The FAIR Principles

Guidelines for publishing reusable data

Mark D Wilkinson

mark.wilkinson@upm.es

2020-10-27

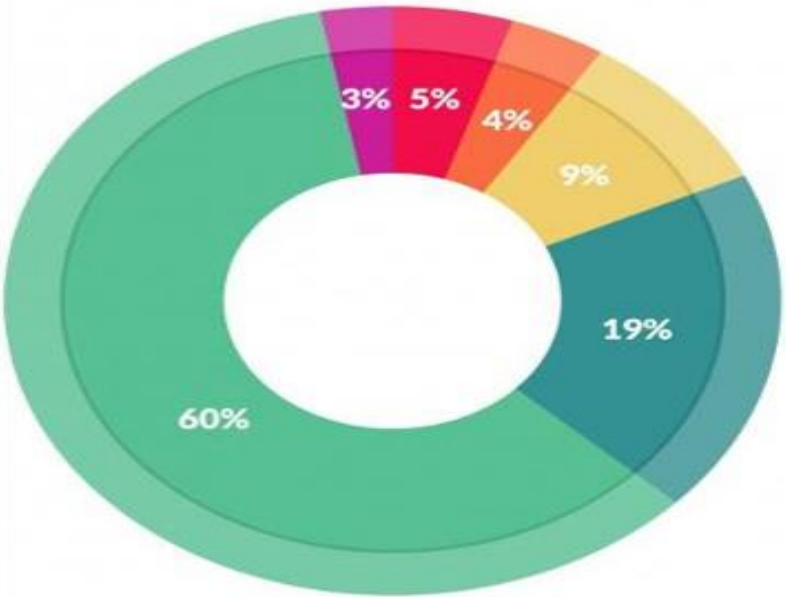
FAIR Data Systems

mea culpa

Most of my examples will come from academia

Most will come from the biological sciences

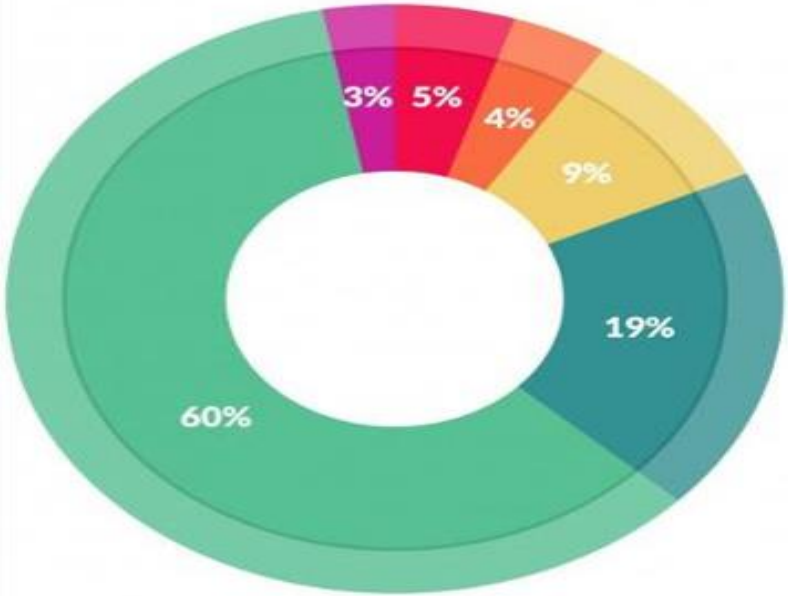
FAIR Data Systems



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%*

http://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says



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%

People who work primarily with data

80% of their time

Finding, filtering, reformatting, and integrating data

FAIR Data Systems

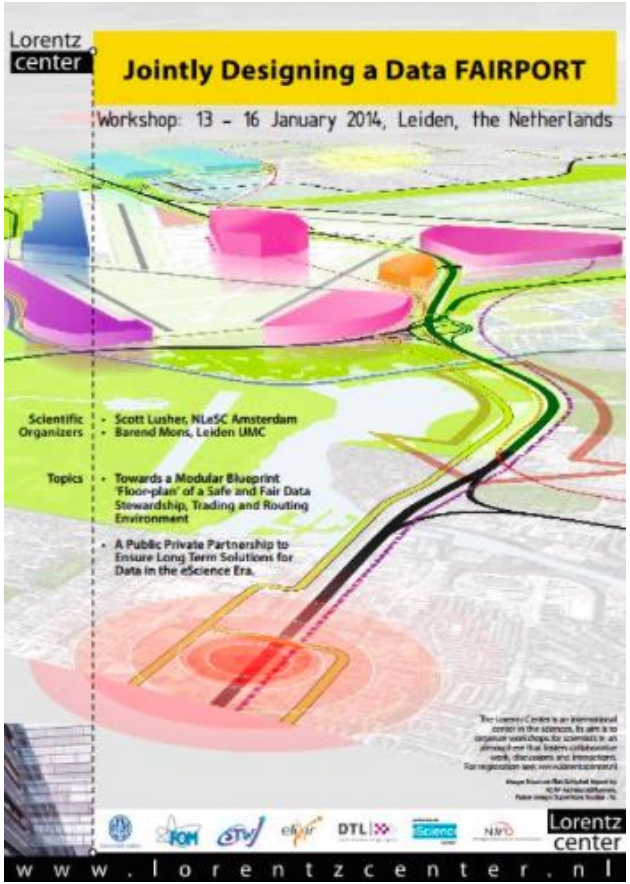
Most data is...

RE-USELESS

A brief history of FAIR

FAIR Data Systems

January, 2014 “unconference”
Jointly Designing a Data FAIRPORT

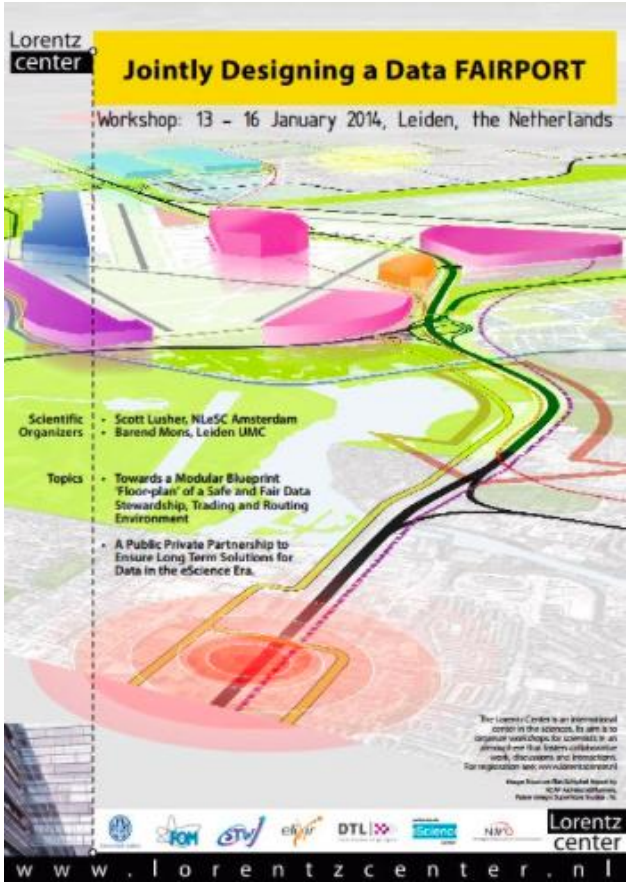


FAIR Data Systems

January, 2014 “unconference”
Jointly Designing a Data FAIRPORT

~31 attendees representing:

- Leading research infrastructures
- Policy institutes
- Publishers
- Semantic web specialists
- Innovators
- Computer scientists
- Experimental (e)Scientists

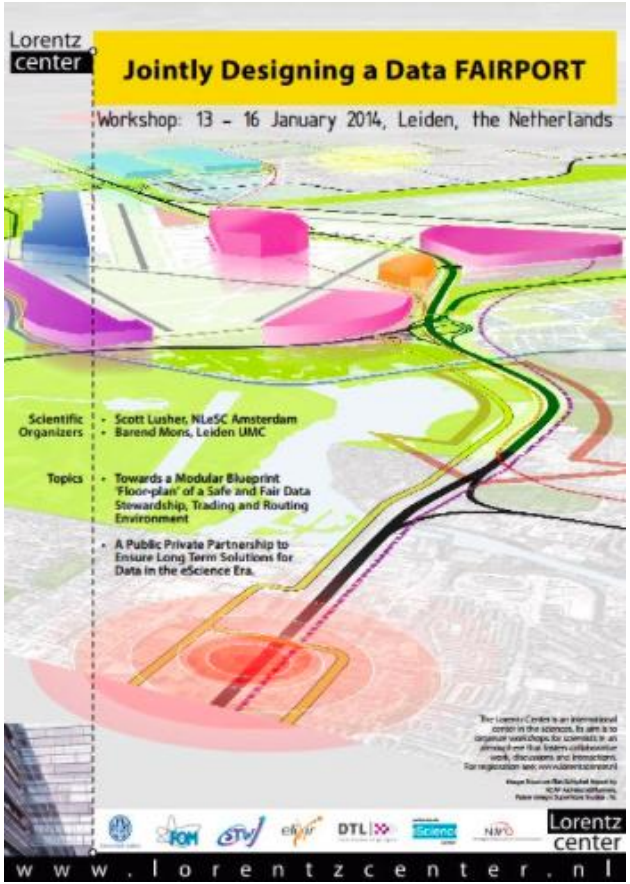


FAIR Data Systems

January, 2014 “unconference”
Jointly Designing a Data FAIRPORT

~31 attendees representing:

- Leading research infrastructures
- Policy institutes
- Publishers
- Semantic web specialists
- Innovators
- Computer scientists
- Experimental (e)Scientists



All participants agreed that a global infrastructure for **professional** data publishing, discovery, exchange and re-use is essential for effective data driven research.

FAIR Data Systems

January, 2014 “unconference”
Jointly Designing a Data FAIRPORT

Initial list of 23 DATA FAIRPORT detailed requirements

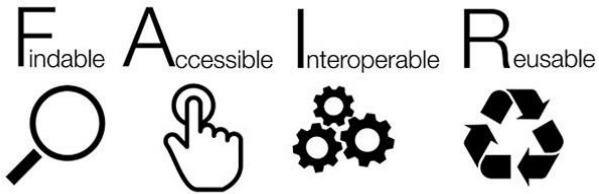
1. Ensure meta data is captured to standards
2. Have citable publication of the data
3. Enable to show that data is used
4. Allow for bootstrapping start functionality:
5. My data is in container X including a bar code on content
6. It can be found and recognised and, if needed, transported
7. Any functionality from user demand can be stacked and provided by a specialist providing systematic data handling compliance
8. Awareness of what is available datawise
9. Identifying relevant data (sets) and Providing information on data (sets); catalog of datasets
10. Providing access to datasets with security and levels of access controls11.Conversion of data (sets) into interoperable format
12. Speeding up data analyses processes
13. Define value returned to the submitter as well as to the community
14.Could take form of certain kinds of similarity search (“you may be interested in these other similar datasets”), statistics, & other computations.
15. Also: citability, funding body impact scores, publishability.
16. Reciprocity of use…
17. Core “bibliographic type” metadata: think of “PubMed for data
18. Cross reference to WHERE the data resides
19. Default data storage where there is not specialist site to store
20. Citable persistent HDL/DOI…
21. Domain--specific metadata plugins e.g. MIAME for arrays, etc
22. Availability metadata –e.g. public/restricted?
23. Optional (but important for our use case) Dataset--specific metada

Through consideration of these core requirements

The acronym FAIR was subsequently coined by **Dr. Barend Mons**

Breaking these requirements into the categories of:

Findability
Accessibility
Interoperability
Reusability



FINDABLE

→ Unambiguous identifiers supported by searchable metadata

ACCESSIBLE

→ Clearly-defined access protocol, preferably machine-actionable

INTEROPERABLE

→ Use shared vocabularies/ontologies in machine-accessible format

REUSABLE

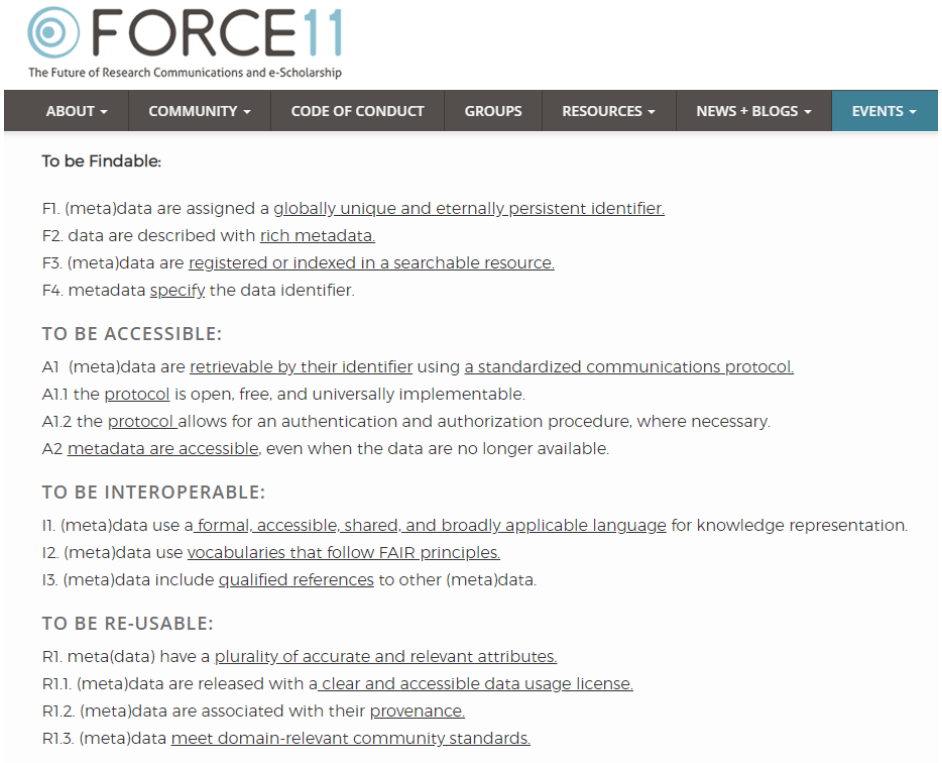
→ Contextual information, allowing proper interpretation

→ Rich provenance information facilitating accurate citation

FAIR Data Systems

The “first edition” of the FAIR Principles was then published for public comment

The venue chosen was the FORCE11 website
(FORCE11 is a community leader supporting research communication)



FAIR Data Systems

In parallel a working group assembled during the NBDC/DBCLS BioHackathon 2015 to further refine/edit the Principles

Editors:

Michel Dumontier

Mark Wilkinson

Contributors:

Mark Thompson

Nick Juty

Gang Fu

Jerven Bolleman

FAIR Data Systems

SCIENTIFIC DATA 2016

Comment | OPEN

The FAIR Guiding Principles for scientific data management and stewardship

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 - Show fewer authors

FAIR Data Systems

nature.com > scientific data > comment > article

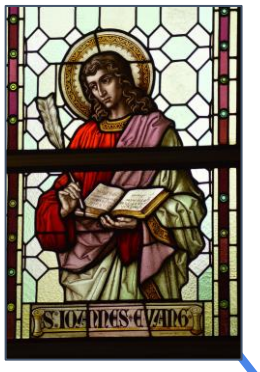
SCIENTIFIC DATA 2016

Created by Royyan Wijaya from Noun Project

Comment | OPEN

The FAIR Guiding Principles for scientific data management and stewardship

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 - Show fewer authors



FAIR Data Systems

Things happened very quickly!

https://ec.europa.eu/research/openscience/pdf/realising_the_european_open_science_cloud_2016.pdf

"Make adequate data stewardship mandatory for all research proposals.

...Horizon 2020, should only support projects that properly address Data Stewardship [and those] **that do not specify FAIR conditions...**

should not be eligible for funding."

FAIR Data Systems

https://ec.europa.eu/research/press/2016/pdf/opendata-infographic_072016.pdf

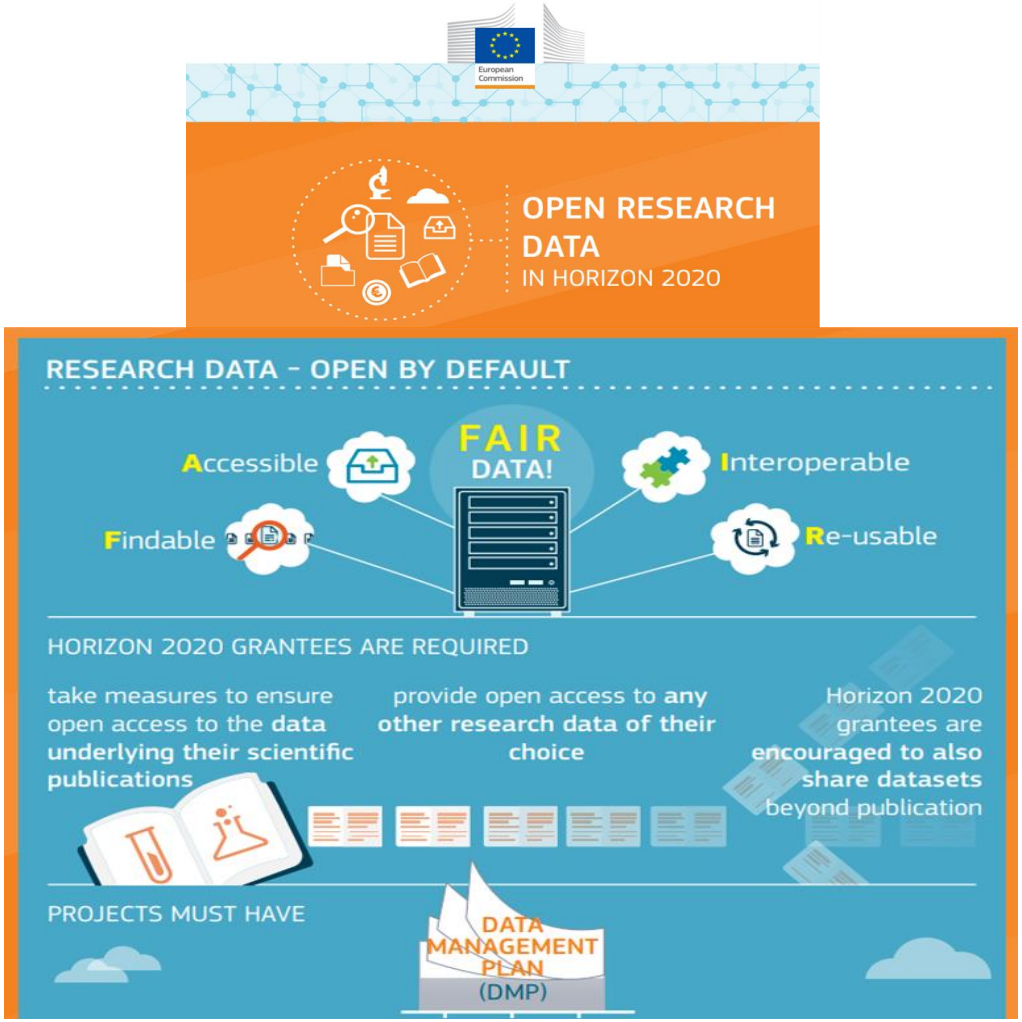
EUROPEAN COMMISSION
Directorate-General for Research & Innovation

H2020 Programme

Guidelines on

FAIR Data Management in Horizon 2020

Version 3.0
26 July 2016



FAIR Data Systems

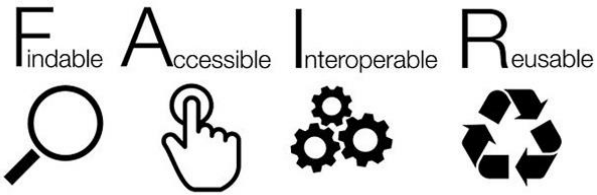
European Commission - Statement

G20 Leaders' Communique Hangzhou Summit

Hangzhou, 5 September 2016

"...we support appropriate efforts to promote open science and facilitate appropriate access to publicly funded research results on findable, accessible, interoperable and reusable (FAIR) principles."

FAIR Data Systems



FINDABLE

→ Unambiguous identifiers supported by searchable metadata

ACCESSIBLE

→ Clearly-defined access protocol, preferably machine-actionable

INTEROPERABLE

→ Use shared vocabularies/ontologies in machine-accessible format

REUSABLE

→ Contextual information, allowing proper interpretation

→ Rich provenance information facilitating accurate citation

FAIR Data Systems

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 Principles in detail

Stopping to smell some roses
along the way...

FAIR Data Systems

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 **explicitly include the identifier of the data** it describes

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

...explicitly include the identifier of the data...

Purpose #1:

- Obviously, it is important to know, unambiguously, what the metadata is describing!

FASTA Formatted Sequence Data

http://doi.org/10.555/fastadatasets.pb11235

>gi|186683441|ref|YP_001862366.2| serine threonine kinase

DLPFFGADLVTLPGGHLIALDMQPLFRDDSAYQAKYTEPILPIFHAHQ

MNSERSDVTLYQPFLDYAIAYMRSRLDLEPYPIPTGFESNSAVVGKGK

AHVQGGNSLQVLNFVIFPHLNYQHLSWGNQEEVVTTSYAFQTAKLRQIRA

GDFPEEAQPFFSPAFLWTRPQETAVVETQVFAAFKDYLKAYLDFVEQAEAV

LRYLRYRAEKDPARGMFKRFYGAEWTEEYIHGFLFDLERKLTVVKTDSQNLVAIKQAQ

>gi|186681228|ref|YP_001864424.1| phycoerythrobilin:ferredoxin oxidoreductase

MNSERSDVTLYQPFLDYAIAYMRSRLDLEPYPIPTGFESNSAVVGKGKNQEEVVTTSYAFQTAKLRQIRA

AHVQGGNSLQVLNFVIFPHLNYDLPFFGADLVTLPGGHLIALDMQPLFRDDSAYQAKYTEPILPIFHAHQ

QHLSWGGDFPEEAQPFFSPAFLWTRPQETAVVETQVFAAFKDYLKAYLDFVEQAEAVTDSQNLVAIKQAQ

LRYLRYRAEKDPARGMFKRFYGAEWTEEYIHGFLFDLERKLTVVK

FAIR Data Systems

…explicitly include the identifier of the data…

Purpose #1:

- Obviously, it is important to know, unambiguously, what the metadata is describing!

Purpose #2:

- Many data formats do not have a “place” to put “arbitrary” information, such as a pointer to the metadata describing that data
- Therefore, the only way to discover the metadata is to search using the data identifier of the

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

retrievable…open, free, and universally implementable protocol

Key: “retrievable” not “resolvable”!!

- The word “retrievable” was chosen carefully!

Key: “protocol” does not imply mechanized protocol!

- Protocol: a set of rules governing the exchange or transmission of data.

Purpose: to support the FAIR retrieval of highly-sensitive data

How???

FAIR Data Systems

This is a FAIR Protocol:

Write down the ID number of the data of interest

Call 555-2368

Ask for Mark

Request a copy of the identified data

Answer some authentication questions

The data will arrive on an encrypted USB key

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

“Qualified references”

Different types of tests for COVID-19

There are many different technologies for COVID-19 testing, some currently available and some still in development. Trackers of the development, regulatory status and commercial release of different types of COVID-19 test are being compiled by Johns Hopkins University and the medical industry news website, 360Dx.

“Qualified references”

Different types of tests for COVID-19

There are many different technologies for COVID-19 testing, some currently available and some still in development. Trackers of the development, regulatory status and commercial release of different types of COVID-19 test are being compiled by Johns Hopkins University and the medical industry news website, 360Dx.

What happens when I click on this blue text?

FAIR Data Systems

“Qualified references”

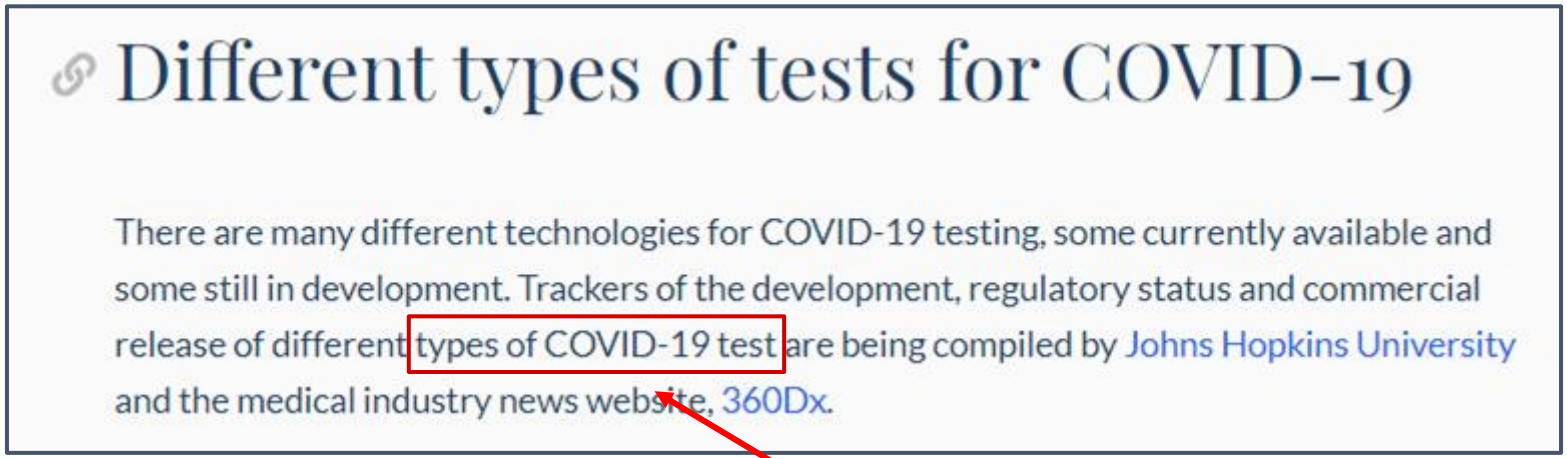
Different types of tests for COVID-19

There are many different technologies for COVID-19 testing, some currently available and some still in development. Trackers of the development, regulatory status and commercial release of different types of COVID-19 test are being compiled by Johns Hopkins University and the medical industry news website, 360Dx.

Do I go to some page about Johns Hopkins?
(which page?)

FAIR Data Systems

“Qualified references”



The context is “types of COVID test… so maybe a page about the test?

FAIR Data Systems

“Qualified references”

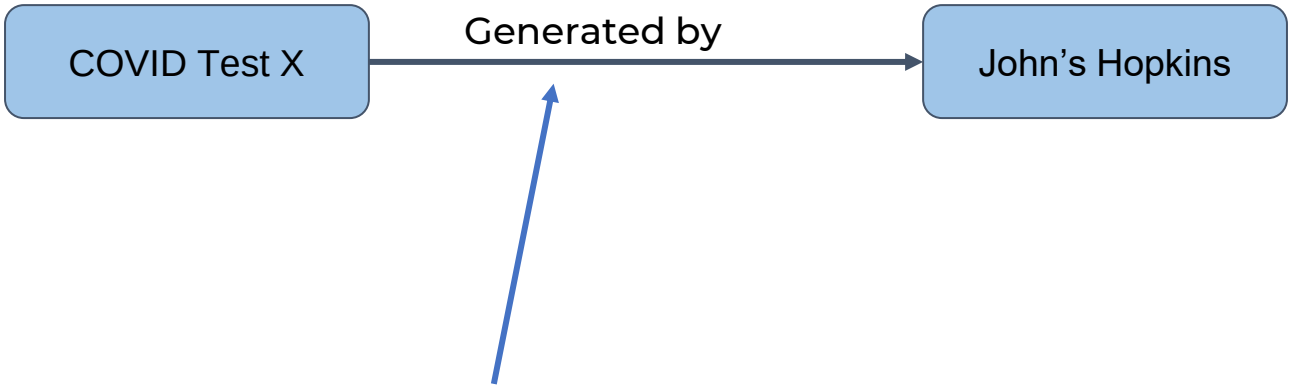
Different types of tests for COVID-19

There are many different technologies for COVID-19 testing, some currently available and some still in development. Trackers of the development, regulatory status and commercial release of different types of COVID-19 test are being compiled by Johns Hopkins University and the medical industry news website, 360Dx.

Answer: You have absolutely no idea!

Because the links in the Web are not “explained”, The Web only functions based on human intuition… and I just demonstrated that we cannot even rely on that!

“Qualified references”



FAIR Data Systems

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 Testing and Evaluation

Measuring FAIRness

FAIR Data Systems

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

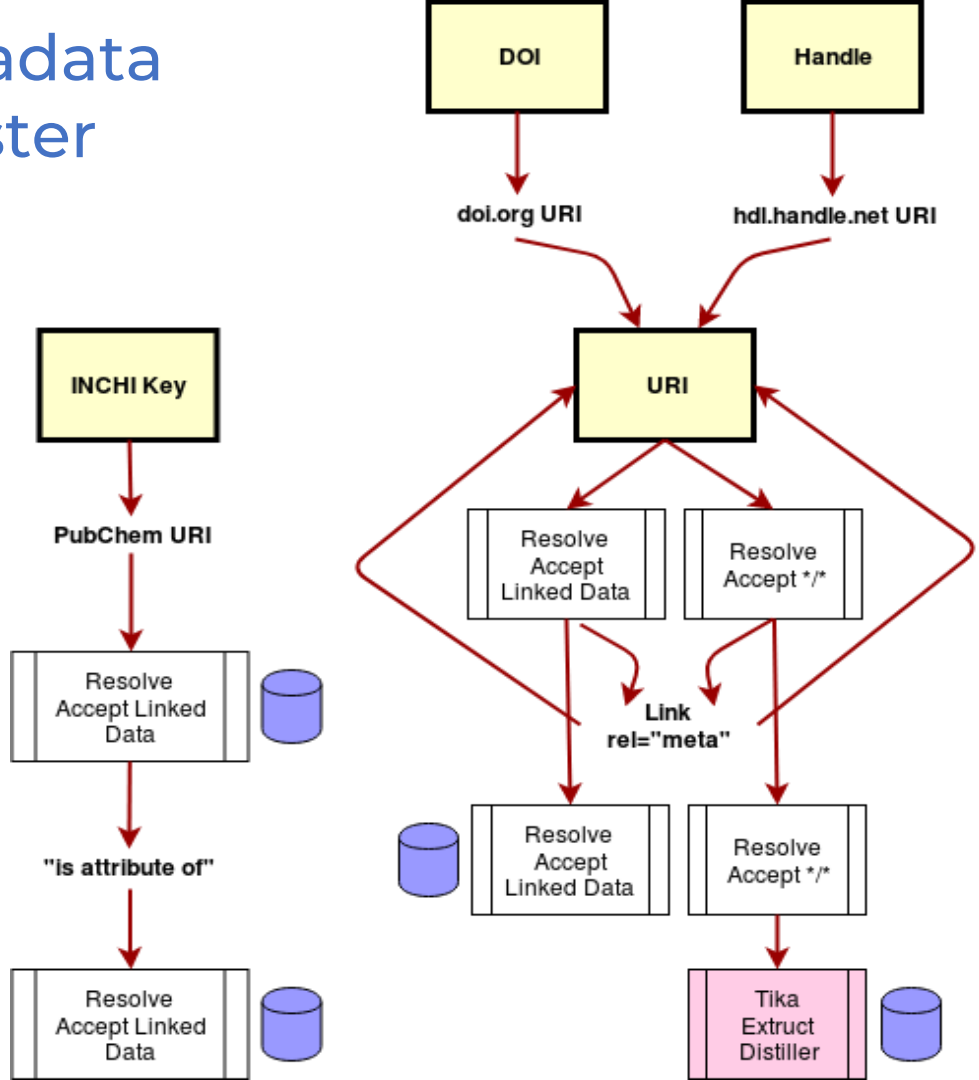
FAIR is as much (or more!) about Metadata as it is about Data...

Testing MI Compliance

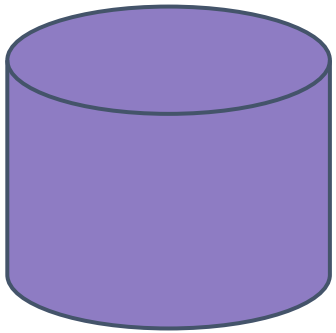
Rule #1: Metadata must exist!

FAIR Data Systems

The Metadata Harvester



FAIR Data Systems



This “blob” contains a mashup of:

- All hash-style structured metadata
- All linked-data style structured metadata

22 Tests have been written that examine this metadata for behaviours expected from Principles F, A, I (and to a limited extent, R)

FAIR Data Systems

The FAIR Evaluator

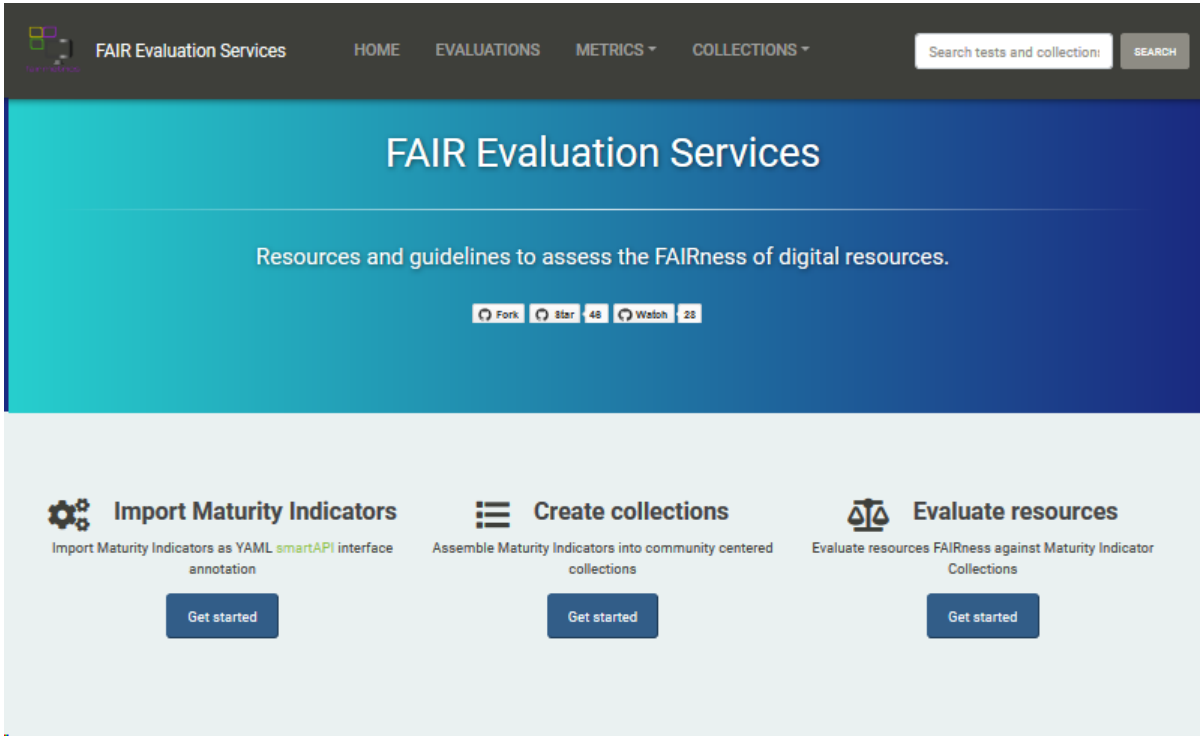
Automating bundles of tests

https://tinyurl.com/FAIREvaluatorPaper

FAIR Data Systems

The Evaluator Online

https://w3id.org/AmIFAIR



FAIR Data Systems

“Collection”-centric Evaluations

- Communities decide which Tests are relevant to them
- These are registered in the Evaluator as a “Collection”
  - Documentation about what Tests are included, and to what communities the Collection would be relevant
- Anyone can execute an evaluation on any Identifier
- Anyone can select any Collection they wish to apply
  - For example, journals may select different evaluation collections than funding agencies, or researchers
- An “Evaluation”, therefore, is the application of an identified collection of Tests tests to a given resource of interest.

FAIR Evaluation Services

Resources and guidelines to assess the FAIRness of digital resources.

Magnaporthe polyadenylation site database homepage against all MIs

Summary:

Description: FAIR Metrics Evaluation: Magnaporthe polyadenylation site database homepage against all MIs; Tested identifier: http://linkeddata.systems/Magnaporthe/polyA_Sites/; generated by https://orcid.org/0000-0001-6960-357X
Resource: http://linkeddata.systems/Magnaporthe/polyA_Sites/
Collection: 6
Observations: Ran 22 tests (19 succeeded, 3 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 Data Systems

Extensive provenance trail for how test result was determined

Helps you understand how to improve!

FAIR METRICS GEN2- METADATA IDENTIFIER EXPLICITLY IN METADATA

Status: Success
Principle tested: F3
Description: Metric to test if the metadata contains the unique identifier to the metadata itself. This is done using a variety of 'scraping' tools, including DOI metadata resolution, the use of the 'extruct' Python tool, and others...
Metric test created on: May 8, 2019 by Mark D Wilkinson (updated on May 8, 2019).
Test executed on: May 21, 2019

Test results

```
INFO: Found a URI.
INFO:  Attempting to resolve http://linkeddata.systems/Magnaporthe/polyA_Sites/ 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 http://linkeddata.systems/Magnaporthe/polyA_Sites/
using HTTP Accept header {"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: parsing as HTML.
INFO:  Using 'extruct' to try to extract metadata from return value (message body) of http://linkeddata.systems
/Magnaporthe/polyA_Sites/.
INFO: the extruct tool found parseable data at http://linkeddata.systems/Magnaporthe/polyA_Sites/
INFO: The response message body component appears to contain JSON::LD::Format.
INFO: Using 'Kellog's Distiller' to try to extract metadata from return value (message body) of
http://linkeddata.systems/Magnaporthe/polyA_Sites/.
WARN: The Distiller tool failed to find parseable data at http://linkeddata.systems/Magnaporthe/polyA_Sites/.
INFO:  Attempting to resolve http://linkeddata.systems/Magnaporthe/polyA_Sites/ using HTTP Headers 'Accept:
*/*'.
INFO: Found html text/html type of content when resolving http://linkeddata.systems/Magnaporthe/polyA_Sites/
using HTTP Accept header {"Accept"=>"*/*"}.
INFO: parsing as HTML.
INFO:  Using 'extruct' to try to extract metadata from return value (message body) of http://linkeddata.systems
/Magnaporthe/polyA_Sites/.
INFO: the extruct tool found parseable data at http://linkeddata.systems/Magnaporthe/polyA_Sites/
INFO: The response message body component appears to contain JSON::LD::Format.
INFO: Using 'Kellog's Distiller' to try to extract metadata from return value (message body) of
http://linkeddata.systems/Magnaporthe/polyA_Sites/.
WARN: The Distiller tool failed to find parseable data at http://linkeddata.systems/Magnaporthe/polyA_Sites/.
INFO: Linked Data Found.  Now searching for the metadata identifier by simple string match in object position
SUCCESS: Found pattern-match in metadata _:b0 http://www.w3.org/2000/10/swap/pim/doc#persistencePolicy
http://linkeddata.systems/Magnaporthe/polyA_Sites/persistence.  This scores as a success, but additional tests
will now be executed.
```

FAIR Data Systems

The FAIR Principles are guidelines for publishing data that is easier to reuse

...especially for machines!

(but not ONLY for machines)

Software is available that allows you to automatically track your journey towards increasingly FAIR data infrastructures

FAIR Data Systems

Thank you for the invitation!

The FAIR Principles

Guidelines for publishing reusable data

Mark D Wilkinson

mark.wilkinson@upm.es

FAIR Data Systems