FAIR
Strategies and technologies for early wins

Mark D Wilkinson
Isaac Peral Senior Investigator
CBGP UPM-INIA/CSIC
Universidad Politécnica de Madrid

mark.wilkinson@upm.es

Keynote to #UNBigData2024
June 10, 2024

https://tinyurl.com/UNBigData2024

doi:10.5281/zenodo.11566326
Framing my presentation

FAIR is intended to help machines help people!

Replace time spent in data discovery/manipulation with time spent on thoughtful exploration of global knowledge

IT WORKS! We can now prove it!

We have also experienced some notable failures...
The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson, Michel Dumontier, JIisbrand 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
The FAIR Guiding Principles...

“\textbf{This necessitates machines to be capable of autonomously and appropriately acting when faced with the wide range of types, formats, and access-mechanisms/protocols that will be encountered during their self-guided exploration of the global data ecosystem.}”

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

When I wrote this paragraph, I was imagining a Web of data discovery and exploration \textbf{agents}.
FAIR is, first, a mechanism to guide automated agents to discovery of task-relevant data. As a consequence, FAIR is, before all else, about metadata.

Notable consequence #1

“This necessitates machines to be capable of autonomously and appropriately acting when faced with the wide range of types, formats, and access-mechanisms/protocols that will be encountered during their self-guided exploration of the global data ecosystem.”

https://www.nature.com/articles/sdata201618
Creating a Web of data that can be **appropriately** (re)used by machines necessitates specific data publishing behaviors by data providers.

These behaviors can be concretely described.

Software is then written to leverage these behaviors to find, access, and correctly reuse the data.

---

"This necessitates machines to be capable of autonomously and appropriately acting when faced with the wide range of types, formats, and access-mechanisms/protocols that will be encountered during their self-guided exploration of the global data ecosystem."
This means that FAIRness is, by definition, measurable by automated agents.
The FAIR Evaluator

The first fully-automated “agent” for testing FAIRness of a resource

Resulted from early pressure from e.g. journal editors who wanted to require FAIRness

https://doi.org/10.1038/s41597-019-0184-5
Any stakeholder provides the URL of any Digital Object, and ~22 distinct tests of “FAIRness” are executed on that Object.
The FAIR Evaluator Harvester

At its core, The Evaluator consists of a “very forgiving” metadata harvesting workflow & library.
FAIR Principle F3: metadata explicitly include the identifier of the data it describes

Explore the harvested metadata for any metadata facet that appears to be a reference to a data record

(there are at least 18 possibilities that are acceptable!)
Evaluator usage

~10,000 FAIRness evaluations run using the public version

Several thousand evaluations run using the private version from my company**

Executions of individual tests are not monitored, but do occur frequently
○ **22** independent FAIR assessment platforms

○ Most are questionnaire-based

○ All of them focus on (primarily) **metadata**

○ **But, there is a problem!** The scores from each platform are different…

---

**FAIR Assessment a cottage industry!**

---

https://fairassist.org
How different can they be?

Comparison of The Evaluator** with F-UJI, on the same Digital Object
(a Catalog record in the Duchenne Parent Project Patient Registry)

Test of: https://w3id.org/duchenne-fdp/catalog/c36b662c-fc4d-4b9f-a833-d4972a6fc395
Mon, 13 Sep 2021 11:08:19 +0000

F Metrics
A Metrics
I Metrics
R Metrics

Metrics release v1.0.26

20/22 Tests Pass

Summary:

Resource PID/URL: https://w3id.org/duchenne-fdp/catalog/c36b662c-fc4d-4b9f-a833-d4972a6fc395
Metric Versions: metrics,v0.4

Findable: 1 of 7
Accessible: 0 of 3
Interoperable: 1 of 4
Reusable: 0 of 10

2/24 Tests Pass
The problem of metadata discovery and interpretation

EOSC calls for an investigation of the FAIR assessment discrepancies
EOSC Task Force on FAIR Metrics and Data Quality

Co-Chairs:
- Mark D Wilkinson
- Chris Schubert
- Carlo Lacagnina (retired)

Established November 2021
Three key outputs v.v. FAIR Testing

**FAIR Assessment Tools: Towards an “Apples to Apples” Comparisons**

Authorship Community:
Mark D Wilkinson1,2, Susanna-Assunta Sansone2,4,5, Marjan Grootveld2,6, Josefine Nordling2,6, Richard Dennis2,7, David Hecker2,7 on behalf of the EOSC FAIR Metrics subgroup

**Report on “FAIR Signposting” and its uptake by the community**

Mark D Wilkinson1,2, Susanna-Assunta Sansone2,4, Marjan Grootveld2,6, Richard Dennis2,7, David Hecker2,7, Robert Huber8, Stian Soiland-Reyes9, Herbert Van de Sompel3, Andreas Czerniak10, Milo Thurston1, Allyson L. Lister1, Alban Gaignard11

**Community-driven Governance of FAIRness Assessment: An Open Issue, an Open Discussion**

Authorship Community:
Mark D. Wilkinson1,2, Susanna-Assunta Sansone2,4, Eva Méndez5, Romain David2,6, Richard Dennis2,7, David Hecker2,7, Mari Kleemola2,9, Carlo Lacagnina3,10, Anastasija Nikiforova2,11, Leyla Jael Castro12

https://doi.org/10.5281/zenodo.7463421

https://doi.org/10.5281/zenodo.7390482

https://doi.org/10.5281/zenodo.10490289
European Union’s Horizon Europe framework programme under grant agreement No. 101130187

Among its deliverables:

- Harmonization of FAIR Assessment tools, workflows, and APIs*
- Bootstrapping a global governance process for FAIR testing*

* These activities are open to non-project participants, so please contact me if you wish to become involved!
An example of a successful large-scale FAIRification initiative

The European Joint Programme on Rare Diseases (EJP-RD)
EJP RD in numbers

+1800 people
35 participating countries
26 EU MS, 7 associated (AM, CH, GE, IL, NO, RS, TK), UK and CA

ALL 24 ERNs

101 M€ Budget
Union contribution: 55 M€ (70% reimbursement rate)

85% of European RD research community (directly or indirectly) involved in EJP RD

91 beneficiaries
10 hospitals
12 research institutes
31 research funding bodies/ministries
27 universities/hospital universities
5 EU infrastructures
5 charities/foundations
EURORDIS

+ 52 linked third parties
+100% associated networks

Coordinated by
Inserm

EJP RD in numbers

Challenge: Make it all FAIR

…but how?
Step 1 - Focus on Metadata that answers the question: “What kind of data does this partner have”?

EJP-RD Metadata Schema:

- Based on DCAT; attempts to follow the European DCAT-AP
- Published by all sites via semi-automated process using Excel spreadsheets
EJP-RD Metadata Publication Platform: FDP

- Open-source, currently distributed via Docker images
- Installs in seconds
- Publishes metadata for human exploration
- Publishes metadata for agent-based exploration
- Provides easy-to-use DCAT editing tools via Web pages
- Automatically registers the new FDP in a central index
Result: a federated network of FDPs representing all biobanks and patient registries

<table>
<thead>
<tr>
<th>Endpoint</th>
<th>Registration</th>
<th>Modification</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="https://w3id.org/ctsr-fdp/">https://w3id.org/ctsr-fdp/</a></td>
<td>15-12-2023, 06:00:00</td>
<td>08-06-2024, 06:00:00</td>
<td>ACTIVE</td>
</tr>
<tr>
<td><a href="https://directory.bbMRI-eric.eu/api/fdp">https://directory.bbMRI-eric.eu/api/fdp</a></td>
<td>04-10-2023, 16:12:15</td>
<td>07-06-2024, 14:00:00</td>
<td>ACTIVE</td>
</tr>
<tr>
<td><a href="https://fdp.wikipathways.org/index.ttl">https://fdp.wikipathways.org/index.ttl</a></td>
<td>27-02-2024, 22:34:01</td>
<td>04-06-2024, 03:02:14</td>
<td>ACTIVE</td>
</tr>
</tbody>
</table>
Result: a federated network of FDPs representing all biobanks and patient registries

It is now possible for a computational agent to automatically explore the metadata of all participants to discover which ones potentially contain data of interest to a rare disease researcher or clinician.
Am I suggesting that there’s no point in making FAIR data?

No…

but there’s little point in working on FAIR data until you get the discovery metadata right!

Some data will never be made FAIR!
However…

EJP-RD also created FAIR Data via an end-user-friendly reusable FAIRification pipeline
Challenge

Need to make all data-focused network partner’s (~50) resources work together

Partners have similar data (patient registry or biobank samples)

Partners have different starting formats

**Generally, they are forbidden from sharing or moving their data**

So the participants are going to have to do the FAIR transformation themselves, on-site, sometimes without even letting the FAIR experts see it!
Step 2: Build a shared FAIR data model
Step 3: Use CSV as a “lingua franca” for all partners

CSV Generated by the participants (easy!)

<table>
<thead>
<tr>
<th>Property</th>
<th>Observation</th>
<th>Value</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diagnosis</td>
<td>ORPHA:98896</td>
<td>-</td>
<td>2021-02-01</td>
</tr>
<tr>
<td>Body Measurement</td>
<td>NCIT_C25208</td>
<td>28kg</td>
<td>2020-04-05</td>
</tr>
<tr>
<td>Laboratory Measurement</td>
<td>NCIT_C399</td>
<td>10mg/L</td>
<td>2020-04-05</td>
</tr>
</tbody>
</table>

Fully automated transformation and “publication” of FAIR data into their FAIR Data Point.
Does this work, in practice?

Five FAIR Data Points for partners representing NMDs
“What is the delay between symptom onset and diagnosis?”

1) Metadata allowed automated agent to discover and interact with participants capable of providing question-relevant data
2) The same query sent to all participants (shared model)
3) Integrate the output

Duchenne Parent Project (DPP)

EURO-NMD (mock data)
Caveat emptor!!

FAIR Data alone is NOT sufficient to achieve interoperability!
This paper compares two independent FAIRification efforts over identical data items.

Interoperability was almost zero!

*Shared models are necessary*
Segue into the panel session coming up next

FLAIR-GG

FAIRification, Linking And Integrated Reuse of Global ex situ plant Germplasm resources

Dr. Santiago Moreno Vasquez
Dr. Mark D. Wilkinson
Oussama Mohammed Benhamed, PhD Candidate
Alberto Camara Ballesteros, PhD Candidate
One of the most complete collections of wild crucifers in the world (1,027 taxa with 4,863 accessions); Seeds of Iberian and Macaronesian endemic species; currently preserves 24% of the threatened flora in Spain
The EURISCO Web catalogue automatically receives data from the European National Inventories (NI). It provides information at the accession level of PGR conserved in European genebanks or other collections. EURISCO is hosted at and maintained by IPK Gatersleben on behalf of the Secretariat. Click here for further information and access.

- **Genesys** (global portal to information about Plant Genetic Resources for FA)
- **GLIS** (Global Information System for PGRFA)

**EURISCO**

*European Search Catalogue for Plant Genetic Resources*

- National focal points
  - (European National Inventories)
- National Genebanks Networks
  - (Crops conserved ex-situ)
A plethora of new stakeholders with highly varying technical experience & varying levels of compliance with emergent FAO descriptors.

**EURISCO**
(European Search Catalogue for Plant Genetic Resources)

National focal points
(European National Inventories on PGR)

National Genebanks Networks
(Crops conserved ex-situ)

- Environmental data
- Legal regulations on territories
- Legal regulations on species
- Conservation status
- Research data on species

Crop Wild Relatives

NGOs
Research centers
Citizens
Universities
A plethora of new stakeholders with highly varying technical experience & varying levels of compliance with FAO descriptors

This is identical to the situation faced by the registries and biobanks in EJP-RD!
A plethora of new stakeholders with highly varying technical experience & varying levels of compliance with FAO descriptors

FAIR Data facilitates communication between partners within the CWR community, and should also improve communication with other indexes such as EURISCO.

EURISCO
(European Search Catalogue for Plant Genetic Resources)

National focal points
(European National Inventories on PGR)

National Genebanks Networks
(Crops conserved ex-situ)

*environmental data
*legal regulations on territories
*legal regulations on species
*conservation status
*research data on species

A plethora of new stakeholders with highly varying technical experience & varying levels of compliance with FAO descriptors

NGOs
Research centers
Citizens
Universities
FLAIR-GG Objectives

Replicate EJP-RDs success with BGV as our first target

…then expand!
FLAIR-GG Status

FAIR Data Point:

- Customized to the Germplasm case
- Automated installer available
- Metadata capture templates available
- We offer to host the FDP for any new network partner to reduce cost-of-entry
FLAIR-GG Status

**Models and Templates**

- Shared data models have been created
- CSV Templates are completed
- Transformation pipeline tests confirm success
- FLAIR-GG infrastructure can be replicated by any partner within minutes.
FLAIR-GG Status

FAIR Data Point Federated Partner Index:

- So far… we’re quite lonely!
- Several seedbank* partners identified who should be onboard within a few weeks
- Everyone is welcome to participate!! Contact me!

* Mainly from REDBAG (Red Española de Bancos de Germoplasma de Plantas Silvestres y Fitorrecursos Autóctonos)
FLAIR-GG Status

FLAIR-GG “Virtual Platform”

- Entrypoint for federated exploration of the partner network
- Drives traffic to partner websites (!!)
- One-click launching of question-specific analytics environments such as map-integration (possible because of FAIR annotations of data services)
FLAIR-GG Next Steps

Begin constructing shared queries to help build data-driven conservation strategies
Conservation Strategy - Breadth:

“What geographic locations have not been sampled by any collection expeditions from our network partners?”
Environment drives intra-species diversity - add resources such as AEMET (Spanish Meteorological Agency), IGME (Geological and Mining Institute), CNIG (Spanish Geographical Agency) to capture environmental information associated to territories where the species lives.

Conservation Strategy - Depth:

“Are there occurrence locations of species X that are within soil types or microclimates for which we lack samples in our germplasm banks?”
Take Home Messages

FAIR is metadata first!!!!

FAIR is measurable, but we need global governance of testing before agencies can trust FAIR quality assessments

FAIR Data does not, alone, lead to interoperability - shared models required!

Rich, high-quality metadata enhances the appropriate reuse of FAIR data

Technologies/strategies allow FAIR experts to assist data owners in creating FAIR data themselves → distribute the effort, rather than centralize/warehouse

Don’t reinvent wheels - projects like EJP-RD and FLAIR-GG have generated a mountain of reusable code and models for FAIRification
Acknowledgements

My numerous and treasured collaborators and co-authors have been cited *in situ* throughout this slide deck.

TED2021-130788B-I00

https://tinyurl.com/UNBigData2024