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
The world is awash with vast amounts of data
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 magnitude of the effect
2. 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.
Data Sharing and Management Snafu in 3 Short Acts

https://www.youtube.com/watch?v=N2zK3sAtr-4
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%
Our ability to reproduce landmark studies is surprisingly low:

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

THE CLINICAL-TRIAL CLIFF

Drug 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*

Cumulative success rate Phase I to launch
Percentage likelihood of moving from Phase I to launch

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

Nature Reviews | Drug Discovery
What hope do we really have to realize

Precision Medicine?
It’s time to completely rethink how we perform research
**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 reproduce original work, resulting in an ineffective translation of research into useful applications
It’s time to completely rethink how we perform research (and how we document and report it)
Rethinking Publishing Scientific 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/
De-centralized knowledge graphs

We need a new *social contract*, supported by *legal* and *technological* infrastructure to make digital resources available in a responsible manner.
Human Machine collaboration will be crucial to our future success
Findable Accessible Interoperable Reusable
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

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

Affiliations | 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

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 P2 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 formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles.
I3. (meta)data include qualified references to other (meta)data.

To be Reusable:
R1. (meta)data are richly described with a plurality of accurate and relevant attributes.
R1.1. (meta)data are released with a clear and accessible data usage license.
R1.2. (meta)data are associated with detailed provenance.
R1.3. (meta)data meet domain-relevant community standards.

http://www.nature.com/articles/sdata201618
Box 2 | The FAIR Guiding Principles

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

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

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

To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

http://www.nature.com/articles/sdata201618
In FAIR, attention to data and their metadata are equally important

**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
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 distinguish and retrieve all forms of digital content and knowledge
- high quality meta(data) to enhance discovery of relevant digital resources
- use of common vocabularies to facilitate query and statistical analysis
- establishment of community standards to reduce the effort in data reuse
- detailed provenance to provide adequate context and to enable reproducibility
- registered in appropriate repositories to fulfill a promise to future content seekers
- simpler terms of use to clarify expectations and intensify innovation
- social and technological commitments to make data ready for intelligent applications
Make scientific data FAIR

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

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). The principles state that research data should be ‘findable, accessible, interoperable and reusable’ (FAIR)². The idea is not new, but aligning this broad community around common data guidelines is a radical step.
Technical infrastructure (generic operations)
Social agreements/contracts (domain-specific content)

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
Drug Discovery Today
Volume 24, Issue 4, April 2019, Pages 933-938

Feature

Implementation and relevance of FAIR data principles in biopharmaceutical R&D

John Wise 1, 2, Alexandra Grebe de Barron 2, Andrea Splendiani 3, Beeta Balali-Mood 4, Drashti Vasant 2, Eric Little 4, Gaspare Mellino 5, Ian Harrow 1, Ian Smith 6, Jan Taubert 7, Kees van Bochove 8, Martin Romacker 5, Peter Walgemoed 9, Rafael C. Jimenez 10, Rainer Winnenburg 11, Tom Plasterer 12, Vibhor Gupta 13, Victoria Hedley 14

Show more

https://doi.org/10.1016/j.drudis.2019.01.008

Under a Creative Commons license

Get rights and content

open access
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.
Why Should *you* Go FAIR?

• Makes it easier for **to use your own data for a new purpose**
• Makes it easier for **other people to find, use and cite your data**, and for them to understand what you expect in return
• Makes it easier/possible for people to **verify your work**
• Ensure that the **data are available in the future**, especially as your may not want the responsibility
• **Satisfy the expectations** around data management from institution, funding agency, journal, my peers
https://www.go-fair.org/
Let’s build the Internet of FAIR data and services
FAIRification process

1. Retrieve non-FAIR data
2. Analyse the retrieved data
3. Define the semantic model
4. Make data linkable
5. Assign license
6. Define metadata for the dataset
7. Deploy FAIR data resource

Questions across multiple sources
Combine with other FAIR data

GO FAIR Fairification: https://www.go-fair.org/fair-principles/fairification-process/
FAIRplus FAIR cookbook: https://fairplus.github.io/cookbook-dev/intro.html
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

3954 Accesses  7 Citations  62 Altmetric  Metrics

http://w3id.org/AmIFAIR

Other schemes: https://fairassist.org
The Linked Open Data Cloud

https://lod-cloud.net/
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)

- **Uniform Resource Identifiers** (URIs) should be used to name and identify individual things.
- **HTTP** URIs should be used to allow these things to be looked up, interpreted, and subsequently "dereferenced".
- Useful information about what a name identifies should be provided through open standards such as **RDF**, **SPARQL**, etc.
- When publishing data on the Web, other things should be referred to using their HTTP URI-based names.
Key Differences between FAIR and 5* Linked Open Data Principles

1. FAIR: Independent of any particular implementation
2. FAIR: Equal emphasis on data and metadata.
3. FAIR: Open where possible, closed where necessary.
4. FAIR: Specific inclusion of licensing and provenance info, for both data and their metadata
Bio2RDF is an open source project that uses semantic web technologies to make it easier to reuse biomedical data.

- **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
Query multiple databases on the biological web of data

Phenotypes of knock-out mouse models for the targets of a selected drug (Imatinib)
An **ontology** for formal knowledge representation.

- Features **essential types and relations** for the rich description of arbitrary (real, hypothesized, virtual, fictional) objects, processes and their attributes.
- Provides **extensions** in the domains of chemistry, biology, biochemistry, and bioinformatics.
- Defines **simple design patterns** to describe and associate qualities, capabilities, functions, quantities, and informational entities including textual, geometrical, and mathematical entities.
- 1552 classes, 211 object properties, and 1 datatype property.
6.2.2 Title

At least one human-readable title should be provided for a dataset using dct:title. Alternative or older titles may be specified using dct:alternative.

For example, to provide a title and alternative title for the ChEMBL dataset:

```xml
<:chembl
  dct:title "ChEMBL"@en;
  dct:alternative "ChEMBL-db"@en;
```

---

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

https://rdfshape.weso.es/
Validate RDF data with ShEx

**RDF Input**

- by Input
- By URL
- By File

Data format
- turtle

Inference
- NONE

Add endpoint

**Shapes graph (ShEx)**

- by Input
- By URL
- By File

Data format
- ShExC

**ShapeMap**

- by Input
- By URL
- By File

Data format
- compact

Validate

---

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

https://rdfshape.weso.es/
Explore we know, and formulate hypotheses about what we don’t
by exploring a probabilistic semantic knowledge graph

And validate them against pipelines for drug discovery

Finding melanoma drugs through a probabilistic knowledge graph.
Reproduce original research

AUC 0.91 across all therapeutic indications

Scripts not available. Feature tables available.

Result: ROCAUC 0.83 ... doesn’t quite match

https://doi.org/10.7717/peerj-cs.281
Examine disease pathophysiology and treatment
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.
FAIR is a part of the solution that will enable arbitrary machines to work with each other.
Summary

Biomedicine is a knowledge driven discipline – it is fueled by access and reuse of data and research findings. We need these to be readily available in a manner accessible to the machines we use in our research.

FAIR demands a new social, legal, ethical, scientific and technological infrastructure:

– How does your research group or community make their data/findings FAIR?
– What support does your university provide you? FAIR data stewards, courses.
– Are you making use of all the data and findings that you could?

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.
Acknowledgements

Dumontier Lab (Maastricht University, Stanford University, Carleton University)
SU: Sandeep Ayyar, Remzi Celebi, Shima Dastgheib, Maulik Kamdar, David Odgers, Maryam Panahiazar, Amrapali Zaveri
CU: Alison Callahan, Jose Toledo-Cruz, Natalia Villaneuva-Rosales

FAIR

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