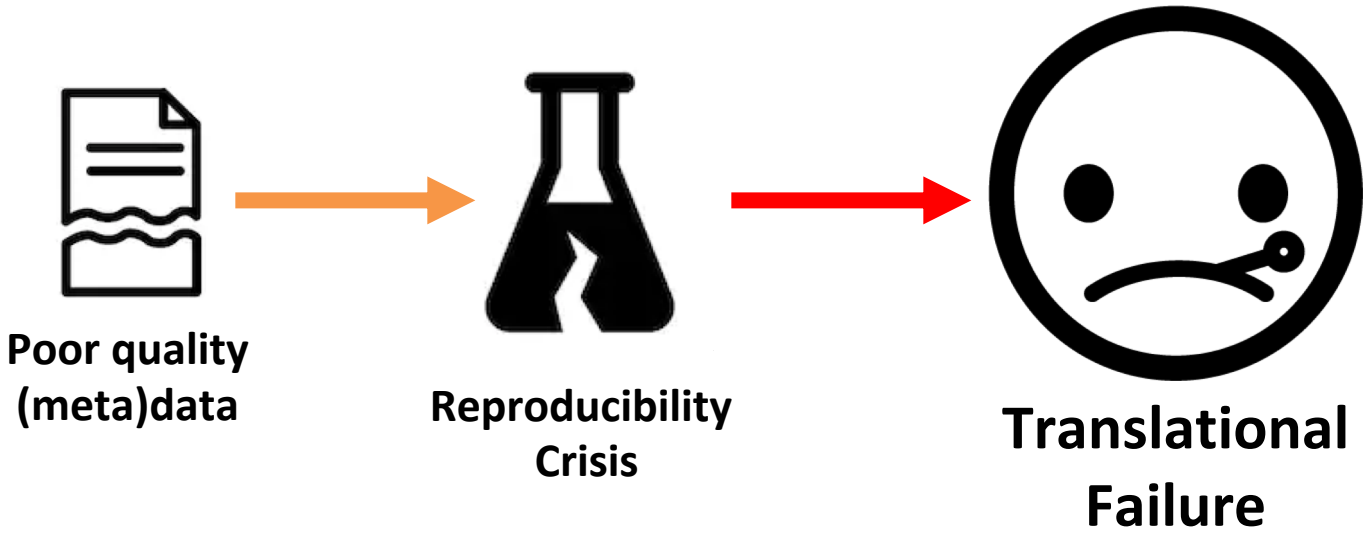
Drug companies are removing more compounds from the pipeline at all levels of testing than ever before.

Probability to launch



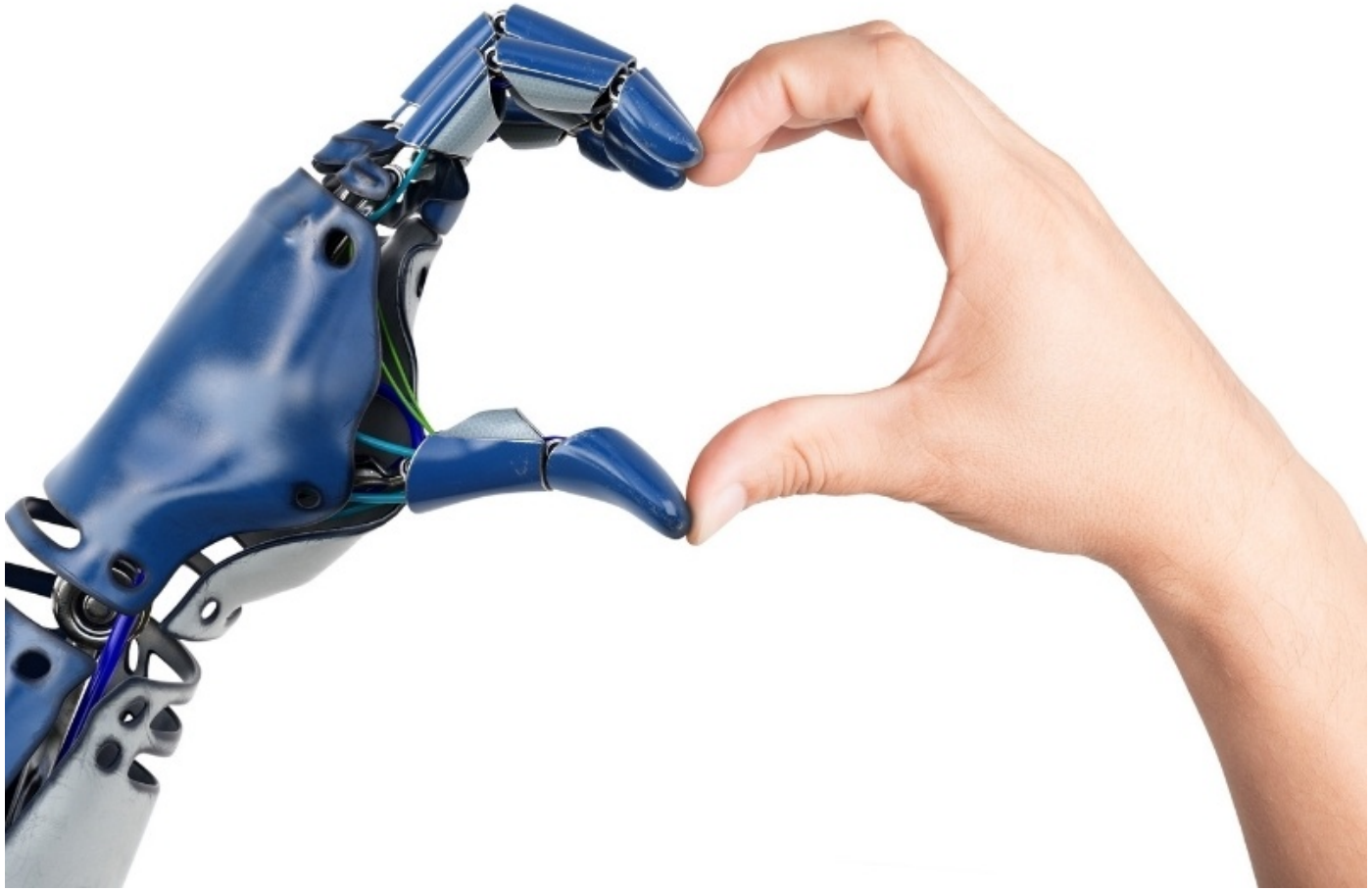


Nature Reviews Drug Discovery **18**, 495-496 (2019)
https://doi.org/10.1038/d41573-019-00074-z


Poor quality
(meta)data
Reproducibility
Crisis
Translational
Failure

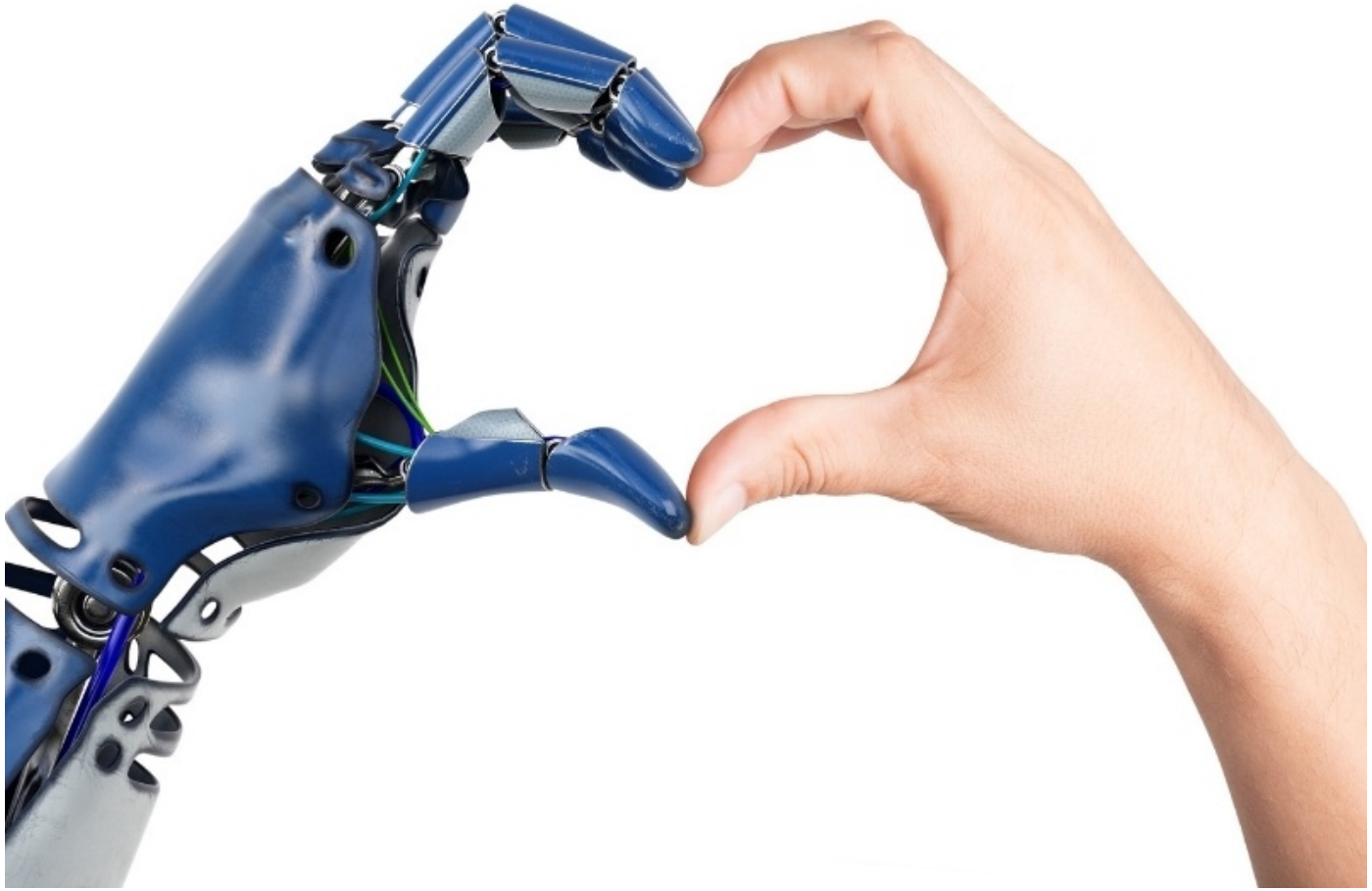
It’s time to completely rethink how we perform and document empirical research

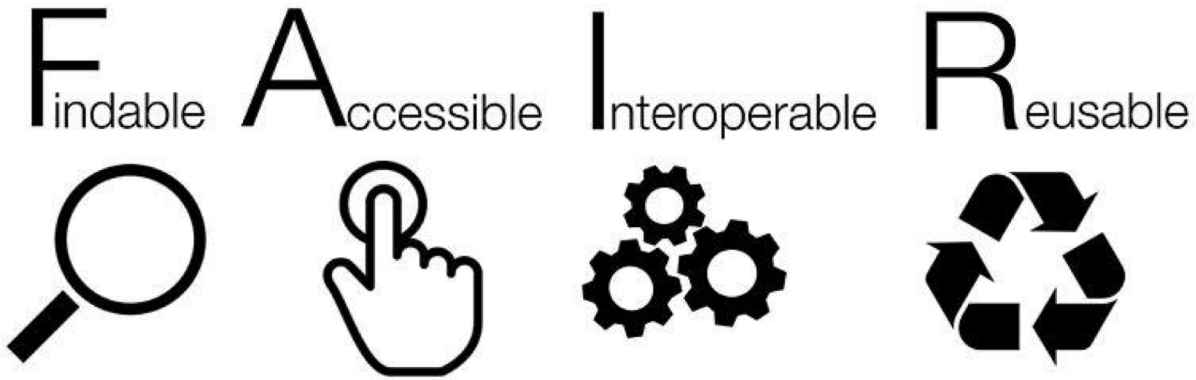
Human Machine collaboration is crucial to our future success

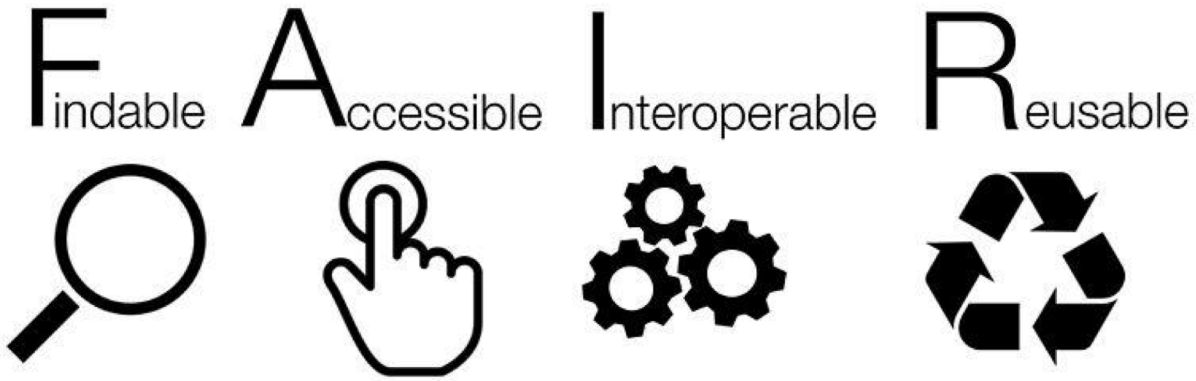


Machines

need to be able to discover and reuse data




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

402k Accesses | **3123** Citations | **1939** Altmetric | Metrics

This article is in the 99th percentile (ranked 42nd) of the 273,306 tracked articles of a similar age in all journals and the 1st percentile (ranked 1st) of the 1 tracked articles of a similar age in *Scientific Data*

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

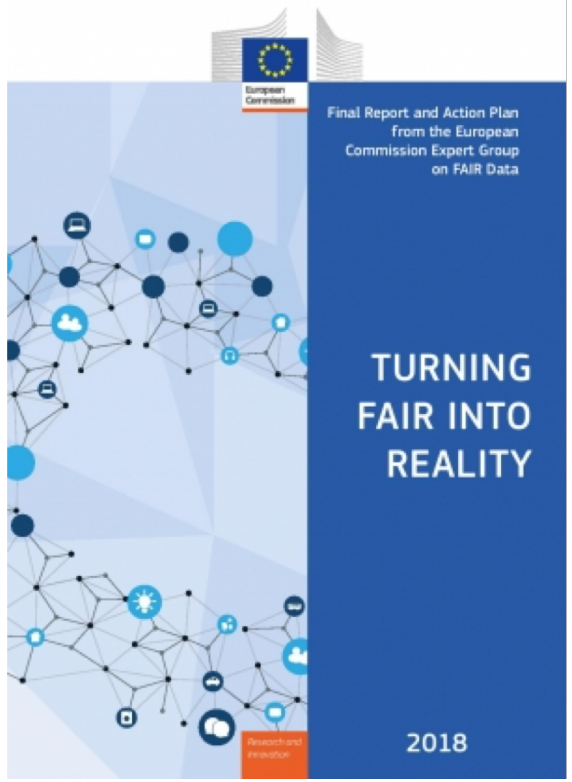
G20 Leaders' Communique Hangzhou Summit

Hangzhou, 5 September 2016

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

Annex 4:
G7 Expert Group on Open Science

Turin, Italy, September 28, 2017



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

FAIR in a nutshell

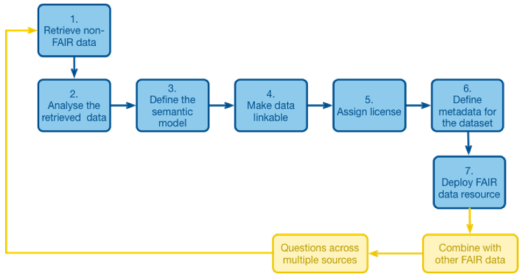
FAIR aims to create **social, scientific** and **economic impact** by facilitating the discovery and reuse of **digital resources** through a set of requirements:

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

Learn how to create and publish FAIR data

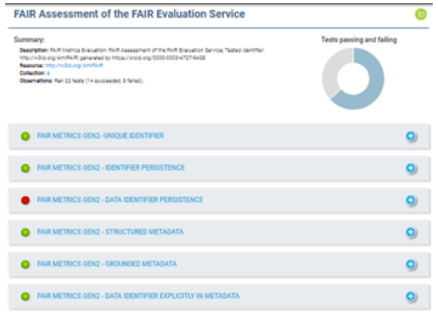
RDMkit

GO FAIR



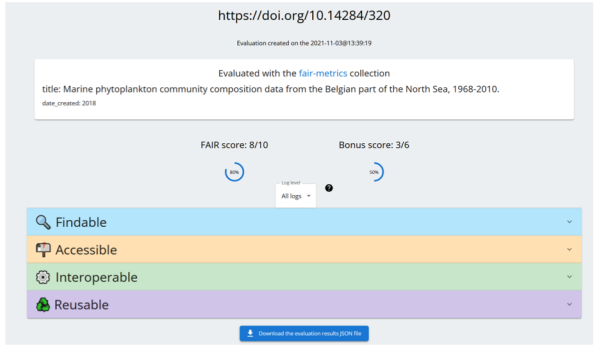
FAIR cookbook

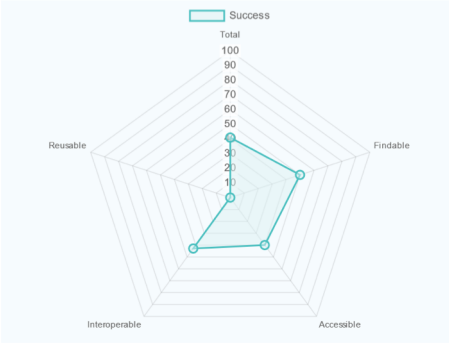
Tools to check the correctness of FAIR Implementations



FAIR Evaluator

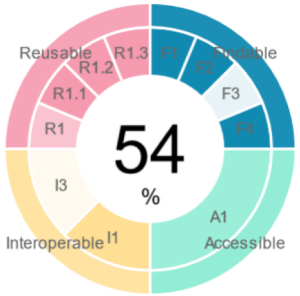
https://w3id.org/AmIFAIR





FAIR Checker

https://fair-checker.france-bioinformatique.fr/base_metrics



F-UJI

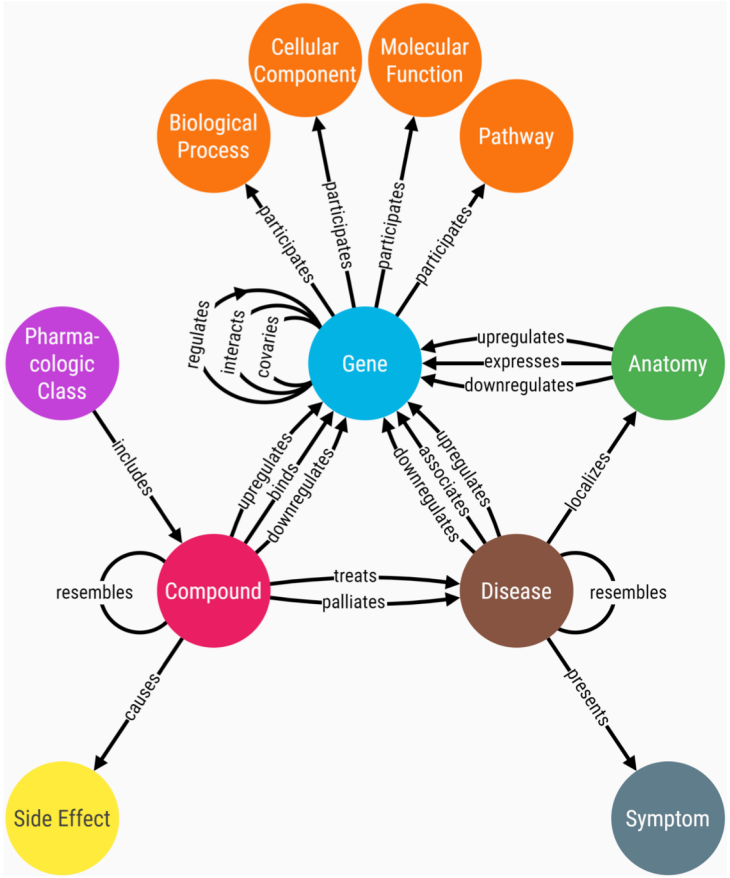
https://www.f-uji.net

FAIR Enough
https://fair-enough.semanticscience.org

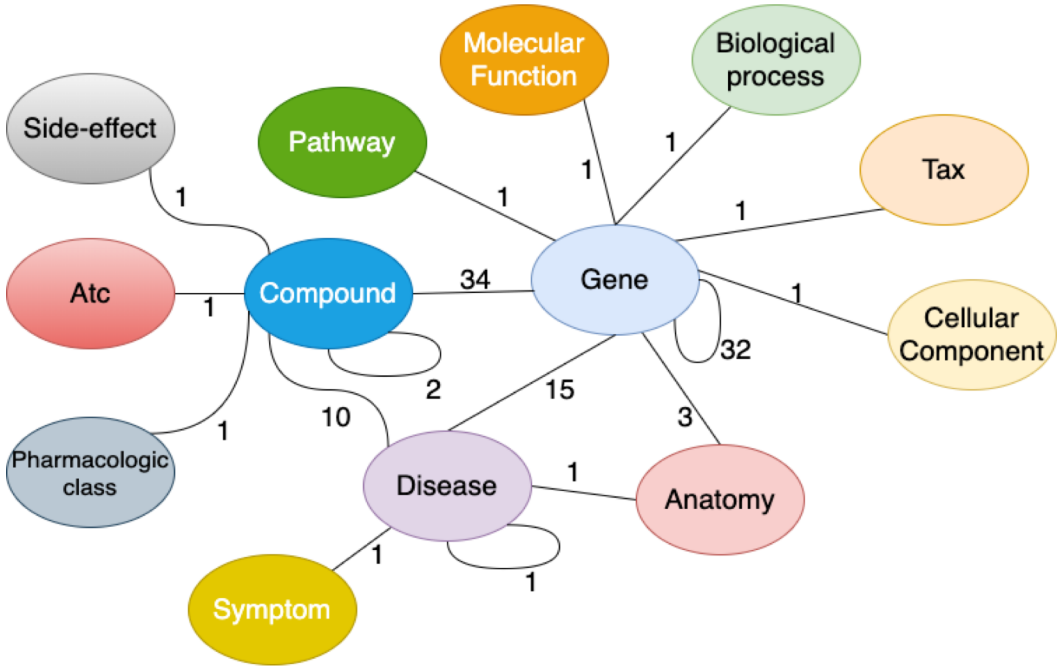
Towards FAIR Knowledge Graphs

A number of biomedical KGs have been developed

See recent review Callahan et al. 2020. Knowledge-Based Biomedical Data Science.

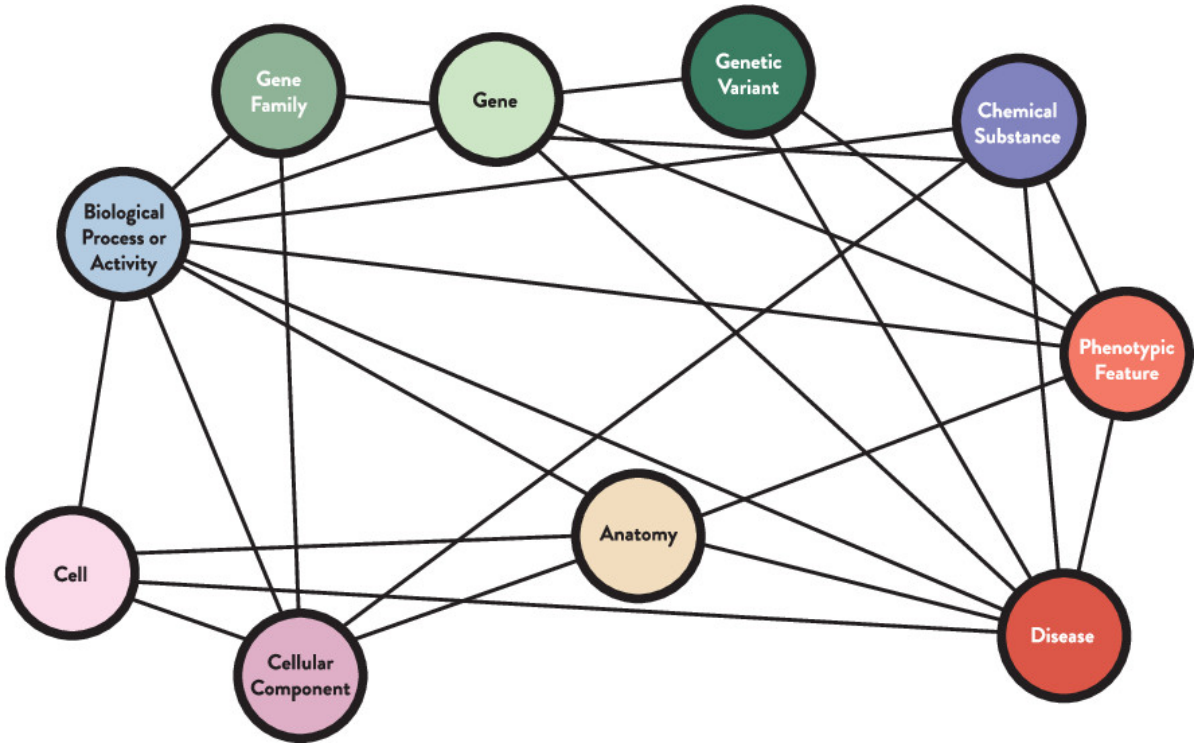


metagraph for [het.io]

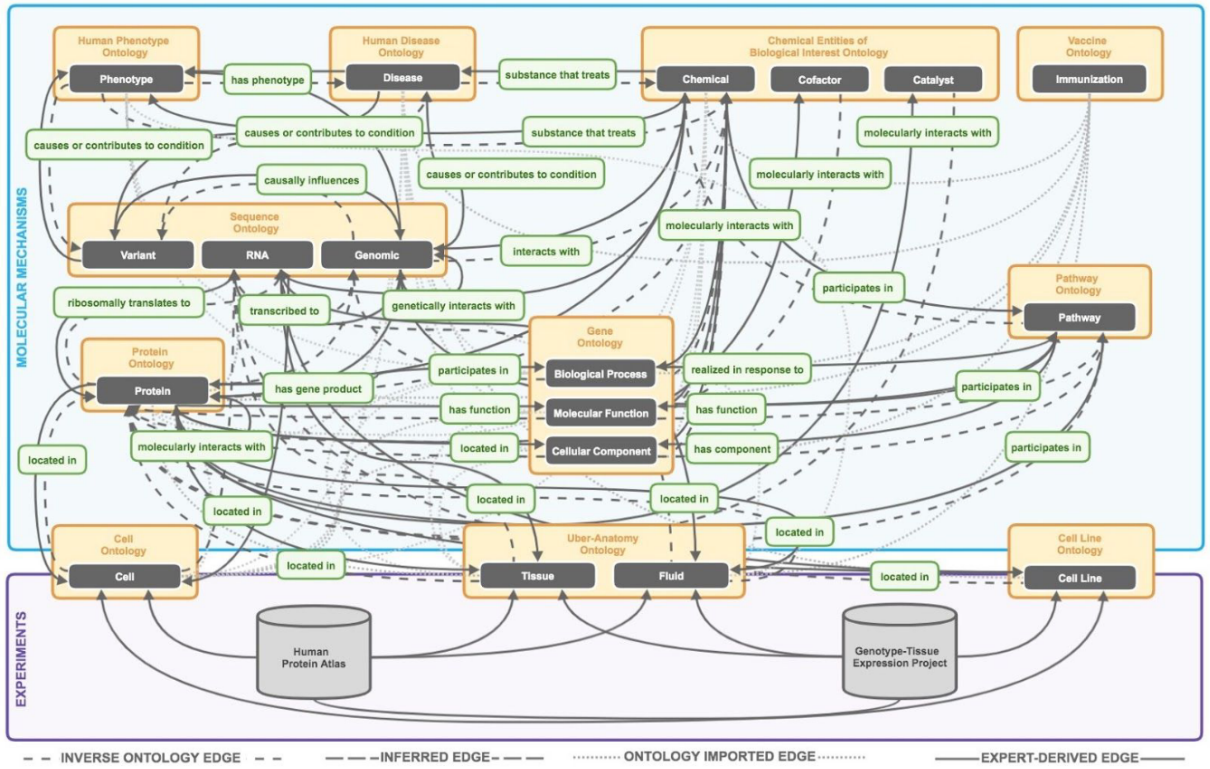


metagraph for [DRKG]

Including ontology guided Knowledge Graphs



ROBOKOP
https://robokopkg.renci.org



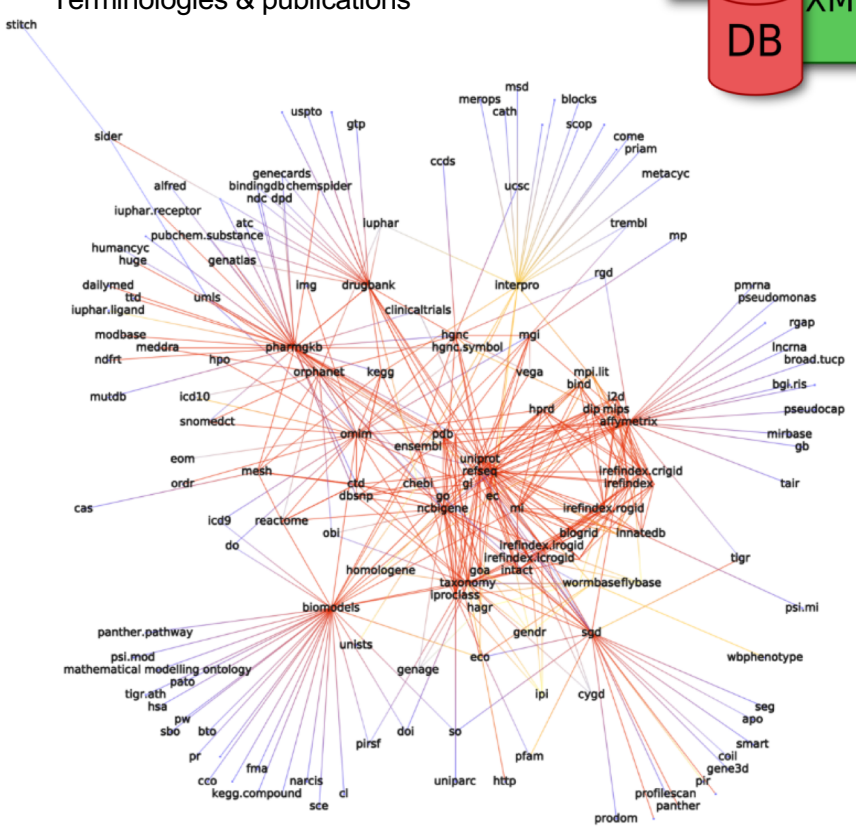
PheKnowLator. Callahan et al.
https://github.com/callahantiff/PheKnowLator

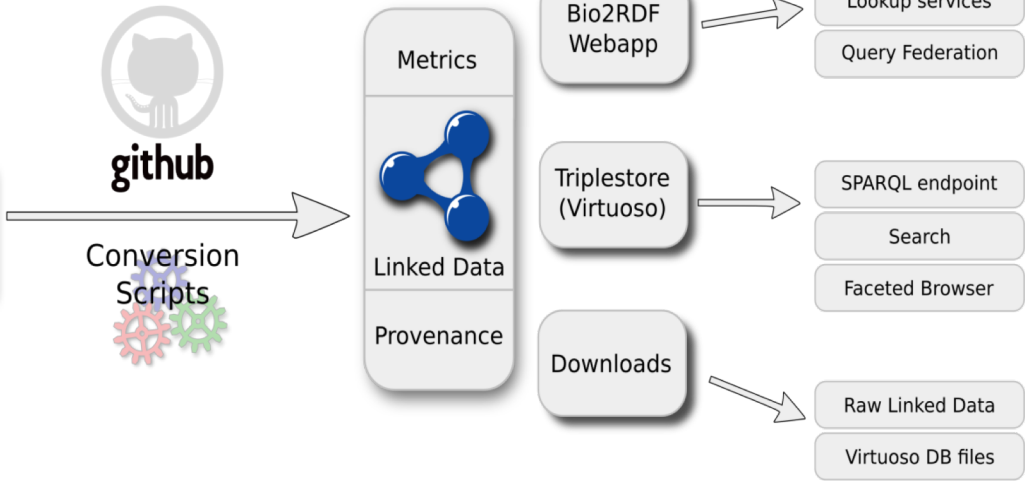
BIO2RDF

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

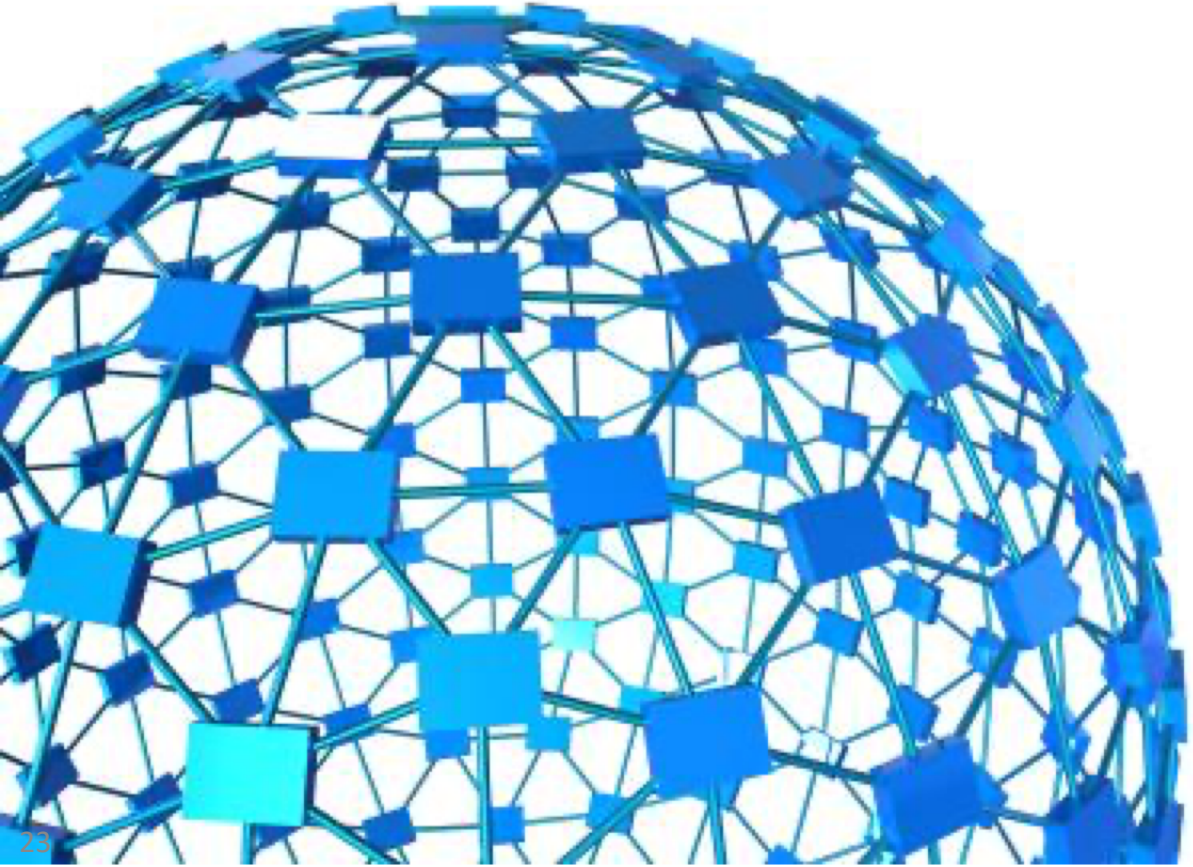
- **30+** biomedical data sources
- **10B+** interlinked statements
- EBI, SIB, NCBI, DBCLS, NCBO, and many others produce this content

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

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

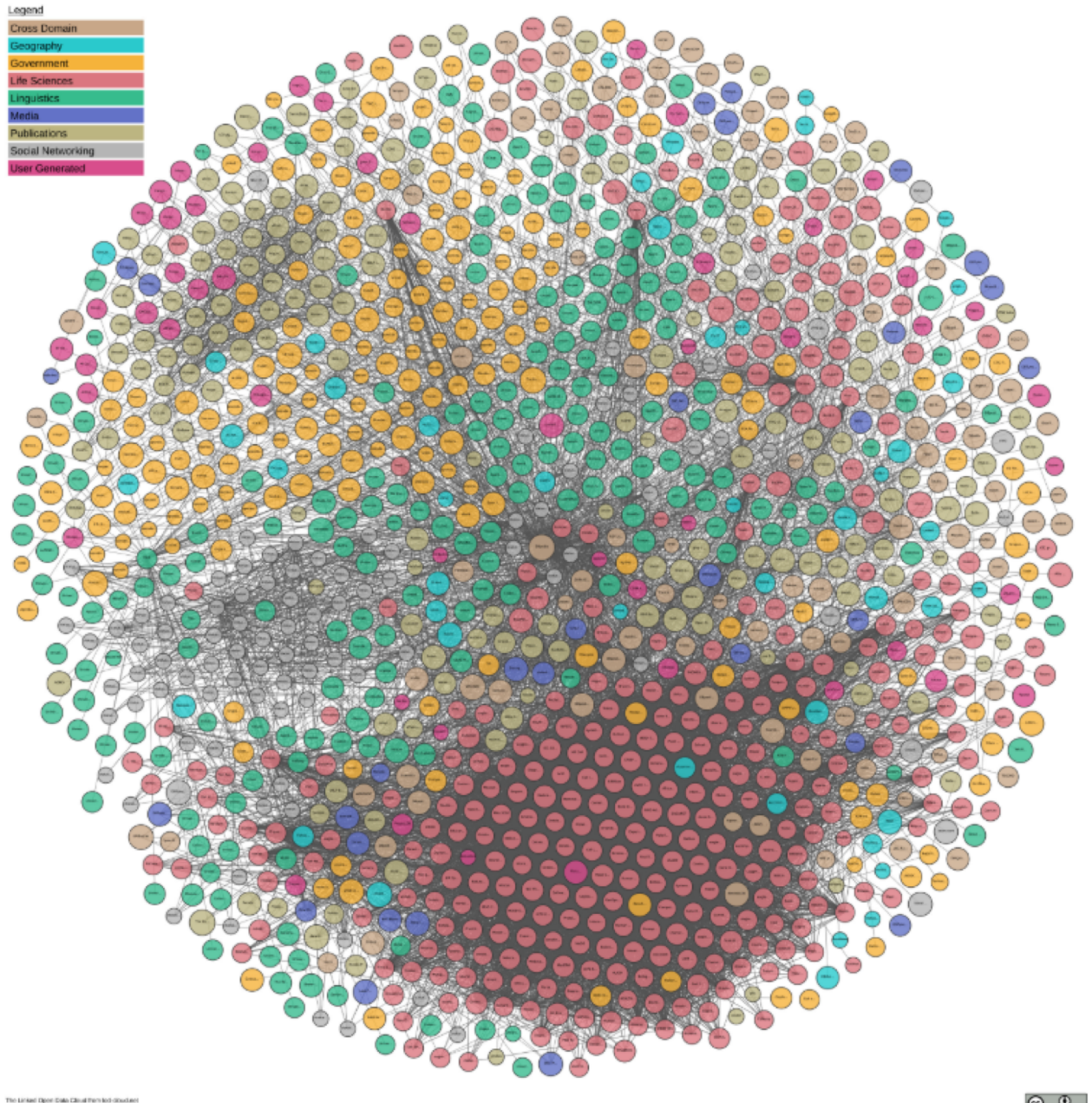
standards for publishing, sharing and querying
facts, expert knowledge and services

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

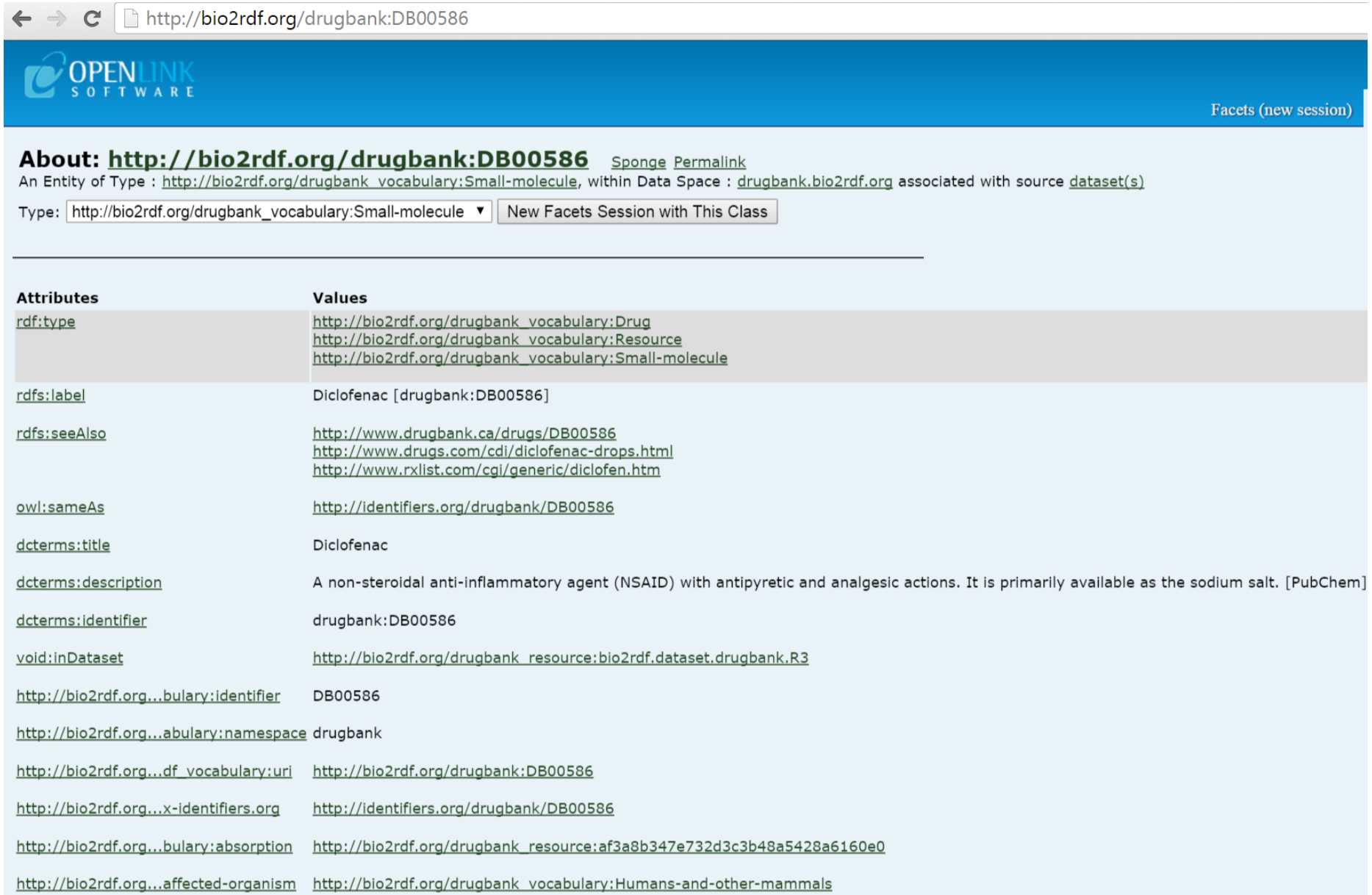


W3C Semantic Web

The Linked Open Data Cloud



https://lod-cloud.net/


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

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

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

Output : Table

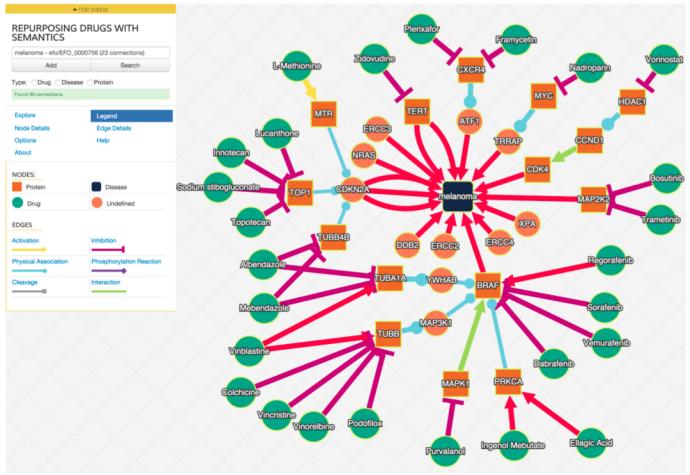
Configure request

```
PREFIX dct: <http://purl.org/dc/terms/>
SELECT DISTINCT ?phenotype_label
WHERE {
  SERVICE <http://drugbank.bio2rdf.org/sparql> {
        ?drug <http://bio2rdf.org/drugbank_vocabulary:target> ?target .
         FILTER(?drug = <http://bio2rdf.org/drugbank:DB00619>)
        ?target <http://bio2rdf.org/drugbank_vocabulary:x-hgnc> ?hgnc .
  }
  SERVICE <http://hgnc.bio2rdf.org/sparql> {
    ?hgnc <http://bio2rdf.org/hgnc_vocabulary:x-mgi> ?marker .
  }
  SERVICE <http://mgi.bio2rdf.org/sparql> {
        ?model <http://bio2rdf.org/mgi_vocabulary:marker> ?marker .
        ?model <http://bio2rdf.org/mgi_vocabulary:allele> ?all .
        ?all <http://bio2rdf.org/mgi_vocabulary:allele-attribute> ?allele_type .
        ?model <http://bio2rdf.org/mgi_vocabulary:phenotype> ?phenotypes .
        FILTER (str(?allele_type) = "Null/knockout")
  }
  SERVICE <http://bioportal.bio2rdf.org/sparql> {
        ?phenotypes rdfs:label ?phenotype_label .
  }
}
```

DRUGBANK HGNC HUGO Gene Nomenclature Committee MGI BioPortal

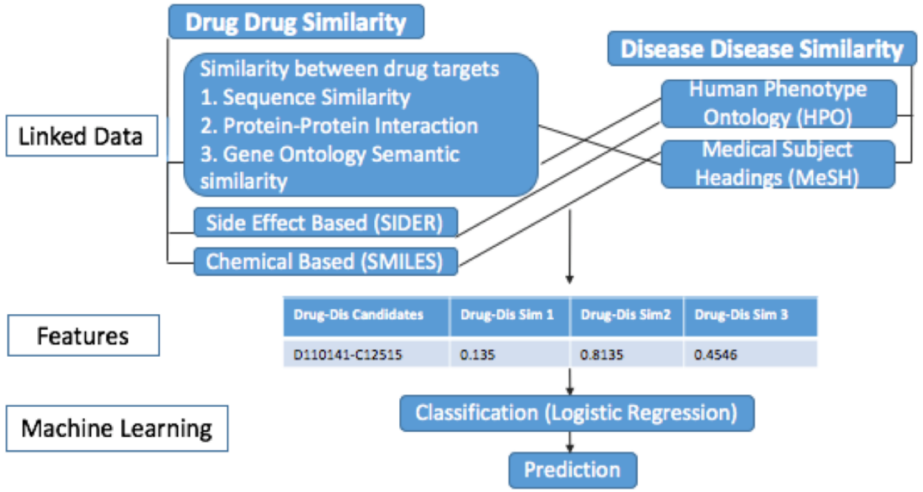
	phenotype_label
1	"hemorrhage [mp:0001914]"@en
2	"intracranial hemorrhage [mp:0001915]"@en
3	"perinatal lethality [mp:0002081]"@en

Exploration: drug-target-disease networks

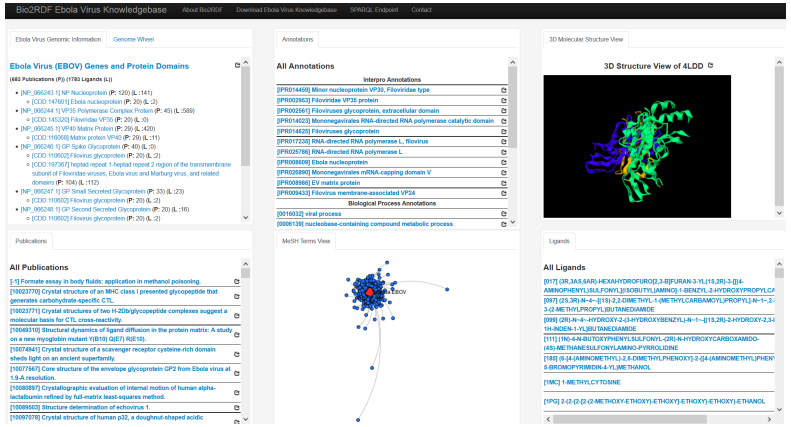


https://doi.org/10.7717/peerj-cs.106

Reproducible ML: new uses for existing drugs



Custom Knowledge Portal: EbolaKB



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

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

Predict new drug applications in a documented and reproducible manner

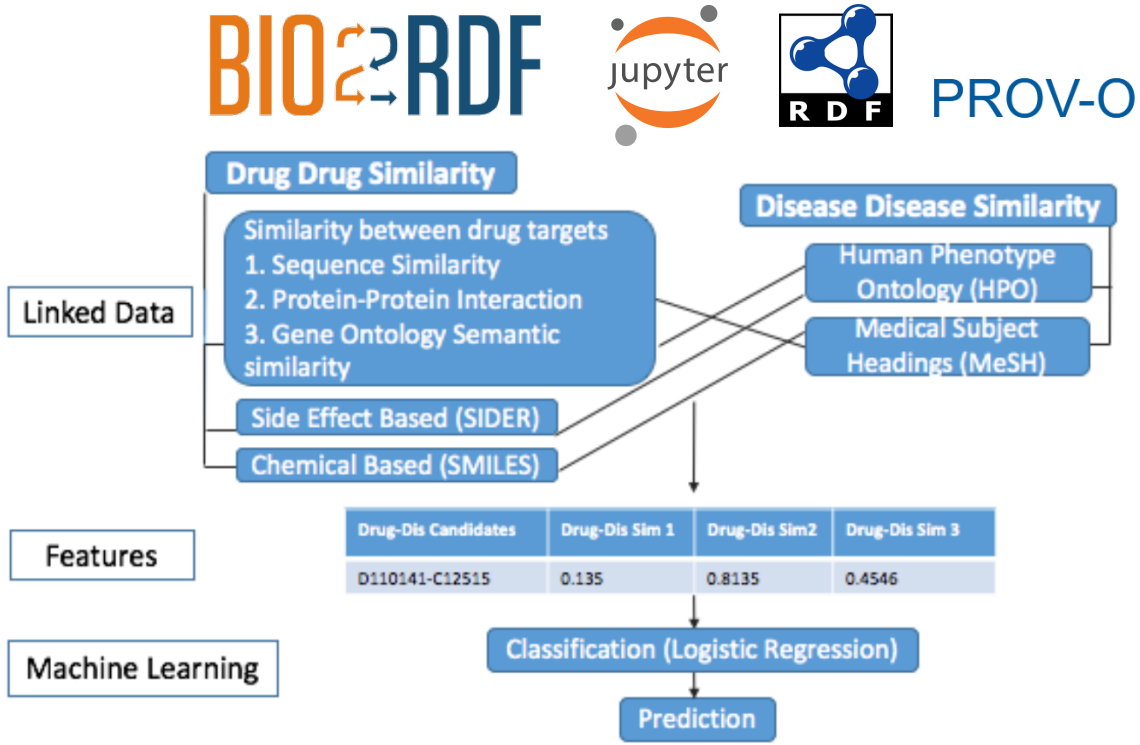
Mol Syst Biol. 2011; 7: 496.
Published online 2011 Jun 7. doi: 10.1038/msb.2011.26
PMCID: PMC3159979

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

Assaf Gottlieb,[1] Gideon Y Stein,[2,3] Eytan Ruppin,[1,2] and Roded Sharan[a,1]

AUC 0.90 across all therapeutic indications

Scripts not available. Feature tables available.
Not reproducible!



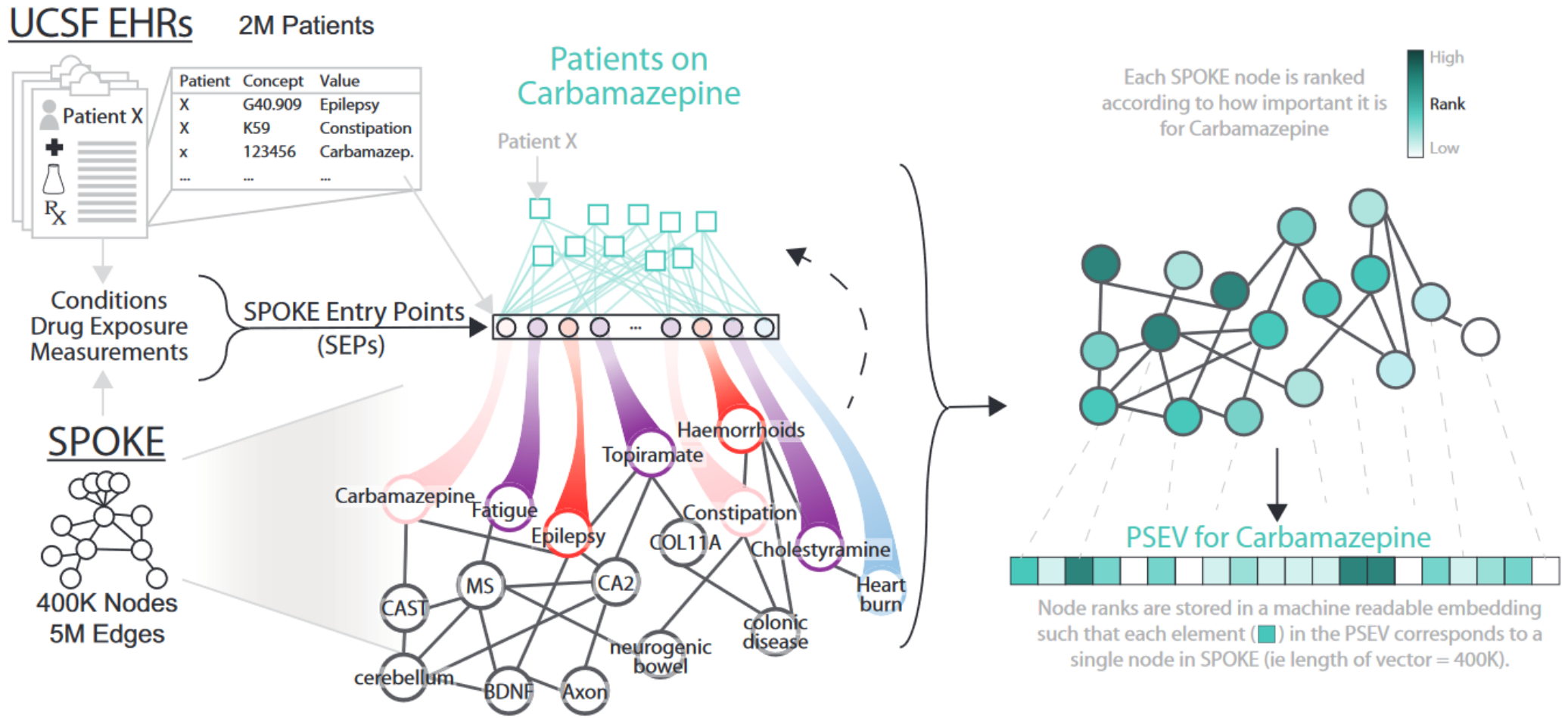
Towards FAIR protocols and workflows: the OpenPREDICT use case

Remzi Celebi[1,*], Joao Rebelo Moreira[2,*], Ahmed A. Hassan[3], Sandeep Ayyar[4], Lars Ridder[5], Tobias Kuhn[2] and Michel Dumontier[1]

Result: ROCAUC 0.83

Celebi R, Rebelo Moreira J, Hassan AA, Ayyar S, Ridder L, Kuhn T, Dumontier M. 2020. Towards FAIR protocols and workflows: the OpenPREDICT use case. PeerJ Computer Science 6:e281 https://doi.org/10.7717/peerj-cs.281

KG-based data fusion for predictive analytics

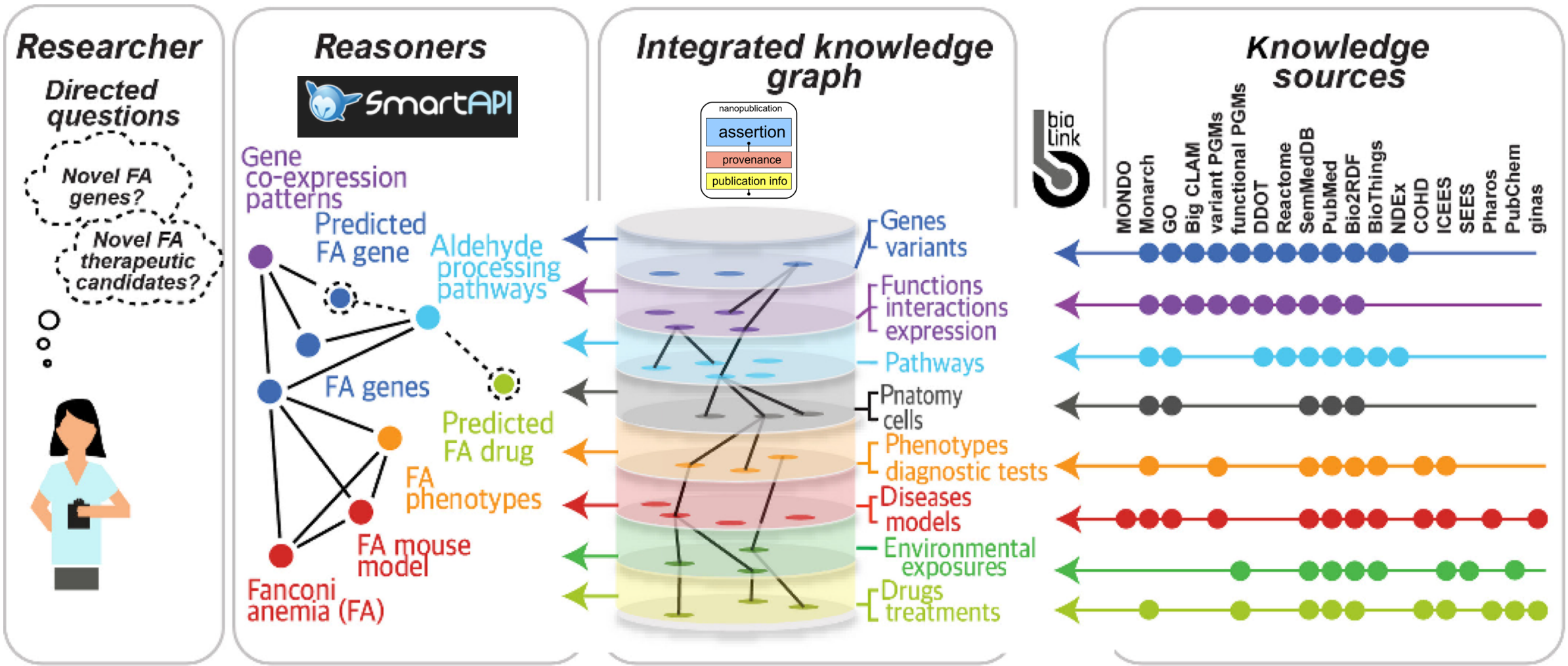


Journal of the American Medical Informatics Association, 29(3), 2022, 424–434

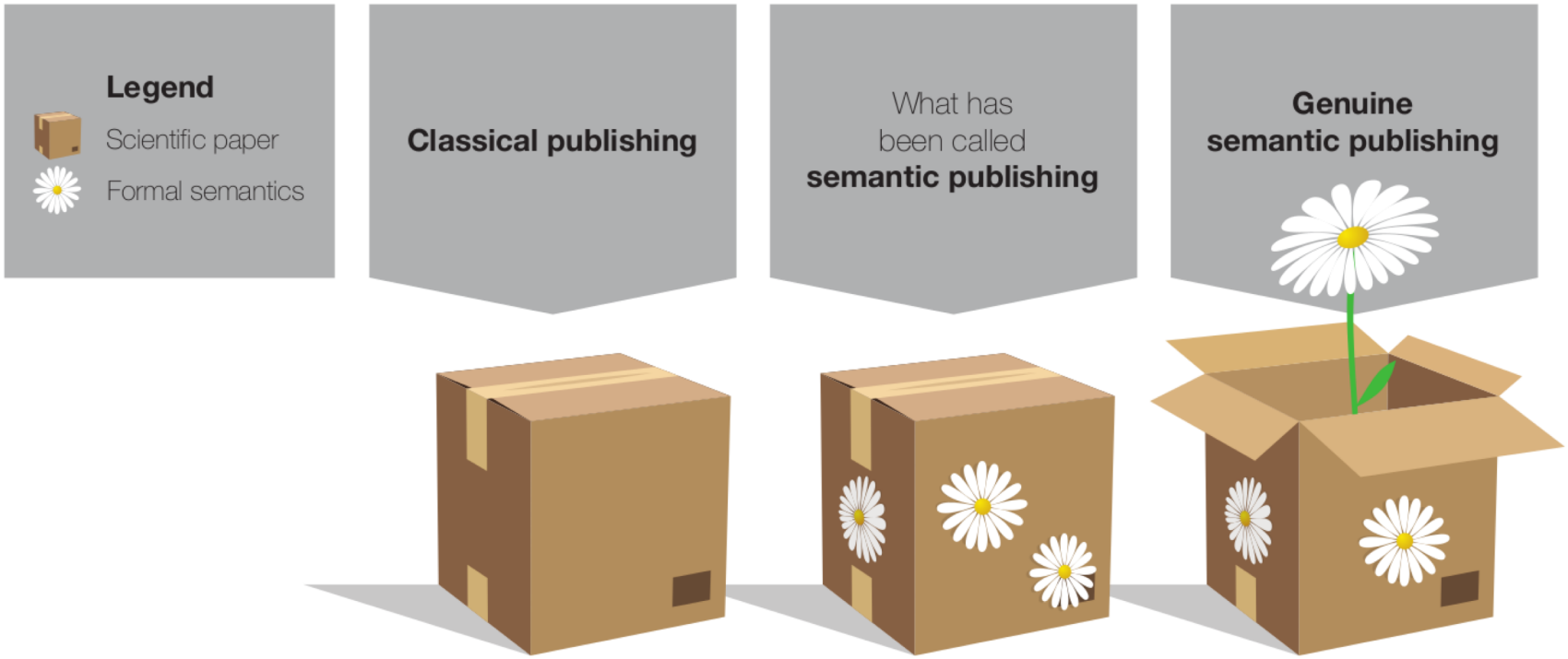
https://doi.org/10.1093/jamia/ocab270

National Center for Advancing Translational Sciences

Biomedical Data Translator



Rethinking Publishing Data-Driven Research



Genuine Semantic Publishing

by Tobias Kuhn and Michel Dumontier

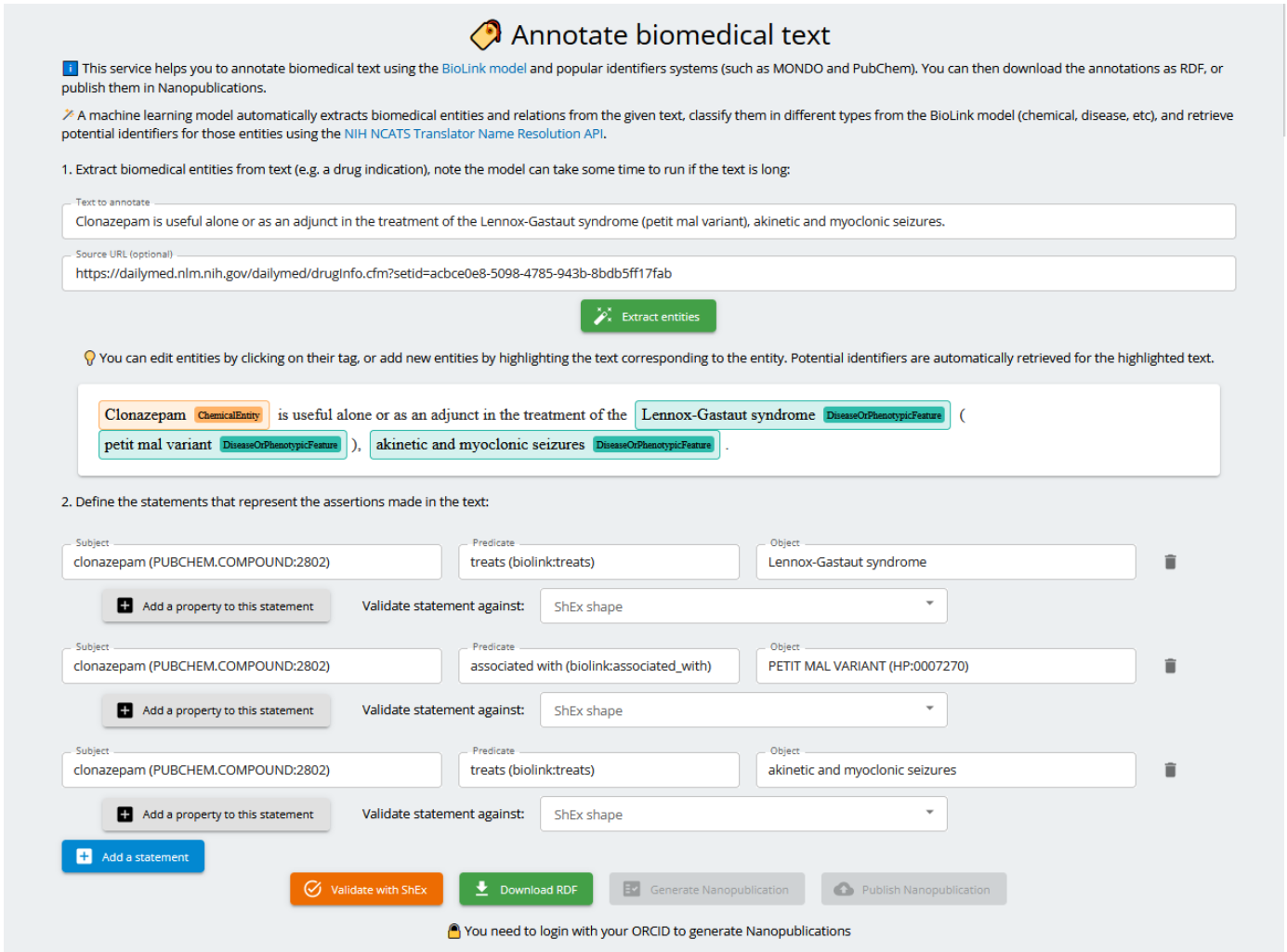
Content:

- 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/

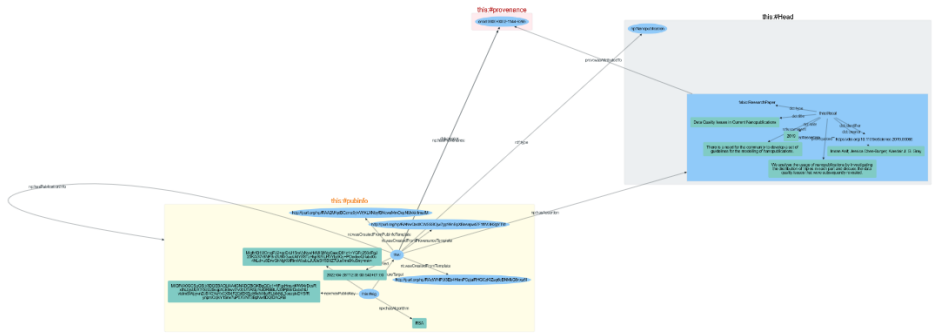
The Personal Scientific Knowledge Graph and Knowledge Collaboratory

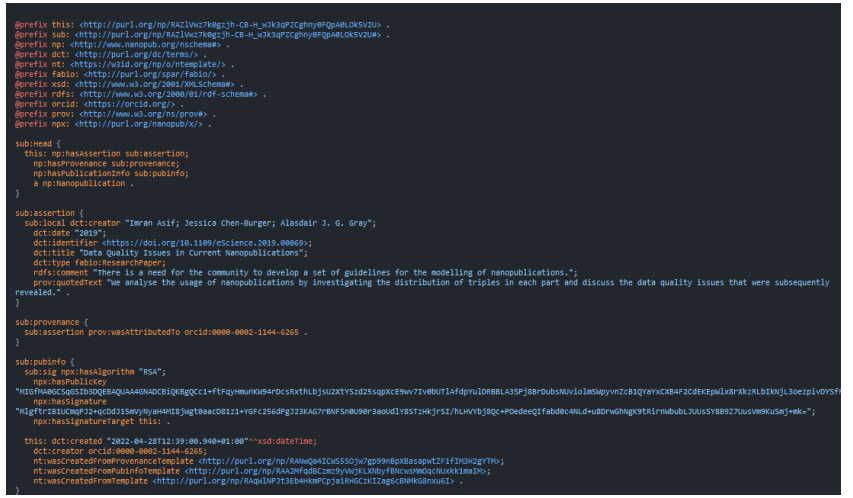
https://collaboratory.semanticscience.org



https://purl.org/np/RAZlVwz7k0gzjh-CB-H_wJk3qPZCghny0FQpA0LOk5V2U

Published on the 2022-04-28T11:39:00.940Z by Imran Asif



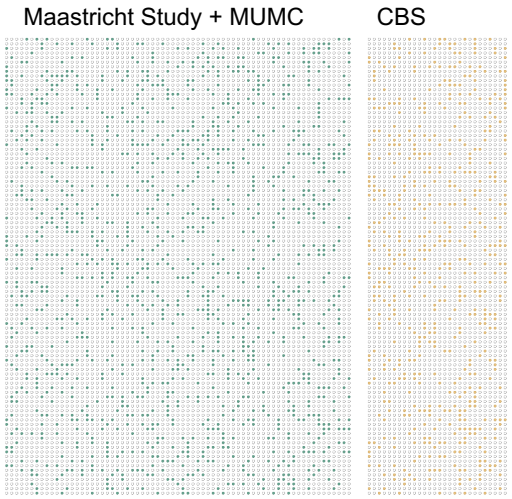


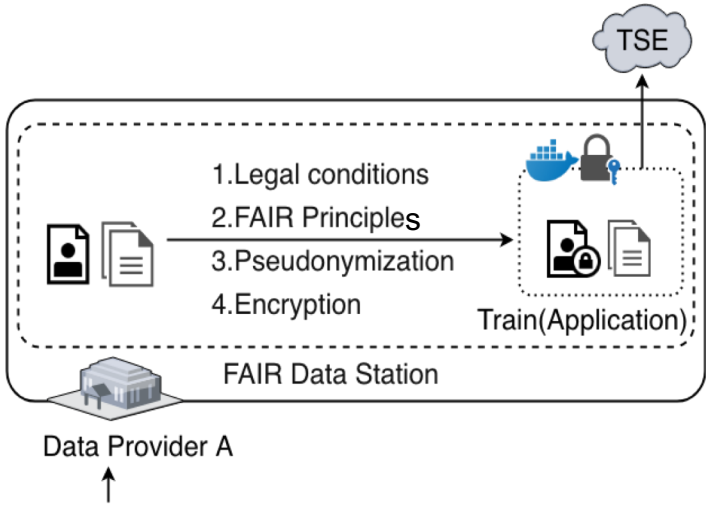
Data Science 5 (2022) 53–56
DOI 10.3233/DS-210047
IOS Press

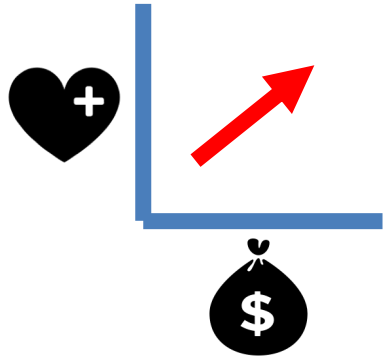
A formalization of one of the main claims of "The FAIR Guiding Principles for scientific data management and stewardship" by Wilkinson et al. 2016[1]

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

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

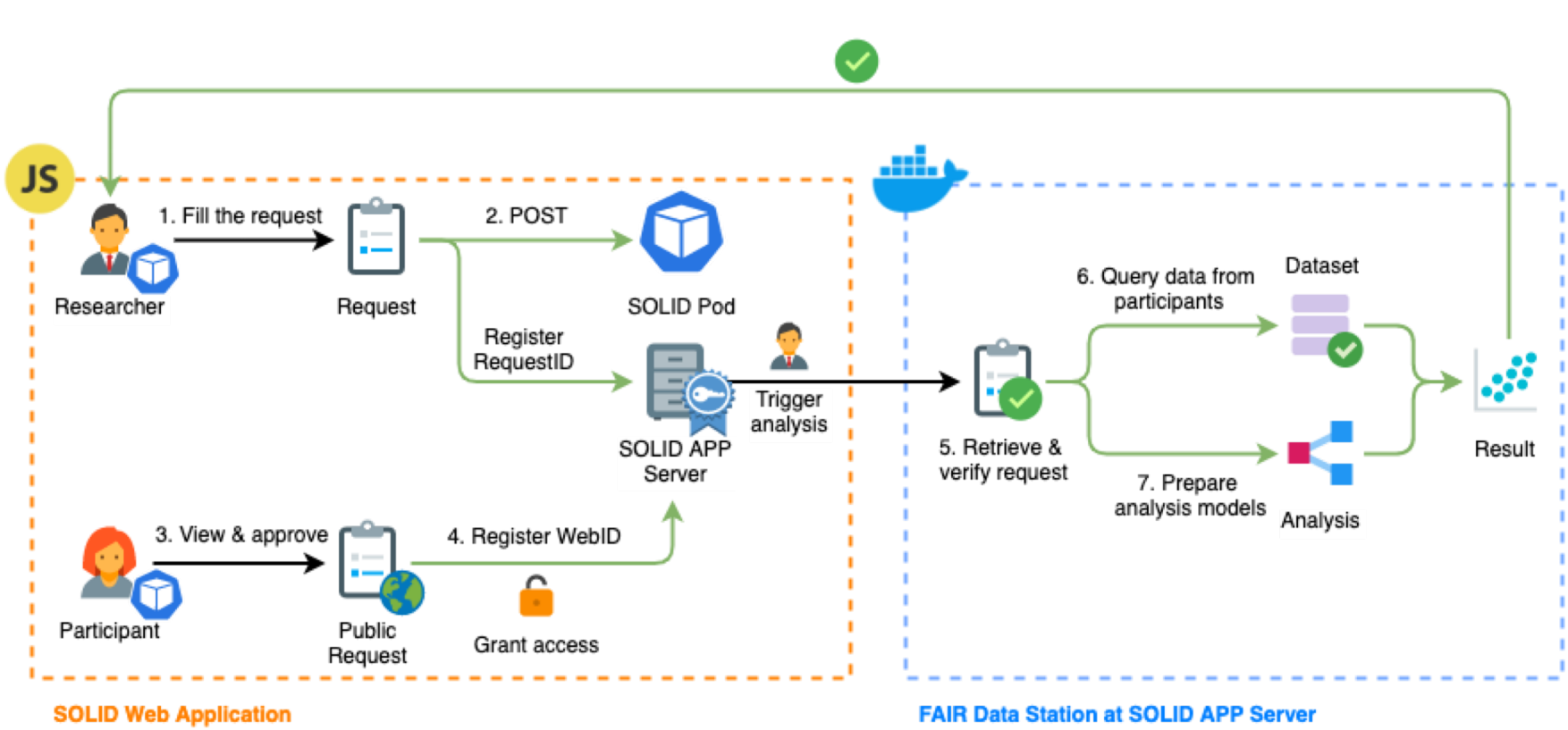


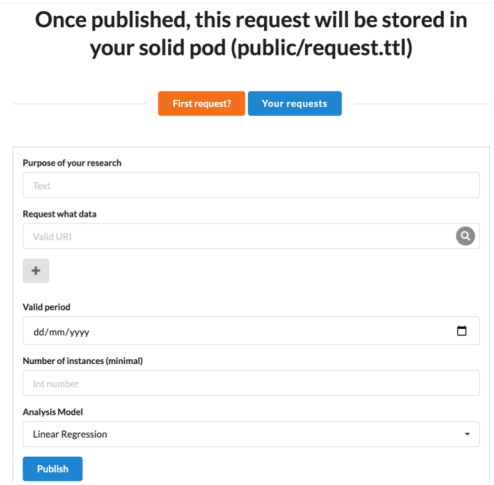


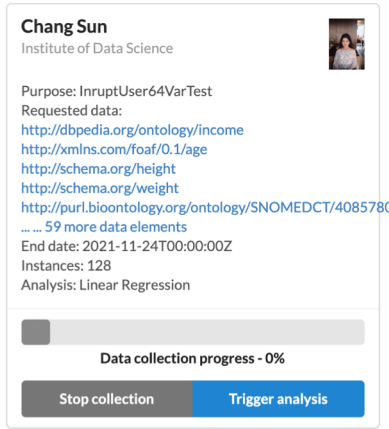


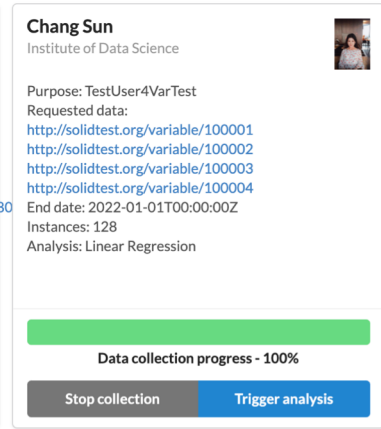
Learn **high confidence** determinants of health in a **privacy preserving** manner over ***vertically partitioned data***. The data are made available through **FAIR data stations** that provide access to *allowable* subsets of data to *authorized* users of *approved* algorithms.

TIDAL: Citizen-centric data platform using SOLID *using personal data for research*









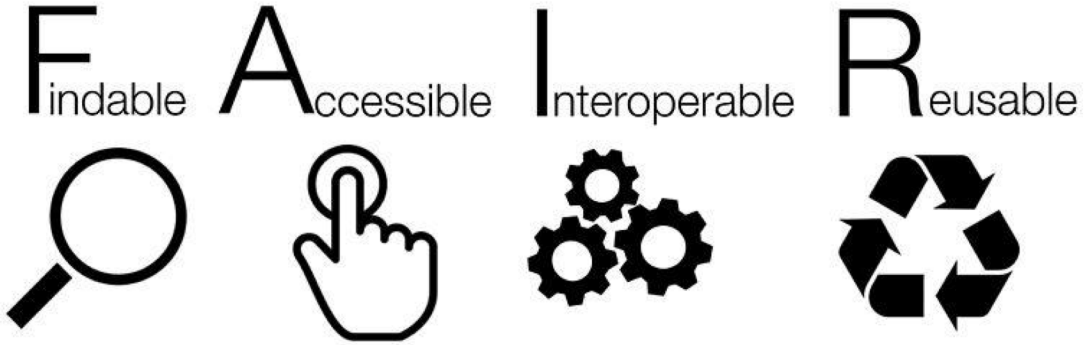
Summary

FAIR 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 to *facilitate* downstream reuse

We need to foster the construction of FAIR knowledge graphs from the source, and in a manner that supports purpose-driven interrogation in a privacy-preserving manner, where needed.

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