

EOSC TF on FAIR Metrics and Data Quality

Chris Schubert - Mark D Wilkinson
(Carlo Lacagnina)
Task Force Co-Chairs

Romain David TF member (**presenter**)
on behalf of the FM&DQ TF members

ERINHA-AISBL - FR

romain.david@erinha.eu

ORCID : 0000-0003-4073-7456



romain_david_13



A set of principles, to ensure that data are shared in a way that enables and enhances reuse by humans and machines

Findable

- F1. (meta)data are assigned a globally unique and eternally persistent identifier.
- F2. data are described with rich metadata.
- F3. (meta)data are registered or indexed in a searchable resource.
- F4. metadata specify the data identifier.

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.

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.

Reusable

- R1. meta(data) have 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 their provenance.
 - R1.3. (meta)data meet domain-relevant community standards.

What the Principles DIDN'T Do

From the 2016 FAIR Principles paper:

*These high-level FAIR Guiding Principles precede implementation choices, and **do not suggest any specific technology, standard, or implementation-solution**; moreover, the **Principles are not, themselves, a standard or a specification**. They act as a guide to data publishers and stewards to assist them in evaluating whether their particular implementation choices are rendering their digital research artefacts Findable, Accessible, Interoperable, and Reusable. We anticipate that **these high level principles will enable a broad range of integrative and exploratory behaviours, based on a wide range of technology choices and implementations.***

ANNEX 1: Horizon 2020 FAIR Data Management Plan (DMP) Template

INTRODUCTION

This Horizon 2020 FAIR DMP template has been designed to be applicable to any Horizon 2020 project that produces, collects or processes research data. You should **develop a single DMP for your project** to cover its overall approach. However, where there are specific issues for individual datasets (e.g. regarding openness), you should clearly spell this out.

FAIR data management

In general terms, your research data should be 'FAIR', that is findable, accessible, interoperable and re-usable. These principles precede implementation choices and do not necessarily suggest any specific technology, standard, or implementation-solution.

ANNEX 1: Horizon 2020 FAIR Data Management Plan (DMP) Template

INTRODUCTION

This Horizon 2020 FAIR DMP template has been designed to be applicable to any Horizon 2020 project that produces, collects or processes research data. You should **develop a single DMP for your project** to cover its overall approach. However, where there are specific issues for individual datasets (e.g. regarding openness), you should clearly spell this out.

FAIR data management

In general terms, your research data should be 'FAIR', that is findable, accessible, interoperable and re-usable. These principles precede implementation choices and **do not necessarily suggest any specific technology, standard, or implementation-solution.**

BUT!!

ANNEX 1: Horizon 2020 FAIR Data Management Plan (DMP) Template

INTRODUCTION

This Horizon 2020 FAIR DMP template has been designed to be applicable to any Horizon 2020 project that produces, collects or processes research data. You should **develop a single DMP for your project** to cover its overall approach. However, where there are specific issues for individual datasets (e.g. regarding openness), you should clearly spell this out.

FAIR data management

In general terms, your research data should be 'FAIR', that is findable, accessible, interoperable and re-usable. These principles precede implementation choices and do not necessarily suggest any specific technology, standard, or implementation-solution.

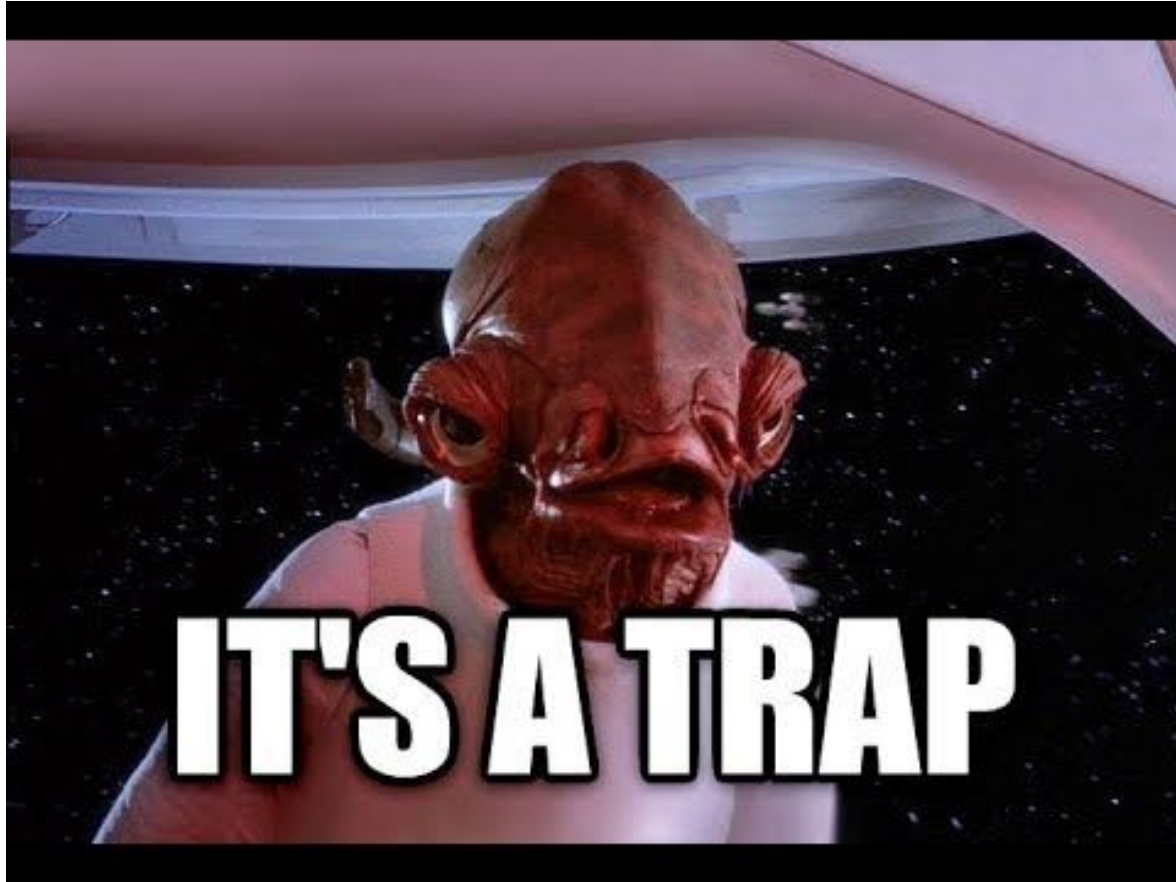
Contrast that with....

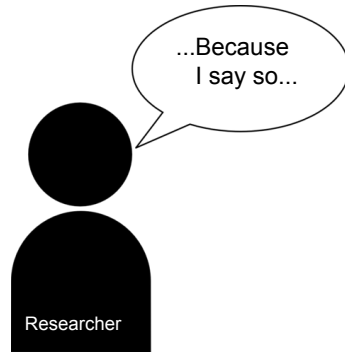
Commission High Level Expert Group on the European Open Science Cloud

Realising the European Open Science Cloud: first report and recommendations

20 June 2016

*Projects...that do not specify FAIR conditions for data...
should not be eligible for funding.*







FAIR assessment a cottage industry!

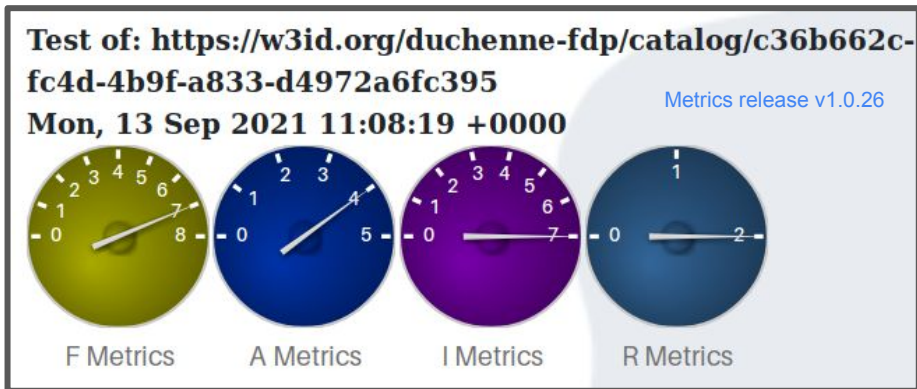
- Suffers from abundance!
 - **23** independent FAIR assessment platforms**
→ (see fairassist.org).
 - Most are questionnaire-based, several automated
 - **Outputs cannot be compared to one another!**

**** Demonstrates that the community of stakeholders are clamoring for a solution!**

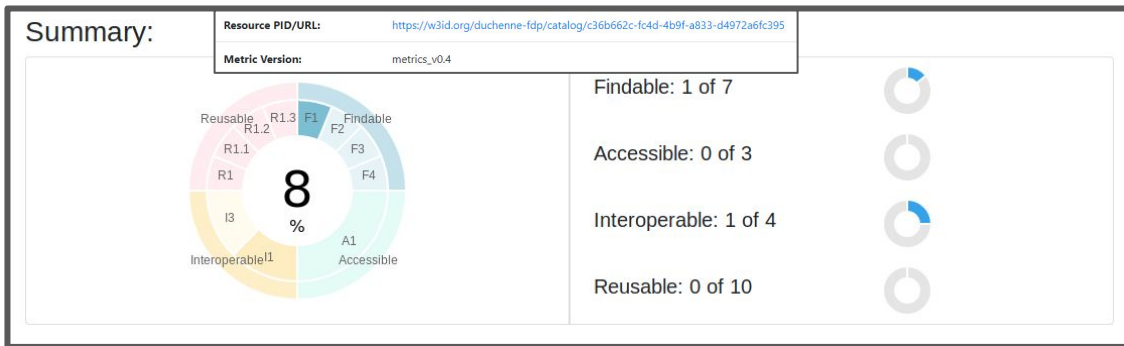
Resource ▾	Execution Type
5 Star Data Rating Tool	Manual - questionnaire
Data Stewardship Wizard	Predictive; based on a manually filled questionnaire
F-UJI	Automated
FAIR Data Self-Assessment Tool	Manual - questionnaire
FAIR Evaluator	Automated
FAIR enough?	Manual - checklist
FAIR-Aware (BETA)	Manual - questionnaire
FAIR-Checker	Automated
FAIRdat	Manual - questionnaire
FAIRness self-assessment grids	Manual - checklist
FAIRshake	Manual - questionnaire, Semi-manual
GARDIAN FAIR Metrics	Manual - checklist
RDA Maturity Model	Manual - checklist

How different can they be?

Comparison of The Evaluator with F-UJI, on the same URI
(a Catalog record in the Duchenne Muscular Dystrophy FAIR Data Point)



20/22 Tests Pass



2/24 Tests Pass

But... which one is *correct*?

But... which one is *correct*?

(The one that gives you the best score, obviously!)



But... which one is **correct**?

(The one that gives you the best score, obviously!)

Will this satisfy **reviewers**?

Will this satisfy **agencies**?
Journal **editors**?

Will this satisfy **businesses** who want to
purchase tools/software that claim to
“be FAIR”?

EOSC Task Force on FAIR Metrics and Data Quality

co-Chairs:
Mark D Wilkinson
Chris Schubert
(formerly Carlo Lacagnina)

Established November 2021

Chairs



Mark Wilkinson
UPM

Carlo Lacagnina
BSC

Members

Aguilar-Gómez, Fernando
CSIC

Al-Zoubi, Raed
ASREN

Bertino, Andrea
SWITCH

Biehlmaier, Oliver
Biozentrum, University of Basel

Cappiello, Cinzia
Politecnico di Milano

David, Romain
ERINHA AISBL

Dennis, Richard
Copenhagen University Library

Gingold, Arnaud
CNRS

Hajič, Jan
Charles University

Hecker, David
German Aerospace Center

Kleemola, Mari
CESSDA ERIC

Kuusniemi, Mari Elisa
OpenAIRE

Nikiforova, Anastasija
BBMRI

Nordling, Josefine
CSC

Papadopoulou, Elli
ATHENA RC

Sansone, Susanna-Assunta
University of Oxford

Schubert, Chris
TU Wien

Smit, Eefke
STM Association

Stryeck, Sarah
Graz University of Technology

Thiemann, Hannes
DKRZ

Velupillai, Sumithra
Swedish Research Council

von Stein, Ilona
DANS

Wright, Louise
EURAMET

Board Liaison



Sarah Jones
GÉANT

Outputs

[Task Force charter](#)



EOSC FAIR Working Group Recommendations on FAIR Metrics for EOSC:

“Support the definition and implementation of evaluation tools; their thorough assessment and evaluation including inclusiveness; comparison of tools (manual, automated); **identification of their biases** and applicability in many different contexts, including thematic ones.”

Check implementation of Metrics v.v.

- established quantitative criteria,
- measurement tools
 - F-UJI, The Evaluator, EOSC Synergy evaluator, AutoFAIR, FAIRshake, FAIRchecker

Exploring the problem @ workshops and hackathons

Creators of all automated FAIR assessment tools came together over 4 sessions

Discussed the bases for the differences in FAIR measurement

Decided that the complexity of **metadata discovery and harvesting** was the most critical problem - they each did it differently!

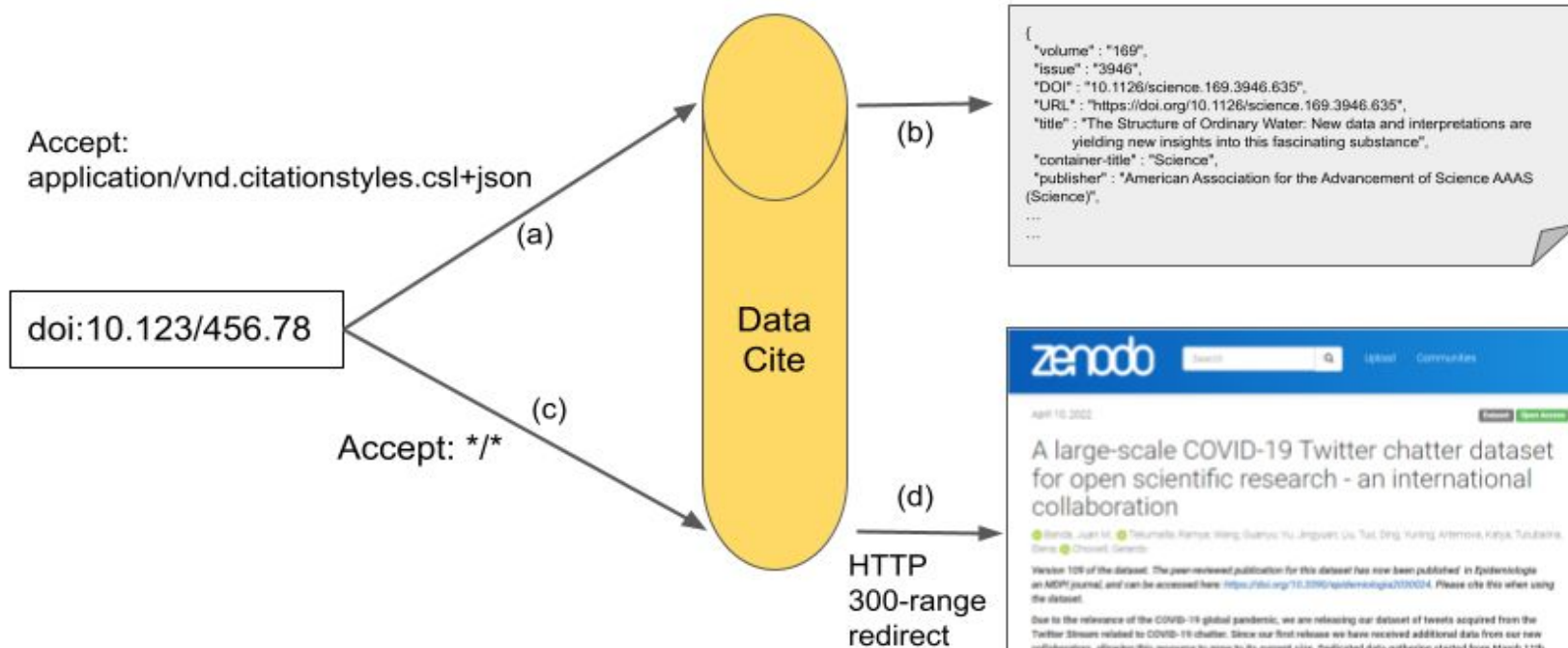
Impossible to compare tests when they are testing different “substrates”!

The problem of metadata discovery and interpretation

Exploration of a single common example: DOIs



Pathway to DOI resolution, including metadata



Eventually leads to a “landing page”

zenodo

April 10, 2022

A large-scale COVID-19 Twitter chatter dataset for open scientific research - an international collaboration

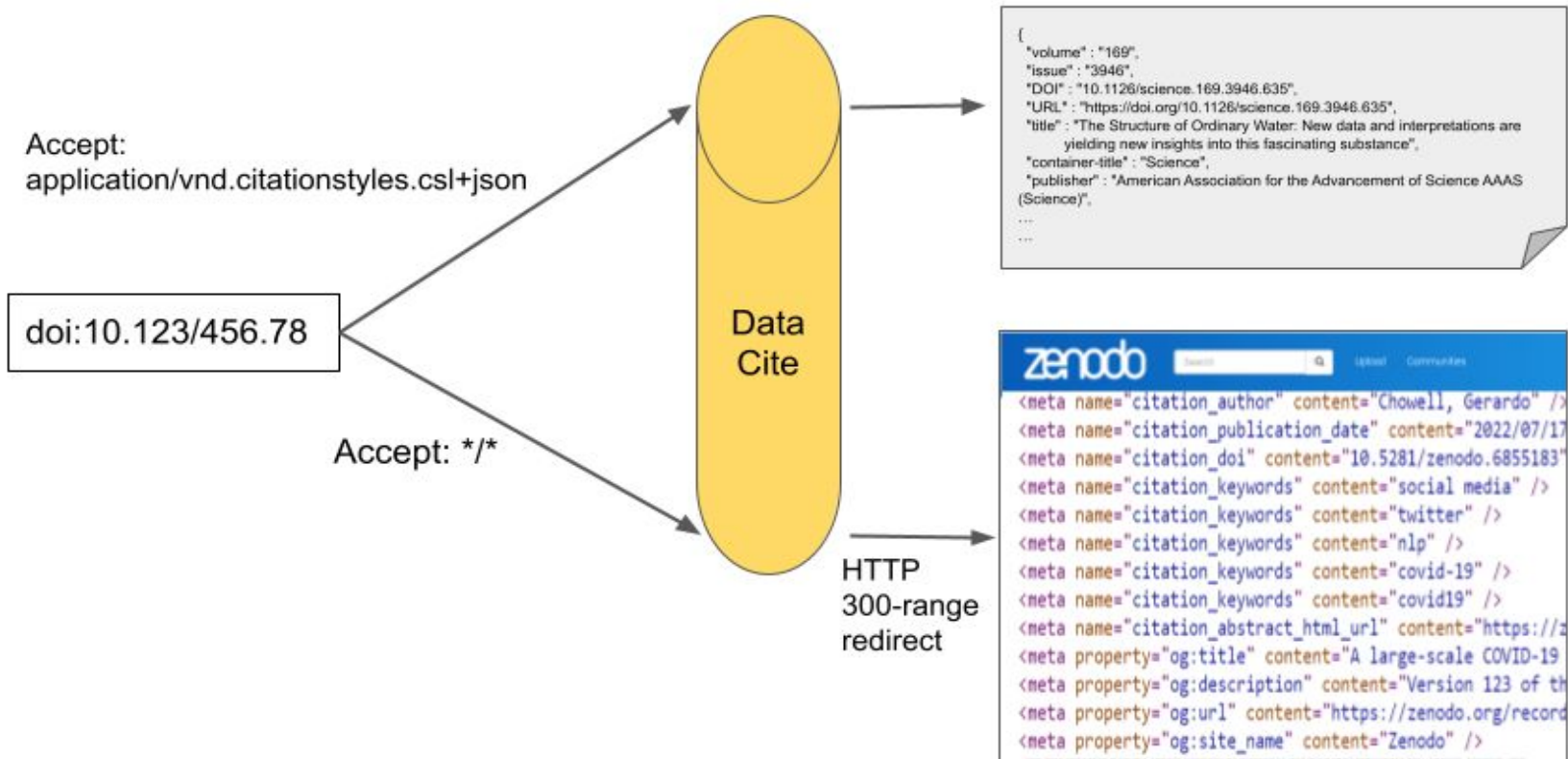
Version 109 of the dataset. The peer-reviewed publication for this dataset has now been published in *Epidemiology and Infection* journal, and can be accessed here <https://doi.org/10.1017/S0950268821000024>. Please cite this when using the dataset.

Due to the relevance of the COVID-19 global pandemic, we are releasing our dataset of tweets acquired from the Twitter Stream related to COVID-19 chatter. Since our first release we have received additional data from our new collaborators, allowing this resource to grow to its current size. Dedicated data gathering started from March 11th yielding over 4 million tweets a day. We have added additional data provided by our new collaborators from January 27th to March 27th, to provide extra longitudinal coverage. Version 10 added ~1.5 million tweets in the Russian language collected between January 1st and May 8th, graciously provided to us by Kalya Artemova (MSU HSE) and Denis Tsubolina (RTU). From version 12 we have included daily hashtags, mentions and emojis and their frequencies in the respective zip files. From version 14 we have included the tweet identifiers and their respective language for the clean version of the dataset. Since version 20 we have included language and place location for all tweets.

The data collected from the stream captures all languages, but the higher prevalence are: English, Spanish, and French. We release all tweets and retweets on the full_dataset for file (1,329,134,697 unique tweets), and a cleaned version with no retweets on the full_dataset-clean file (343,273,315 unique tweets). There are several practical reasons for us to leave the retweets, tracking important tweets and their dissemination is one of them. For NLP tasks we provide the top 1000 frequent terms in frequent_terms.csv, the top 1000 bigrams in frequent_bigrams.csv, and the top 1000 trigrams in frequent_trigrams.csv. Some general statistics per day are included for both datasets in the



Pathway to DOI resolution, including metadata



Landing page embedded metadata

Pathway to DOI resolution, including metadata

HTML “Typed Links”

```
<link rel="canonical" href="https://zenodo.org/record/6438032">
<link rel="alternate" type="application/zip" href="https://zenodo.org/record/6438032/files/emojis.zip">
<link rel="alternate" type="text/csv" href="https://zenodo.org/record/6438032/files/frequent_bigrams.csv">
<link rel="alternate" type="text/csv" href="https://zenodo.org/record/6438032/files/frequent_terms.csv">
<link rel="alternate" type="text/csv" href="https://zenodo.org/record/6438032/files/frequent_trigrams.csv">
<link rel="alternate" type="text/tab-separated-values" href="https://zenodo.org/record/6438032/files/full_d
<link rel="alternate" type="application/gzip" href="https://zenodo.org/record/6438032/files/full_dataset_cl
<link rel="alternate" type="text/tab-separated-values" href="https://zenodo.org/record/6438032/files/full_d
<link rel="alternate" type="application/gzip" href="https://zenodo.org/record/6438032/files/full_dataset.ts
<link rel="alternate" type="application/zip" href="https://zenodo.org/record/6438032/files/hashtags.zip">
<link rel="alternate" type="application/zip" href="https://zenodo.org/record/6438032/files/mentions.zip">
```

“If the `alternate` keyword is used with the `type` attribute, it indicates that the referenced document is a reformulation of the current document in the specified format.”

Too many sources of ambiguity

The metadata harvester has to guess what to do at many steps

There is overlap between the DataCite-sourced metadata and Zenodo metadata

The use of typed links leaves ambiguity

The interpretation of the “landing page” itself is ambiguous

- Some DOIs resolve directly to data, this one resolves to a landing page
- What, then, does the DOI represent? The landing page, or the data?

There is no way to support provider-sourced metadata (the most important stuff!)

This is just one example!

A harmonized approach is needed

We need to define a metadata publishing paradigm that will:

1. Support all publishers (both large and small; i.e. low complexity!)
2. Support the agents that are exploring them
3. Be unambiguous
4. Work on all types of digital object
 - a. “Traditional” data
 - b. Software
 - c. Workflows
5. Provide access to the most important metadata: that of the data creator!

Decision from the EOSC Workshops & Hackathons



FAIR Metrics and Data Quality
Task Force

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

10.5281/zenodo.7463421

“FAIR Signposting”

Three things are necessary for successful traversal of a FAIR Record:

1. Unambiguous identification of the GUID for the record
2. Unambiguous identification of the metadata record(s)
3. Unambiguous identification of the data record(s)

Using the well-established technology of “Links”, we defined a subset of Link relation types that can address these three requirements

Workshop and Hackathon Attendees

Mark D Wilkinson

Herbert Van de Sompel

Susanna-Assunta Sansone

Marjan Grootveld

Josefine Nordling

Richard Dennis

David Hecker

Erik Schultes

Andreas Czerniak

Stian

Allyson Lister

Milo Thurston

Philippe Rocca-Serra

Leonidas Pispiringas

Tim Smith

Sonia Barbosa

Wilko Steinhoff

Avi Ma'ayan

Carole Goble

Ceilyn Boyd

Kristian Garza

Soilan Peter Doorn

Alban Gaignard

Thomas Rosnet

Antonis Lempesis

Luiz Bonino

Michel Dumontier

Vincent Emonet

Robert Huber

Barbara Magagna

Marie-Dominique Devignes

Table 1: Link Relations used by FAIR Signposting	
Relation	Usage
cite-as	A one-to-one relationship between the entity and its globally unique identifier
describedby	A one-to-many relationship between the entity and all known metadata records about that entity
item	A one-to-many relationship between an entity representing a deposit and the data file(s) it contains.

These links can appear in:

- The body of the HTML (“Typed Links”)
- The Headers of the HTTP message (“Link Headers”)

Therefore can be used on both Web pages, as well as other non-HTML digital objects`

Starting Point:

Web Search
Bookmark
DOI resolution
Other ID resolution




...



April 10, 2022

Follow Open Access

A large-scale COVID-19 Twitter chatter dataset for open scientific research - an international collaboration

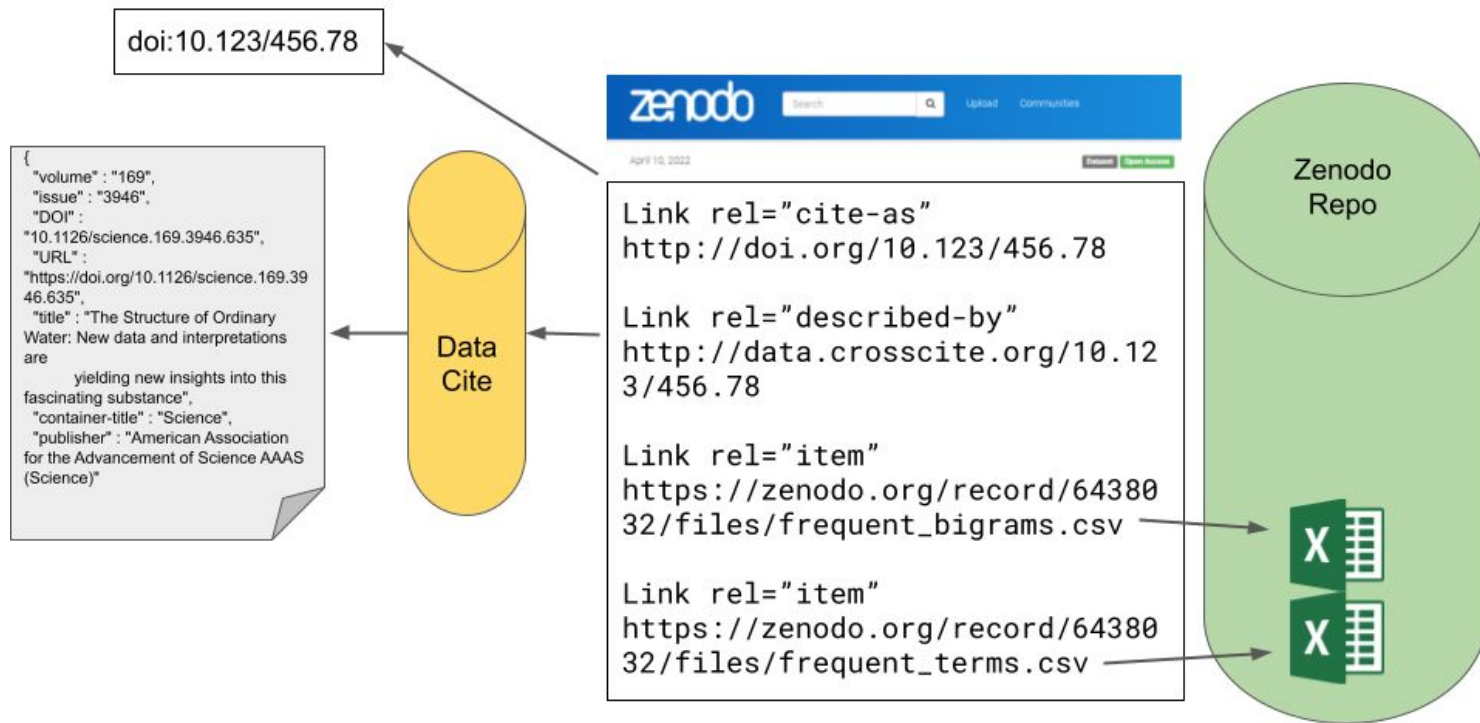
 Banda, Juan M.,
  Tekumalla, Ramya,
 Wang, Guanyu,
 Yu, Jingyuan,
 Liu, Tuo,
 Ding, Yuning,
 Artemova, Katya,
 Tutubalina, Elena,
  Chowell, Gerardo

Version 109 of the dataset. The peer-reviewed publication for this dataset has now been published in *Epidemiologia* an MDPI journal, and can be accessed here: <https://doi.org/10.3390/epidemiologia2030024>. Please cite this when using the dataset.

Due to the relevance of the COVID-19 global pandemic, we are releasing our dataset of tweets acquired from the Twitter Stream related to COVID-19 chatter. Since our first release we have received additional data from our new collaborators, allowing this resource to grow to its current size. Dedicated data gathering started from March 11th yielding over 4 million tweets a day. We have added additional data provided by our new collaborators from January 27th to March 27th, to provide extra longitudinal coverage. Version 10 added ~1.5 million tweets in the Russian language collected between January 1st and May 8th, gracefully provided to us by: Katya Artemova (NRU HSE) and Elena Tutubalina (KFU). From version 12 we have included daily hashtags, mentions and emojis and their frequencies the respective zip files. From version 14 we have included the tweet identifiers and their respective language for the clean version of the dataset. Since version 20 we have included language and place location for all tweets.

The data collected from the stream captures all languages, but the higher prevalence are: English, Spanish, and French. We release all tweets and retweets on the `full_dataset.tsv` file (1,329,136,097 unique tweets), and a cleaned version with no retweets on the `full_dataset-clean.tsv` file (343,273,315 unique tweets). There are several practical reasons for us to leave the retweets, tracing important tweets and their dissemination is one of them. For NLP tasks we provide the top 1000 frequent terms in `frequent_terms.csv`, the top 1000 bigrams in `frequent_bigrams.csv`, and the top 1000 trigrams in `frequent_trigrams.csv`. Some general statistics per day are included for both datasets in the

FAIR Signposting Harvesting Workflow

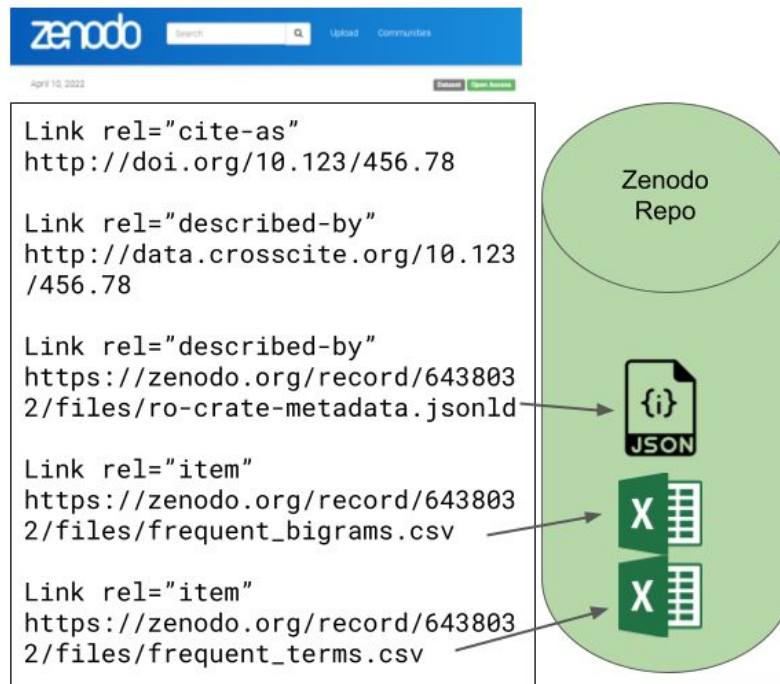


The “purpose” of the Landing Page is now unambiguous. It is a “broker” pointing at all other entities required by a FAIR record

Better yet!!

There is (finally!) an unambiguous way to support a data provider's own contextual metadata about the record they have deposited!

(Here I am pointing to a metadata record published using the newly established RO-Crate specification)



File Icon by Mohit Gandhi

FAIR Signposting Harvesting Workflow

HTTP
Link Headers

```
Link rel="cite-as"  
https://upload.wikimedia.org/wikipedia/commons/9/91/Mona\_Lisa\_vectorized.svg  
  
Link rel="described-by"  
https://commons.wikimedia.org/wiki/File:Mona\_Lisa\_vectorized.svg#metadata
```

Starting Point:

- Web Search
- Bookmark
- DOI resolution
- Other ID resolution
- ...



Sebastian Wallroth, CC0, via Wikimedia Commons

We can do the same thing without a landing page through Link Headers, thus supporting all kinds of digital object

Benchmarks for Apples-to-Apples FAIR Signposting

These are the Apples-to-Apples FAIR Signposting benchmark tests for tools to verify parsing and compliance with the FAIR Signposting profile.

Benchmarks

- 01-http-describedby-only/
- 02-html-full/
- 03-http-citeas-only/
- 04-http-describedby-iri/
- 05-http-describedby-citeas/
- 06-http-citeas-describedby-item/
- 07-http-describedby-citeas-linkset-json/
- 08-http-describedby-citeas-linkset-txt/
- 09-http-describedby-citeas-linkset-json-txt/
- 10-http-citeas-not-perma/
- 11-http-describedby-iri-wrong-type/
- 12-http-item-does-not-resolve/
- 13-http-describedby-with-type/
- 14-http-describedby-citeas-linkset-json-txt-conneg/
- 15-http-describedby-no-conneg/
- 16-http-describedby-conneg/
- 17-http-citeas-multiple-rels/
- 18-html-citeas-only/

We have 34 Benchmark tests

positive examples and
negative examples

that we can use to challenge the various
metadata harvesting workflows to
ensure that they truly are all working in
exactly the same way

The first step in harmonization of FAIR
assessments...

What do we see in FAIR's future?



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

Authorship Community:

Mark D. Wilkinson^{1,3}
Susanna-Assunta Sansone^{2,4}
Eva Méndez⁵
Romain David^{2,6}
Richard Dennis^{2,7}
David Hecker^{2,8}
Mari Kleemola^{2,9}
Carlo Lacagnina^{1,10}
Anastasija Nikiforova^{2,11}
Leyla Jael Castro¹²

1. Co-Chair, EOSC Task Force on FAIR Metrics and Data Quality
2. Member, EOSC Task Force on FAIR Metrics and Data Quality



Personal opinions on the next-steps for FAIR

There is a lot at-stake for FAIR Stakeholders - They will be judged on their FAIRness!

Therefore we need FAIR to fulfil its original objective of being Professional,

- Ensure it is considered **trustworthy, objective, valid, and achievable**

To do this we (all stakeholders) need to agree on some form of **governance**

The EOSC Task Force on FAIR Metrics and Data Quality has just issued a whitepaper describing a proposed governance model for FAIR assessments (available soon!) and an invitation to **join the founding stakeholders group that will establish the charter and continuity plan for a FAIR assessment governance body.**

FAIR Governance Model Whitepaper:

Mark D. Wilkinson
Susanna-Assunta Sansone
Eva Méndez
Romain David
Richard Dennis
David Hecker
Mari Kleemola
Carlo Lacagnina
Anastasija Nikiforova
Leyla Jael Castro

Personal opinions on the next-steps for FAIR

What could FAIR governance look like?

Top-down? Who is at the top? Who would be a trusted, arms-length third party with sufficient knowledge?

Bottom-up? Community-driven? Stakeholder-driven?
Stakeholders have vested interests... will they sufficiently agree? Isn't that what we already have?

Mixed? W3C model with open, but member-vetted, new memberships?

Testing-only? Is it enough to govern only the assessment/testing aspect of FAIR? Do the Principles themselves need governance? (The FAIR4RS process suggests the existing Principles may not be sufficient!)

Personal opinions on the next-steps for FAIR

Task Force Activity: Surveys

Questionnaires covering various aspects of the following issues:

- 1) Are people aware of the FAIR Principles
- 2) Are people aware that there are FAIR Evaluation tools?
- 3) Are people aware that they will be evaluated (whether they want to be or not!)
- 4) How do they feel about being evaluated
- 5) Are they aware of the evaluation tools, and how they work
- 6) What would have to happen to increase their level of comfort with being evaluated?
 - a) Rigorous peer-review of tools?
 - b) A trusted governance body
 - c) Participation of their community members in the governance process
- 7)

Acknowledgements

Memberships, Affiliations, and Supporters of these works

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



European Research Infrastructure
on Highly Pathogenic Agents



Instituto Nacional de Investigación
y Tecnología Agraria y Alimentaria



CBGP
CENTRO DE BIOTECNOLOGÍA
Y GENÓMICA DE PLANTAS



EXCELENCIA
SEVERO
OCHOA
| 2022 - 2025
| 2017 - 2021



CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS



BGV-UPM

"César Gómez Campo"



European Union's Horizon 2020
grant agreement N°825575



TED2021-130788B-I00

