FAIRSFAIR

Fostering Fair Data Practices in Europe

Domain agnostic metadata schemas

(and how machines (FAIR assessment tools) can find them)

Robert Huber
(MARUM, Universität Bremen)
rhuber@uni-bremen.de

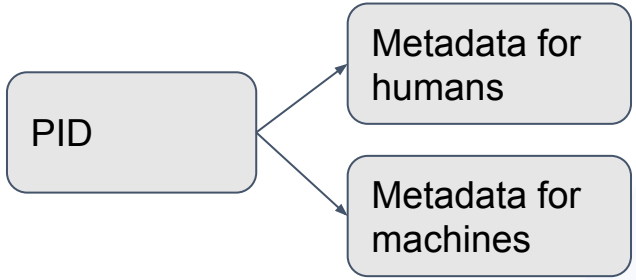
EOSC-Nordic FAIRification webinar, FAIR step 3
April 29, 2021

FAIRsFAIR "Fostering FAIR Data Practices in Europe" has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 Grant agreement 831558

FAIR principles to be considered for domain agnostic metadata provision

FAIR is for machine & humans
And for data & metadata

F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata



"The emphasis placed on FAIRness being applied to both human-driven and machine-driven activities, is a specific focus of the FAIR Guiding Principles that distinguishes them from many peer initiatives (discussed in the subsequent section)." (Wilkinson et al, 2016)

Metadata for machines and humans

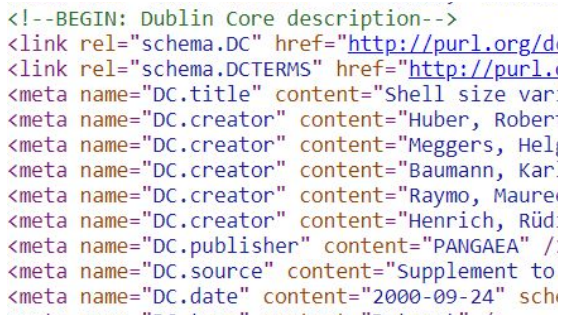
PID resolves to a human readable landing page

- Embedding metadata in HTML page (landing page)
- Retrieving machine metadata using content negotiation
- Providing machine readable links in HTML (signposting, typed links)

Metadata for machines and humans

Embedded:

Accept:text/html



Typed links:

Accept:text/html

Content negotiation:

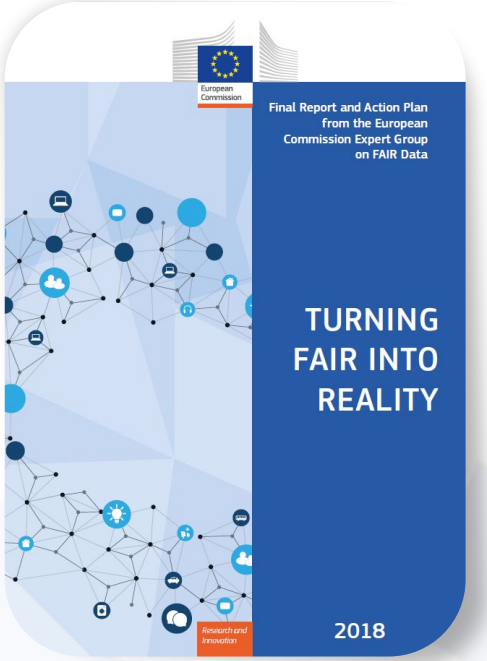
Accept:application/ld+json

FAIRsFAIR - FAIR Data Assessments

- FAIR assessment implementation comprises the development of two main components – assessment metrics and tool.

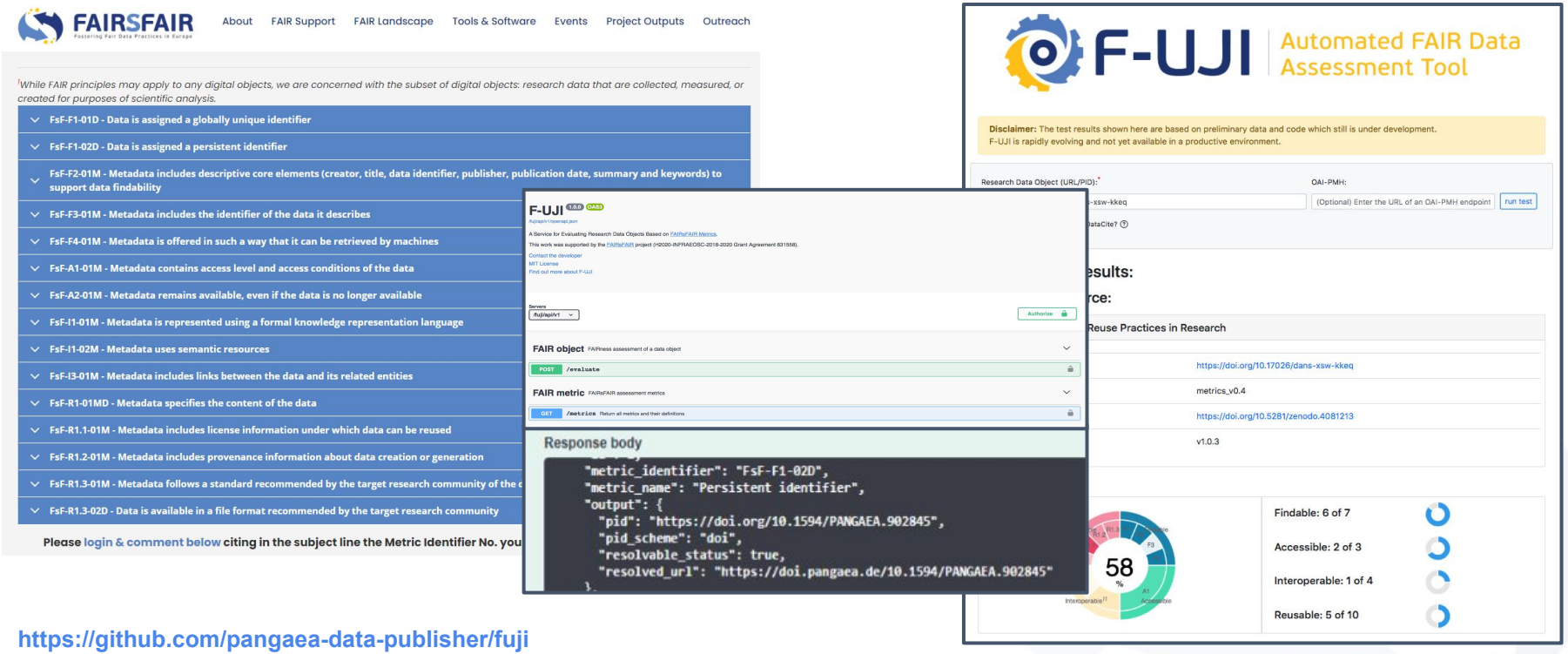
Priority Recommendations
Rec. 8: Facilitate automated processing
Rec. 12: Develop metrics for FAIR Digital Objects

Supporting Recommendations
Rec. 25: Implement FAIR metrics to monitor uptake



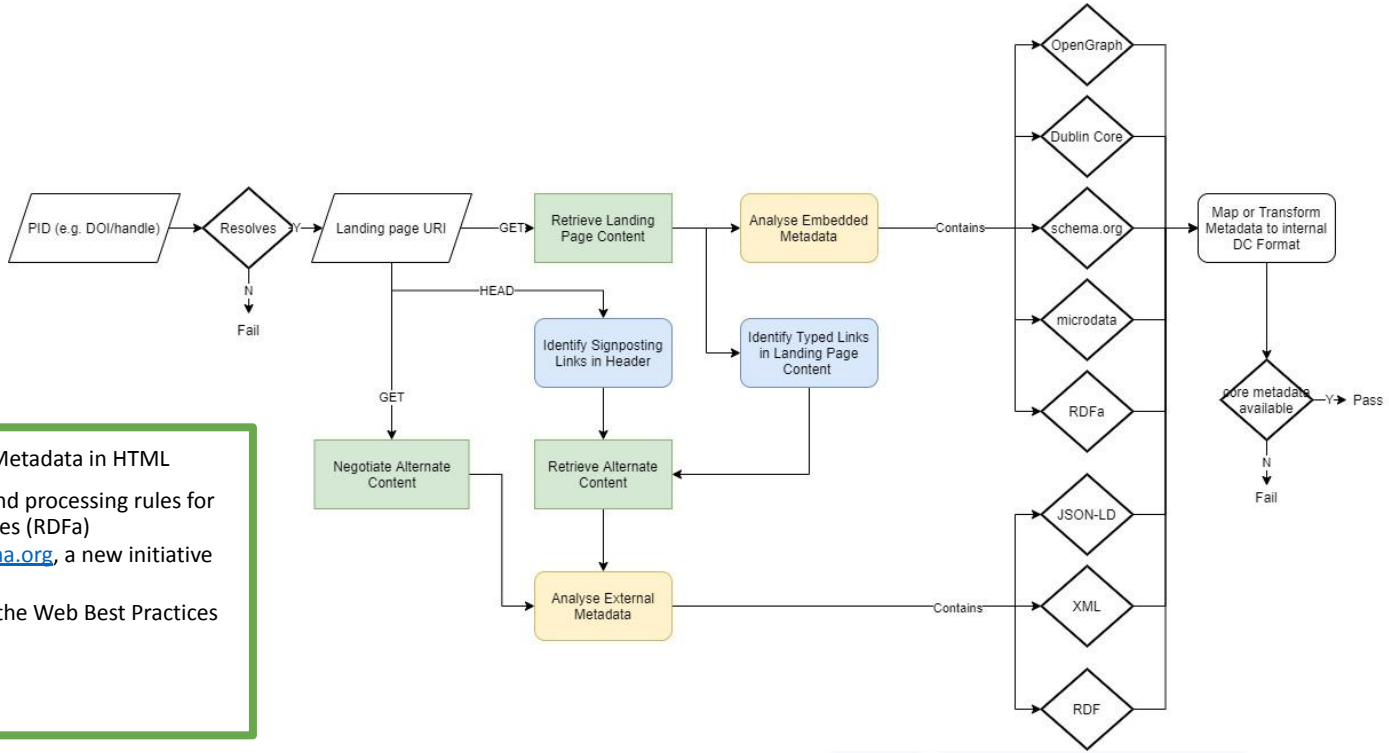
European Commission Expert Group on FAIR Data. 2018. 'Turning FAIR into Reality: Final Report and Action Plan from the European Commission Expert Group on FAIR Data.'
https://doi.org/10.2777/1524

F-UJI -FAIRsFAIR FAIR metric assessment tool



https://github.com/pangaea-data-publisher/fuji
https://www.f-uji.net

F-UJI: metadata discovery (F2)



RFC2731: Encoding Dublin Core Metadata in HTML

W3C recommendation: Syntax and processing rules for embedding RDF through attributes (RDFa)
...Today we're announcing schema.org, a new initiative from Google, Bing and Yahoo!...
W3C recommendation: Data on the Web Best Practices
RFC5988: web linking
Signposting.org
Datacite

Domain agnostic metadata schemas

Widely used:

- Dublin Core
- DCAT
- Schema.org
- (DataCite schema)

Dublin Core

- Metadata standard and vocabulary
- Can be expressed in:
  - XML (schema ...)
  - RDF
  - Microdata, RDFa
  - XHTML (`<meta name="DC.element" content="Value" />`)
- Protocol / Interface:
  - OAI-PMH
  - SPARQL
- Pro:
  - embedding in landing page
- Con:
  - Very generic, hard to link data

```
<?xml version="1.0" encoding="utf-8"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:dc="http://purl.org/dc/elements/1.1/">
  <rdf:Description>
    <dc:title>Homage to Catalonia,</dc:title>
    <dc:creator>Orwell, George,1903-1950.</dc:creator>
    <dc:type>text</dc:type>
    <dc:publisher>London, Secker and Warburg</dc:publisher>
    <dc:date>[1938]</dc:date>
    <dc:language>eng</dc:language>
  </rdf:Description>
</rdf:RDF>
```

ENDORSED BY:

Schema.org

schema.org

- Generic metadata vocabulary and schemas (e.g. Dataset)
- Can be expressed in:
  - RDF
  - JSON-LD


```
{
  "@context": "http://schema.org/",
  "@id": "https://doi.org/10.1594/PANGAEA.907686",
  "@type": "Dataset",
  "identifier": "https://doi.org/10.1594/PANGAEA.907686",
  "url": "https://doi.pangaea.de/10.1594/PANGAEA.907686",
  "creator": [
    {
      "@id": "https://orcid.org/0000-0003-3000-0020",
      "@type": "Person",
      "name": "Robert Huber",
      "familyName": "Huber",
      "givenName": "Robert",
      "identifier": "https://orcid.org/0000-0003-3000-0020",
      "email": "rhuber@uni-bremen.de"
    },
    {
      "@type": "Person",
      "name": "Robert Darga",
      "familyName": "Darga",
      "givenName": "Robert"
    },
    {
      "@type": "Person",
      "name": "Hans Lauterbach",
      "familyName": "Lauterbach",
      "givenName": "Hans"
    }
  ],
  "name": "Roundness of grains from sediments from the surroundings of Lake Tüttensee (Bavaria, Germany)",
  "publisher": {
    "@type": "Organization",
    "name": "PANGAEA",
    "disambiguatingDescription": "Data Publisher for Earth & Environmental Science",
```


- Protocol / Interface:
  - SPARQL
  - Google Dataset Search ;)
- Pro:
  - Suitable for embedding in landing page (human & machine)
  - SEO included
- Con:
  - Difficult to validate

ENDORSED BY: Google Bing yahoo!

DCAT (v2)

DCAT

- Generic data cataloging metadata vocabulary
- Can be expressed in:
  - RDF
  - JSON-LD
- Protocol / Interface:
  - SPARQL
- Pro:
  - Designed for data catalogs
  - Uses Dublin Core etc.
  - Extensible
  - Widely used incl. Open Data (gov)
  - Accepted by Google Dataset Search
- Con:
  - Difficult to validate

ENDORSED BY: W3C®

https://github.com/pangaea-data-publisher/fuji

https://www.f-uji.net

2021-02-03