



Why is FAIR Evaluation failing?



Presentation at the EOSC Symposium

Mark D Wilkinson

mark.wilkinson@upm.es





FAIR Assessment a cottage industry

- **22** independent FAIR assessment platforms
- Most are questionnaire-based
- **Outputs cannot be compared to one another!**

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

<https://fairassist.org>

How different can they be?

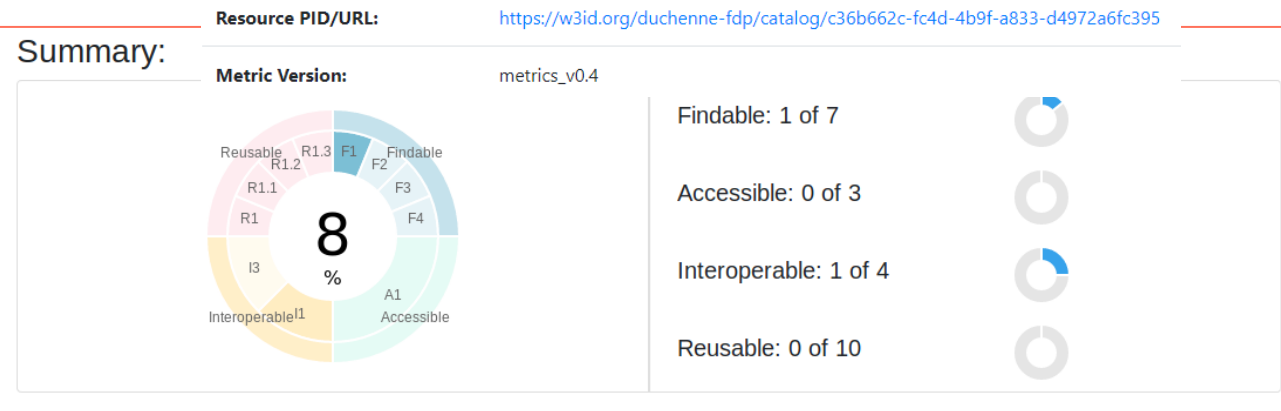
Comparison of The Evaluator with F-UJI, on the same URI
(a Catalog record in the Duchenne Muscular Dystrophy FAIR Data Point)

Test of: <https://w3id.org/duchenne-fdp/catalog/c36b662c-fc4d-4b9f-a833-d4972a6fc395> Metrics release v1.0.26

Mon, 13 Sep 2021 11:08:19 +0000



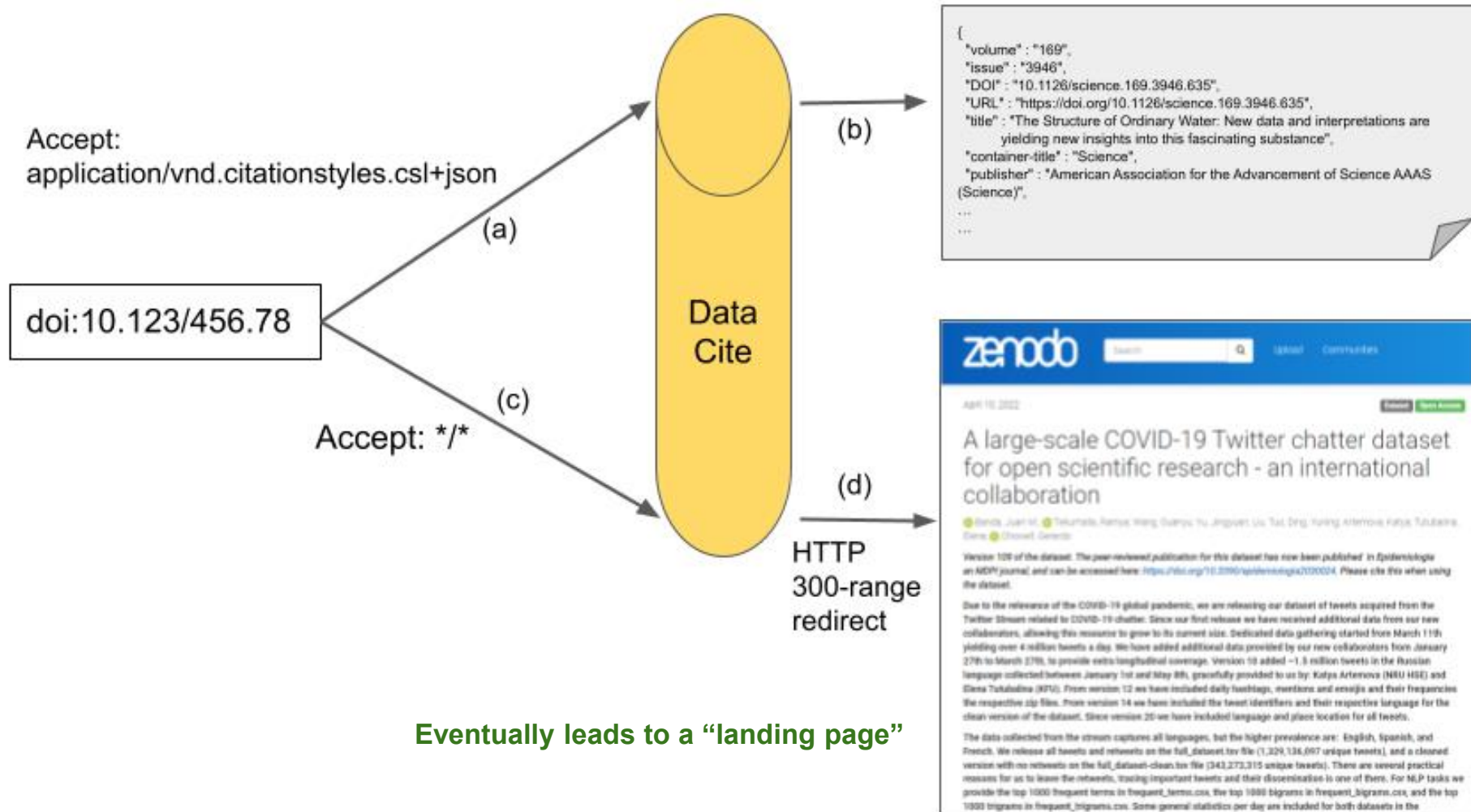
20/22 Tests Pass



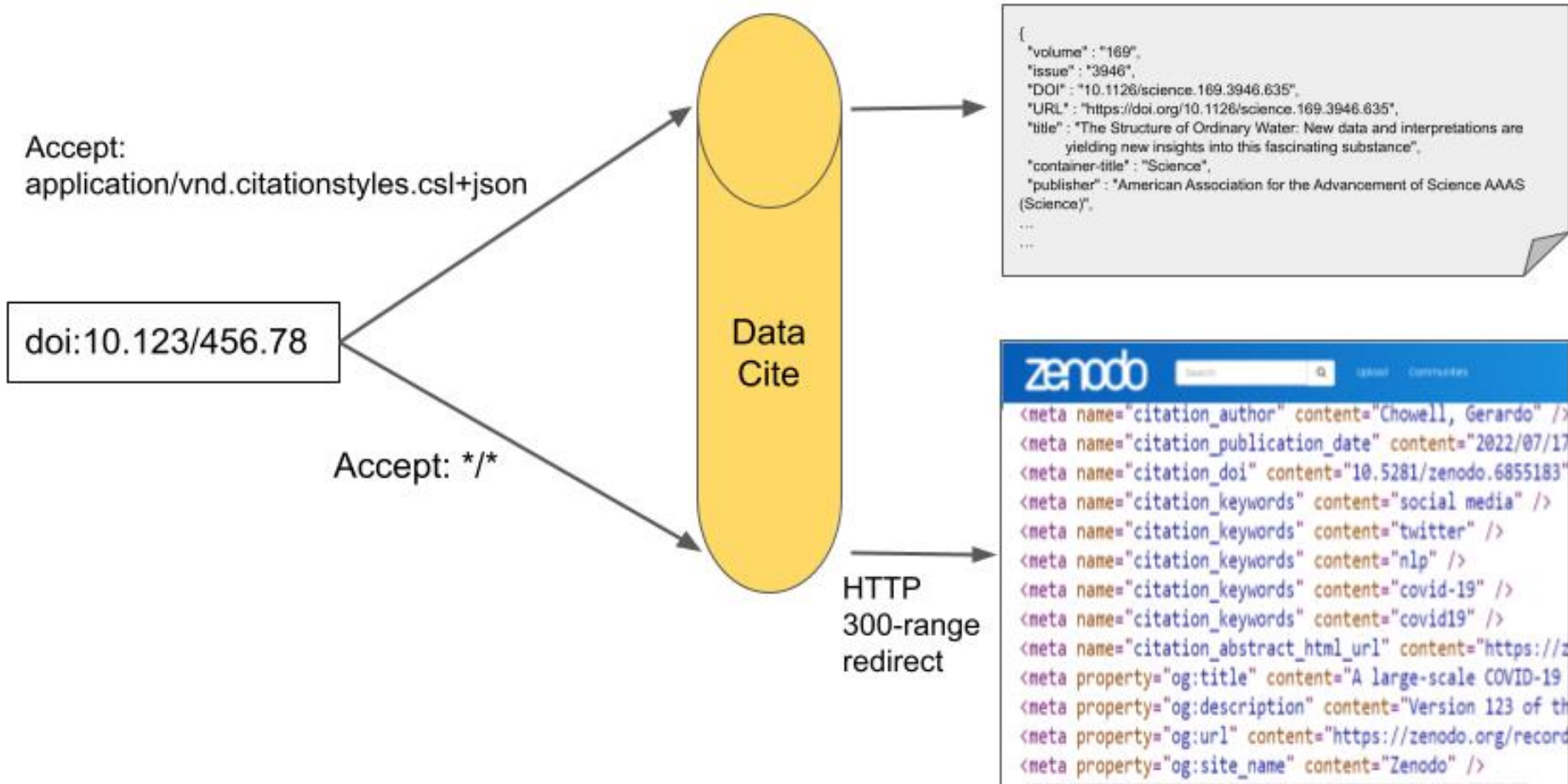
2/24 Tests Pass



Typical DOI resolution



Typical DOI resolution



Landing page embedded metadata

Typical DOI resolution

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.”

+

+

+

+

Many sources of ambiguity

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

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

The use of typed links leaves ambiguity due to different interpretations of the spec

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! (and DOI is perhaps the most widely recognized scholarly identifier)

Output from the EOSC Workshops & Hackathons



FAIR Metrics and Data Quality
Task Force

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

“FAIR Signposting”

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

- Unambiguous **identification of the GUID** for the record
- Unambiguous **identification of the metadata** record(s)
- Unambiguous **identification of the data** record(s)

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

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 Soiland-Reyes
Allyson Lister
Milo Thurston
Philippe Rocca-Serra

Leonidas Pispiringas
Tim Smith
Sonia Barbosa
Wilko Steinhoff
Avi Ma'ayan
Carole Goble
Ceilyn Boyd
Kristian Garza
Peter Doorn
Alban Gaignard
Thomas Rosnet

Antonis Lempesis
Luiz Bonino
Michel Dumontier
Vincent Emonet
Robert Huber
Barbara Magagna
Marie-Dominique Devignes

FAIR Signposting

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



Signposting workflow

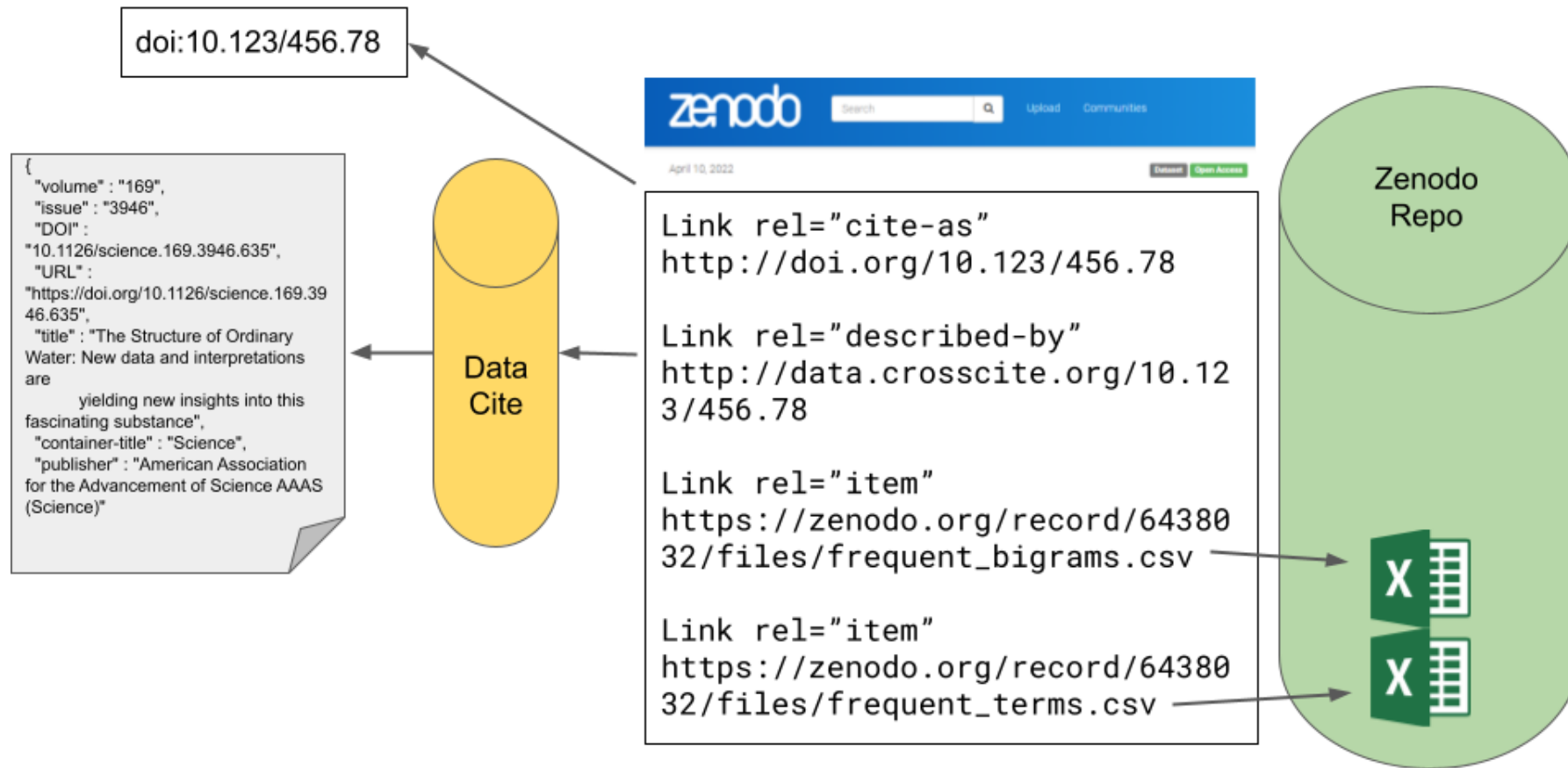
Starting Point:

Web Search
Bookmark
DOI resolution
Other ID resolution
...



The screenshot shows the Zenodo website interface. At the top, there is a blue header with the Zenodo logo, a search bar, and links for 'Upload' and 'Communities'. Below the header, the date 'April 10, 2022' is displayed on the left, and 'Dataset' and 'Open Access' labels are on the right. The main title of the dataset is 'A large-scale COVID-19 Twitter chatter dataset for open scientific research - an international collaboration'. Below the title, the authors are listed: 'Sando, Juan M.; Tekumalla, Ramiya; Wang, Quanyu; Yu, Jingyuan; Liu, Tao; Ding, Yuning; Artemova, Katya; Tutubalina, Elena; Chowell, Gerardo'. A paragraph of text follows, stating: 'Version 109 of the dataset. The peer-reviewed publication for this dataset has now been published in Epidemiologia an MDPJ journal, and can be accessed here: <https://doi.org/10.3390/epidemiologia2030024>. Please cite this when using the dataset.' Another paragraph explains the dataset's relevance and updates: '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: 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.' A final paragraph provides details on the data collection: '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'.

Signposting workflow



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

Signposting workflow

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 RO-Crate specification)



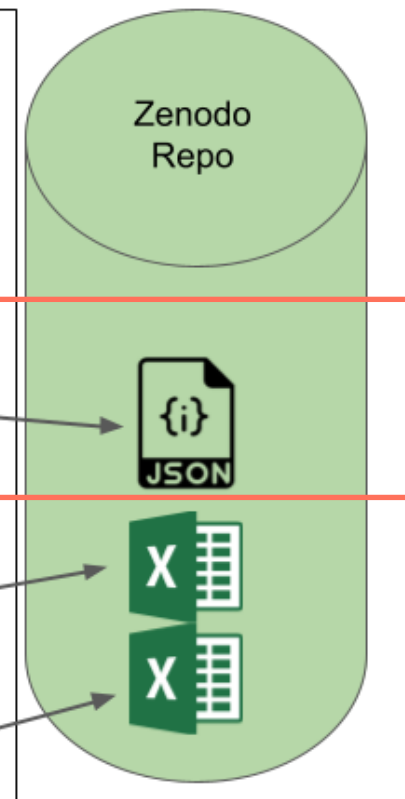
```
Link rel="cite-as"  
http://doi.org/10.123/456.78
```

```
Link rel="described-by"  
http://data.crosscite.org/10.123/456.78
```

```
Link rel="described-by"  
https://zenodo.org/record/643803  
2/files/ro-crate-metadata.jsonld
```

```
Link rel="item"  
https://zenodo.org/record/643803  
2/files/frequent_bigrams.csv
```

```
Link rel="item"  
https://zenodo.org/record/643803  
2/files/frequent_terms.csv
```



File Icon by Mohit Gandhi

Signposting 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






Professionalism

We have **34 Benchmark tests**

positive examples and
negative examples



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

Next FAIR Assessment hackathon in the last week of
September

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/](#)

FAIR Signposting “in the wild”

Signposting has already been added to the latest (5.14) release of Dataverse

Meeting between Hackathon attendees and the NIH [Generalist Repository Ecosystem Initiative](#) (GREI) is scheduled for October

- Presenting the idea to e.g. Zenodo, Figshare, etc.

Signposting on the TODO list for the reference implementation of the FAIR Data Point

Fin

