

F-UJI: An Automated Assessment Tool for Improving the FAIRness of Research Data

Anusuriya Devaraju & Robert Huber {adevaraju, rhuber}@marum.de (on behalf of Task 4.5)





- Task 4.5 (FAIR Data Assessments: Pilots)
- FAIR Assessment of Research Data
 - Data Assessment Metrics
 - F-UJI: An Automated FAIR Data Assessment Tool
- Pilot Repositories & Results
- Conclusions

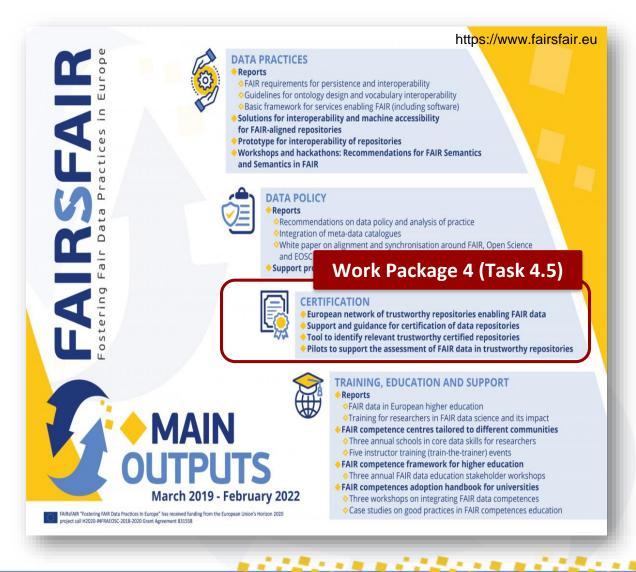


- Task 4.5 (FAIR Data Assessments: Pilots)
- FAIR Assessment of Research Data
 - Data Assessment Metrics
 - F-UJI: An Automated FAIR Data Assessment Tool
- Pilot Repositories & Results
- Conclusions



Fostering Fair Data Practices in Europe (FAIRsFAIR)

- Aims to supply practical solutions for the use of the FAIR principles throughout the research data life cycle.
- Starting date: 01.03.2019
- Duration: 36 months
- 22 partners from 8 Member States.





Task 4.5 – FAIR Data Assessments: Pilots

- The goal of the task is to pilot the FAIR assessment of research data from trustworthy data repositories.
- Research data...



• KPI 4.2: Metrics and badging scheme for assessment of FAIRness of individual datasets in trusted repositories tested and applied to 100 datasets in 5 CoreTrustSeal certified repositories.



Task 4.5 – FAIR Data Assessments: Pilots

 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



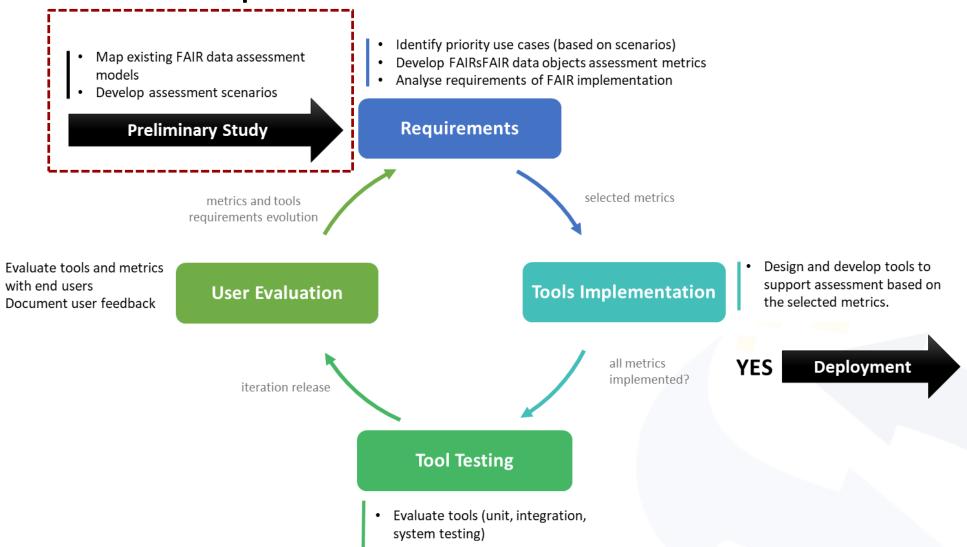
European Commission Expert Group on FAIR Data. 2018. 'Turning FAIR into Reality: Final Report and Action Plan from the European Commission Expert Group on FAIR Data.' https://doi.org/10.2777/1524



- Task 4.5 (FAIR Data Assessments: Pilots)
- FAIR Assessment of Research Data
 - Data Assessment Metrics
 - F-UJI: An Automated FAIR Data Assessment Tool
- Pilot Repositories & Results
- Conclusions



Iterative Development

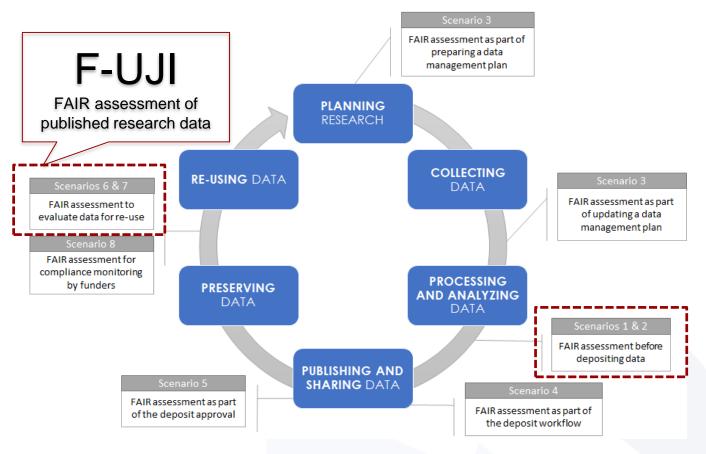




Stakeholders and Scenarios

For more information, see D4.1 Draft Recommendations on Requirements for Fair Datasets in Certified Repositories, https://doi.org/10.5281/zenodo.3678715





FAIR stakeholders; figure is derived from 8.3 Stakeholder Groups Assigned Actions (EC Expert Group on FAIR Data, 2018). Dotted lines represent the stakeholders of the FAIRsFAIR Task 4.5.

Research data lifecycle; figure adapted from (Mosconi et al., 2019) and scenarios of FAIR assessment of datasets therein.

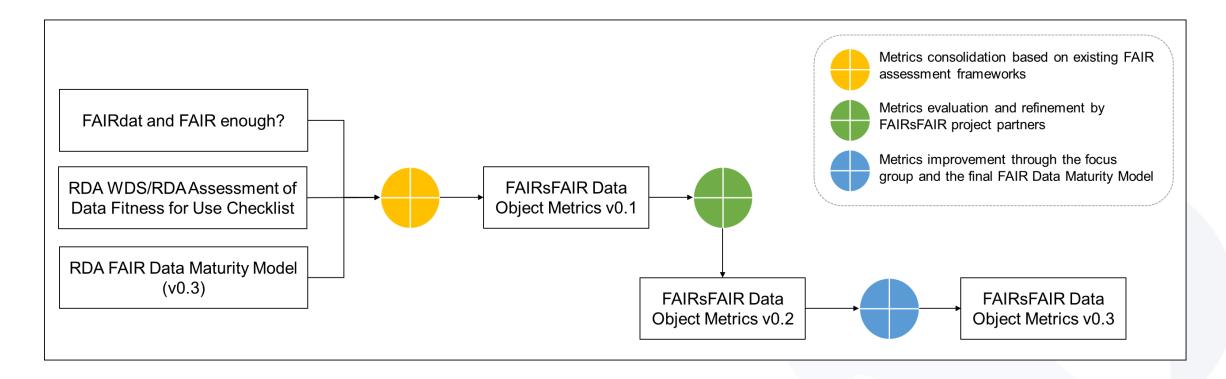


- Task 4.5 (FAIR Data Assessments: Pilots)
- FAIR Assessment of Research Data
 - Data Assessment Metrics
 - F-UJI: An Automated FAIR Data Assessment Tool
- Pilot Repositories & Results
- Conclusions



FAIRsFAIR Data Assessment Metrics

- There are 15 core metrics (v0.3) developed based on existing work:
 - RDA FAIR Data Maturity Model, Fairdat/FAIR Enough?, WDS/RDA Assessment of Data Fitness
- Correspond to a part of or the whole of a FAIR principle.





 The metric specification follows the template modified from (<u>Wilkinson et al.</u>, <u>2018</u>).

COMMENTS

Related Resources

- · Examples of metadata recommendations:
 - o EOSC EDMI metadata properties, https://eosc-edmi.gi/hub.io/properties
 - W3C Recommendation Data on the Web Best Practices, https://www.w3.org/TR/dwbp/#metadata
 - o W3C Data Catalog Vocabulary, https://www.w3.org/TR/vocab-dcat-2/
- · Sites that provide a list of metadata standards:
 - o FAIRsharing standards, https://fairsharing.org/standards/
 - o DCC List of Metadata Standards, http://www.dcc.ac.uk/resourcet/metadata-standards/list
 - RDA Metadata Directory (based on the DCC list), http://rd-alliance.github.io/metadata-directory/
- Examples of domain agnostic metadata standards for describing research date:
 - Dublin Core Metadata Initiative (DCMI) Metadata Terms, https://www.w3.org/TR/dwbp/#bib-DCTERMS
 - DataCite Metadata Schema, https://doi.org/10.14454/7xq3-zf69
 - o Schema.org, https://schema.org/Dataset
 - o Data Catalog Vocabulary (DCAT), https://www.w3.org/TR/vocab-dcat-2/

Known Limitations/Constraints

- The assessment assumes that the identifier resolves to a landing page (e.g., html) that contains the metadata of the data. Landing page may not necessarily be an html page.
- . Data providers may use different standards to expose the metadata of their data.
- The metadata records maintained by a data provider might not be accessible, due to, e.g., broken link of the landing page, proprietary metadata standard used, and restricted metadata.

FIELD	DESCRIPTION
Metric Identifier	FsF-F2-01M
Metric Name	Metadata includes descriptive core elements (creator, title, data identifier, publisher, publication date, summary and keywords) to support data findability.
Description	Metadata is descriptive information about a data object. Since the metadata required differs depending on the users and their applications, this metric focuses on core metadata. The core metadata is the minimum descriptive information required to enable data finding, including citation which makes it easier to find data. We determine the required metadata based on common data citation guidelines (e.g., DataCite, ESIP, and IASSIST), and metadata recommendations for data discovery (e.g., EOSC Datasets Minimum Information (EDMI), DataCite Metadata Schema, W3C Recommendation Data on the Web Best Practices and Data Catalog Vocabulary). This metric focuses on domain-agnostic core metadata. Domain or discipline-specific metadata specifications are covered under metric FsF-R1.3-01M. A repository should adopt a schema that includes properties of core metadata, whereas data authors should take the responsibility of providing core metadata.
FAIR Principle	F2. Data are described with rich metadata
CoreTrustSeal Alignment	R13. The repository enables users to discover the data and refer to them in a persistent way through proper citation
ASSESSMENT	
Requirement(s)	Data identifier (IRI, URL) Machine-accessible and readable metadata
Method	Use the data identifier to access its metadata document. Parse or retrieve core metadata, e.g., through one or more options below, combine the results and then verify presence/absence of the core elements in the metadata. Structured data embedded in the landing page of the identifier (e.g., Schema.org, Dublin Core and OpenGraph meta tags) Typed Links in the HTTP Link header; for more information, see https://signposting.org/conventions/ If the identifier specified is a persistent identifier, use it to retrieve the metadata of the data from its PID provider, e.g., see DataCite Content Resolver at https://datacite.org/content.html
COMMENTS	

化苯乙基乙烯化 化苯乙基乙烯化 化苯乙基乙烯

While FAIR principles may apply to any digital objects, we are concerned with the subset of digital objects: research data that are collected, measured, or created for purposes of scientific analysis.

- ✓ FsF-F1-01D Data is assigned a globally unique identifier
- FsF-F1-02D Data is assigned a persistent identifier
- FsF-F2-01M Metadata includes descriptive core elements (creator, title, data identifier, publisher, publication date, summary and keywords) to support data findability
- FsF-F3-01M Metadata includes the identifier of the data it describes
- FsF-F4-01M Metadata is offered in such a way that it can be retrieved by machines
- FsF-A1-01M Metadata contains access level and access conditions of the data
- FsF-A2-01M Metadata remains available, even if the data is no longer available
- FsF-I1-01M Metadata is represented using a formal knowledge representation language
- ✓ FsF-I1-02M Metadata uses semantic resources
- FsF-I3-01M Metadata includes links between the data and its related entities
- ∨ FsF-R1-01MD Metadata specifies the content of the data
- FSF-R1.1-01M Metadata includes license information under which data can be reused
- FsF-R1.2-01M Metadata includes provenance information about data creation or generation
- FSF-R1.3-01M Metadata follows a standard recommended by the target research community of the data
- FsF-R1.3-02D Data is available in a file format recommended by the target research community

Please login & comment below citing in the subject line the Metric Identifier No. you are referring to - e.g. "FsF-R1.3-01M"

Object Assessment Metrics v0.3

We would love to hear your feedback!

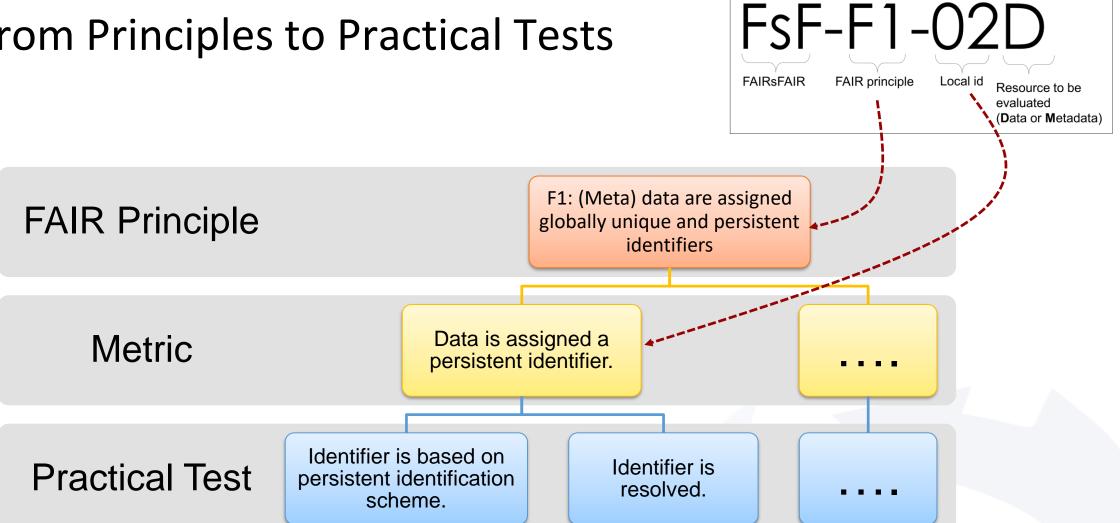
https://www.fairsfair.eu/fairsf air-data-object-assessmentmetrics-request-comments



- Task 4.5 (FAIR Data Assessments: Pilots)
- FAIR Assessment of Research Data
 - Data Assessment Metrics
 - F-UJI: An Automated FAIR Data Assessment Tool
- Pilot Repositories & Results
- Conclusions



From Principles to Practical Tests





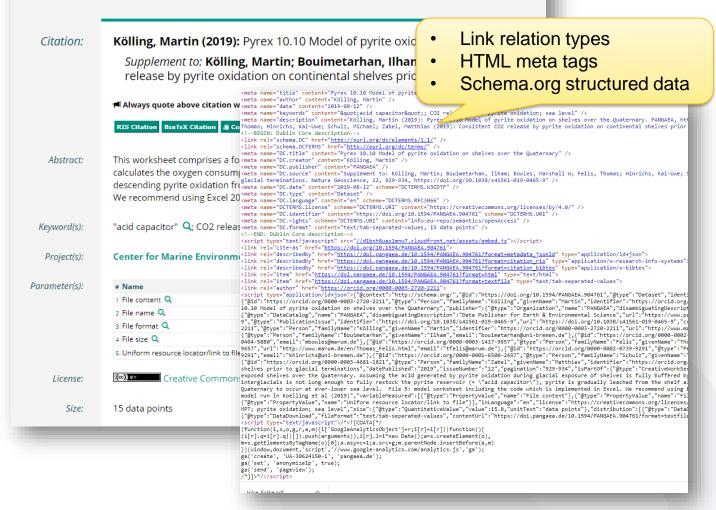
Resources

- Metadata
- Data file(s)
- External resources
 - PID provider service
 - r3data.org
 - SPDX license list
 - RDA Metadata Standards Catalog
 - LOV, LOD
 - ISO/TR 22299 (Digital file format recommendations for long-term storage)
 - Wolfram scientific formats
 - more



PANGAEA.

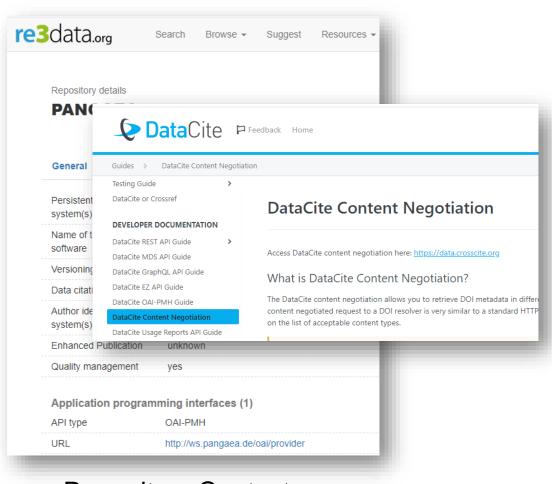
Data Publisher for Earth & Environmental Science

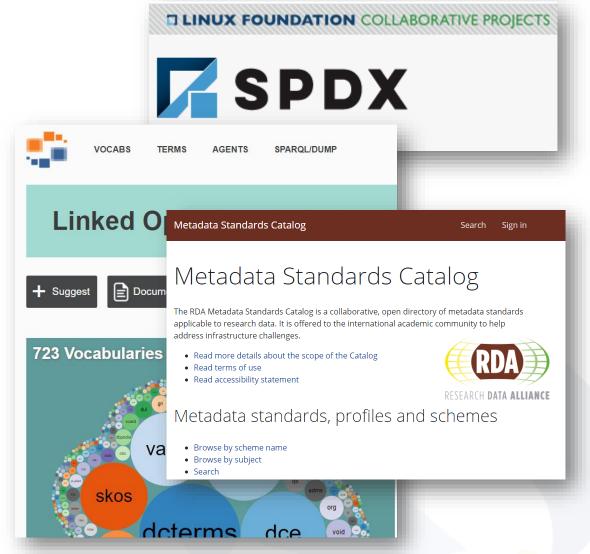




Assessment Enabling Services

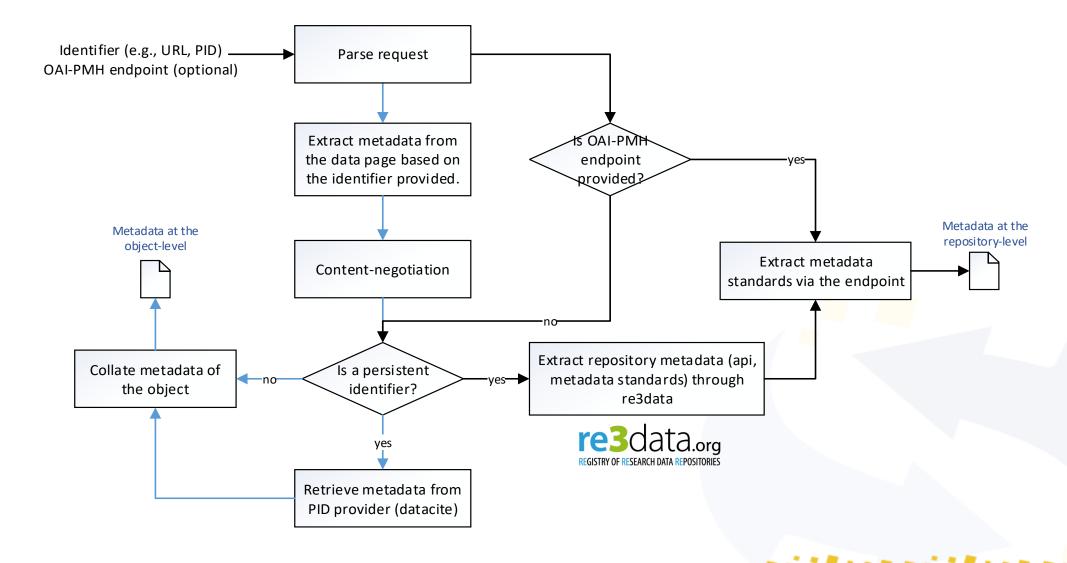
'Lookup' Services







High Level Flow (Data Gathering)



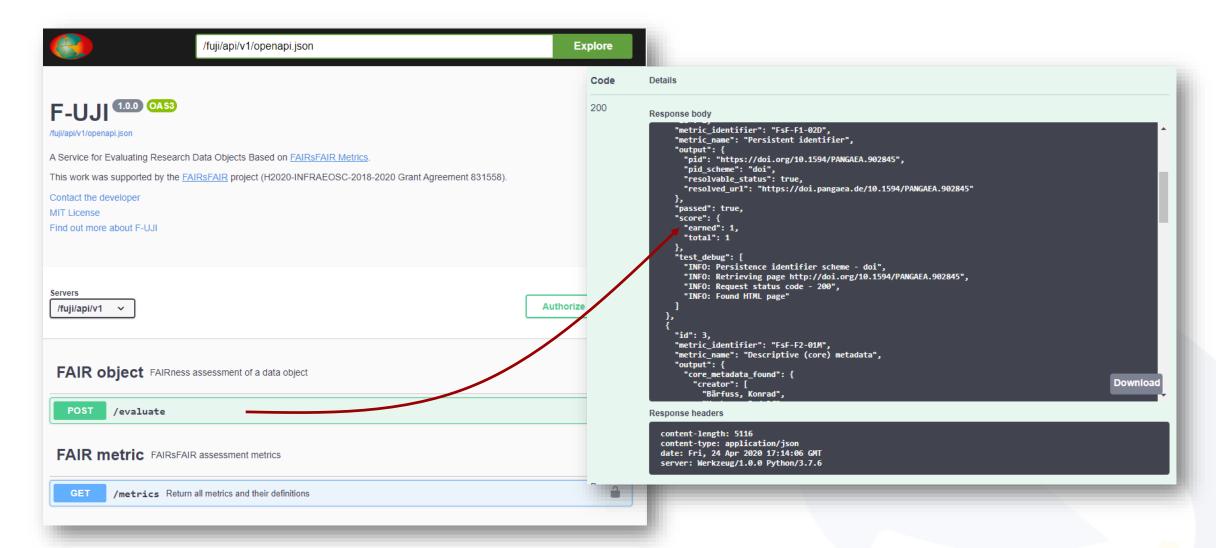


<u>Demo</u>

https://doi.org/10.1594/PANGAEA.908011



F-UJI – An Automated FAIR Data Assessment Tool





Request and Response

```
"object_identifier": "https://doi.org/10.1594/PANGAEA.908011",
    "test_debug": true,
    "oaipmh_endpoint" :"http://ws.pangaea.de/oai/provider"
}
```

```
"test debug": [
 "INFO: Formats of structured metadata embedded in HTML markup dict keys(['json-ld', 'opengra
  "INFO: Extract metadata from Schema.org JSON-LD (Embedded)",
  "INFO: Extract DublinCore metadata from html page",
 "INFO: Retrieving page https://doi.pangaea.de/10.1594/PANGAEA.908011",
  "INFO: Content negotiation accept=application/xml, text/xml;q=0.5, status=406",
 "INFO: Extract metadata from Generic XML (Negotiated)",
  "INFO: Expected XML but content negotiation responded: html",
  "INFO: Extract datacite metadata",
 "INFO: Retrieving page <a href="http://doi.org/10.1594/PANGAEA.908011"">http://doi.org/10.1594/PANGAEA.908011</a>",
  "INFO: Content negotiation accept=application/vnd.datacite.datacite+json, status=200",
 "INFO: Extract metadata from Linked Data (RDF)",
  "INFO: Retrieving page https://doi.pangaea.de/10.1594/PANGAEA.908011",
  "INFO: Content negotiation accept=text/turtle, application/turtle, application/x-turtle;q=0
 text/rdf+n3;q=0.9, status=406",
  "INFO: Expected RDF Graph but received - text/html",
 "INFO: Linked Data metadata UNAVAILABLE",
 "INFO: Type of object described by the metadata - Dataset",
  "INFO: Required core metadata ['creator', 'title', 'publisher', 'publication date', 'summar'
"test status": "pass"
```

```
"id": 3,
"metric identifier": "FsF-F2-01M",
"metric name": "Metadata includes descriptive core elements (creator, titl
keywords) to support data findability.",
"output": {
 "core metadata found": {
    "creator": [
      "Robert Huber"
    "keywords": "Event label, Neogloboquadrina pachyderma sinistral, maxim
   investigations, Kasten corer, CTD/Rosette, Scanning electron microscop
   ARK-II/5, ARK-VI/2, ARK-V/2, ARK-VII/1, ARK-VIII/3, SO82, Meteor (1986)
   Karbonatflüsse seit dem Pliozän: Rekonstruktion polarer und atlantisch
    "object identifier": "https://doi.org/10.1594/PANGAEA.908011",
   "publication date": "2019-11-01",
   "publisher": "PANGAEA",
   "summary": "This data set contains unpublished measurements of the max
   Neogloboquadrina pachyderma sin. carried out on surface sediment sampl
   "title": "Maximum diameter of Neogloboquadrina pachyderma sinistral fr
  "core metadata source": [
   "Schema.org JSON-LD (Embedded)",
    "Embedded DublinCore",
    "Datacite Search"
  "core metadata status": "all metadata"
"score": {
 "earned": 2,
 "total": 2
 test debug": [
  "INFO: Formats of structured metadata embedded in HTML markup dict keys(
  "INFO: Extract metadata from Schema.org JSON-LD (Embedded)",
```

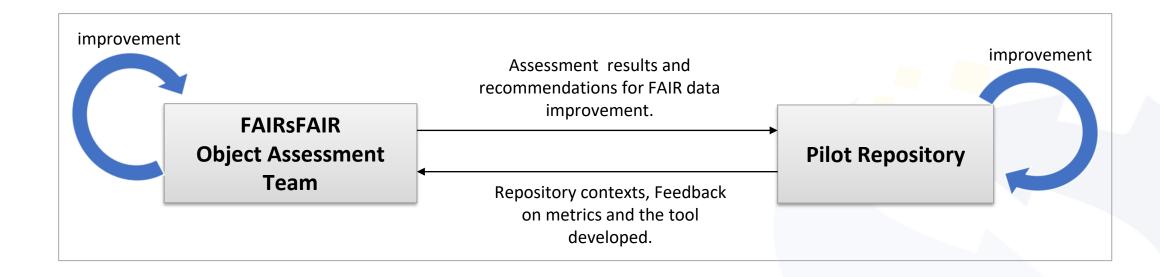


- Task 4.5 (FAIR Data Assessments: Pilots)
- FAIR Assessment of Research Data
 - Data Assessment Metrics
 - F-UJI: An Automated FAIR Data Assessment Tool
- Pilot Repositories & Results
- Conclusions



Progress Towards the End Goal

- KPI 4.2: Metrics and badging scheme for assessment of FAIRness of individual datasets in trusted repositories tested and applied to 100 datasets in 5 CoreTrustSeal certified repositories.
- Automated assessment + consultancy with pilot repositories.





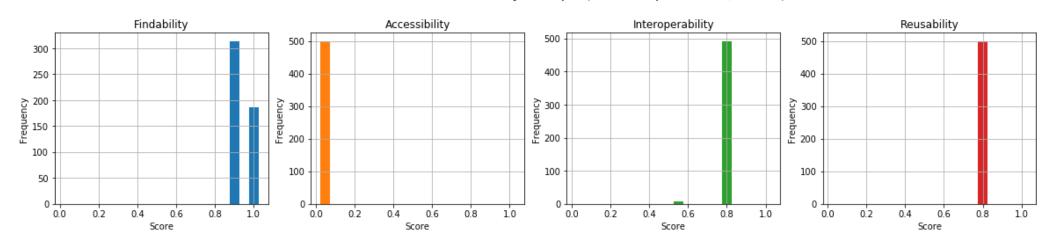
Collaborating Repositories

Repository	Certification	Subject Areas	Datasets Evaluated (as of 25.09.2020)	FAIR Data Improvement	Repository Contact
<u>PANGAEA</u>	CoreTrustSeal, WDS Regular Member	Earth and Environmental Science	500	Completed (1 st iteration)	Uwe Schindler Michael Diepenbroek
Phaidra-Italy	CoreTrustSeal	Cultural Heritage	500	Ongoing	Yuri Carrer Cristiana Bettella GianLuca Drago Giulio Turetta
CSIRO Data Portal	CoreTrustSeal	Multiple disciplines	500	Ongoing	Mikaela Lawrence Dom <mark>in</mark> ic Ho <mark>gan</mark> Cynthia Love
World Data Centre for Climate (WDCC)	CoreTrustSeal, WDS Regular Member	Earth System Science	500	Ongoing	Amandine Kaiser Andrej Fast Hannes Thiemann
<u>DataverseNO</u>	CoreTrustSeal	Multiple disciplines	500	Ongoing	Philipp Conzett (UiT/DataverseNO) Gustavo Durand, Julian Gautier (Harvard/Dataverse)

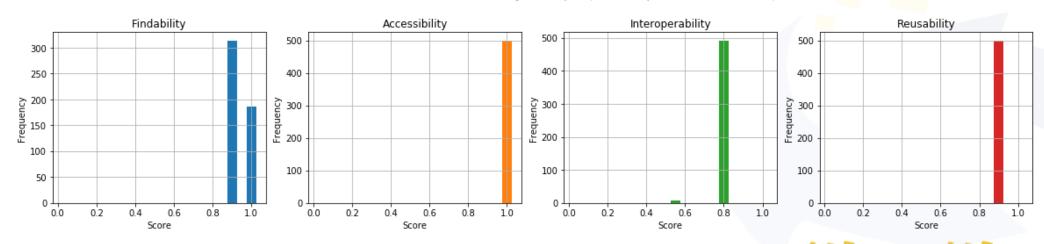


Before and After

FAIR Scores of PANGAEA Datasets By Principle (Before Improvement, n=500)

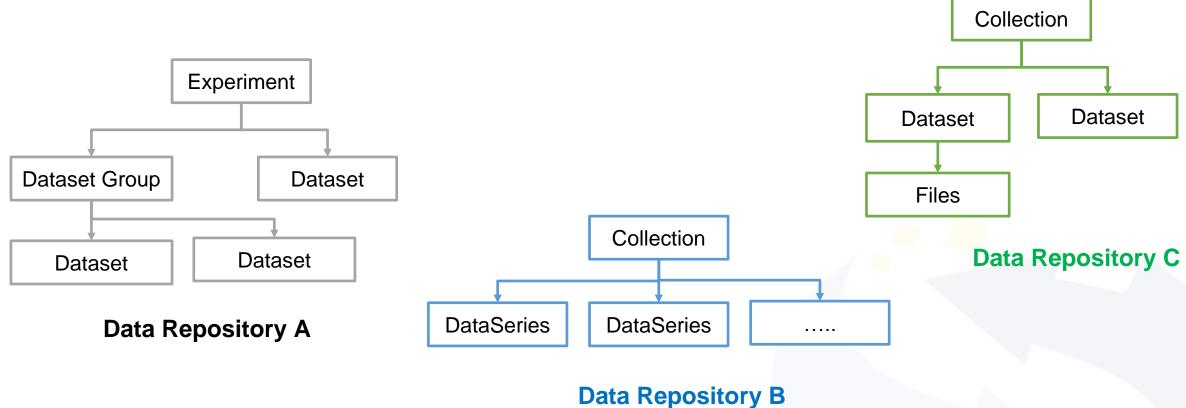


FAIR Scores of PANGAEA Datasets By Principle (After Improvement, n=500)



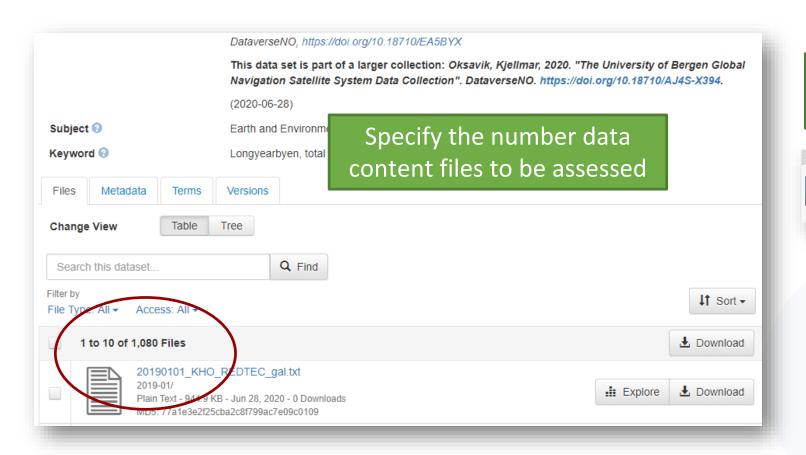


Identify data objects to be assessed





Performance matters!



Cache external resources
(selected) locally.

TLINUX FOUNDATION COLLABORATIVE PROJECTS

SPDX

Metadata Standards Catalog

Linked Open Data Cloud

Linked Open Vocabularies (LOV)

医克尔德氏反射 医克尔德氏反射 医克尔德氏虫虫



Restricted data can be FAIR

```
"id": 14.
                                                     "metric identifier": "FsF-R1.3-02D",
                                                     "metric name": "Data is available in a file format recommended by the target re
                                                      community.",
"id": 4.
                                                     "output": [],
"metric identifier": "FsF-F3-01M",
                                                     "score": { 📟 }.
"metric_name": "Metadata includes the identifier of "test_debug": [
"output": { ( ),
                                                       "INFO: File format specified - text/html",
                                                       "INFO: File format specified - text/tab-separated-values",
"score": { ( ),
                                                       "WARNING: Could not perform file format checks as data content identifier(s)
"test debug": [
                                                        unavailable/inaccesible"
  "INFO: Object identifier specified https://doi.orj.
  "INFO: Number of object content identifier found - 2",
  "WARNING: Content identifier https://doi.pangasa.de/10.1594/PANGAEA.570754?format=html
    inaccessible, HTTPError code 401",
  "WARNING: Content identifier https://doi.pangaea.de/10.1594/PANGAEA.570754?format
    =textfile inaccessible, HTTPError code 401"
```



- Extend ways of retrieving metadata
 - Metadata Aggregators
 - PID provider (datacite) √
 - Other potential aggregators? B2FIND?
 - HTML-embedded Data
 - JSON-LD (schema.org) embedded in a <script> tag√
 - JSON-LD (schema.org) dynamically ingested into the page through JavaScript code
 - Microdata √
 - RDFa ✓
 - Typed Links
 - HEAD/GET request
 - HTML link element √
 - Harvesting protocol supported by a repository
 - OAI-PMH√



- Task 4.5 (FAIR Data Assessments: Pilots)
- FAIR Assessment of Research Data
 - Data Assessment Metrics
 - F-UJI: An Automated FAIR Data Assessment Tool
- Pilot Repositories & Results
- Conclusions



Conclusions

- Our approach → use-case driven, iterative, reuse & adapt.
- Continuous improvement is essential. We need more pilot repositories;)
- What's Next?
 - Badging implementation
 - Continue liaising with other projects/institutions (EOSC Synergy, EOSC-Nordic, DataverseNL, GOFAIR, ...) with regard to FAIR object assessment.



To remember about measuring FAIR data...

- Metrics for research data focus on generally applicable data/metadata characteristics until domain/community-driven criteria have been agreed.
- F. A. I. R. are not new to data repositories!
- Must take into account context (e.g., disciplinary practices, data structures, types) and data infrastructure – human should be in the loop!
- Automated assessment saves efforts, but not all components of the research data ecosystem are machine-friendly; some aspects (rich, plurality, accurate, relevant) specified in FAIR principles still require human mediation and interpretation.
- FAIR assessment must go beyond the object itself. FAIR enabling (trustworthy)
 for repositories/services evolves in parallel.



Task 4.5:

Anusuriya Devaraju, Robert Huber, Mustapha Mokrane, Jerry de Vries, Patricia Herterich, Linas Cepinkas, Vesa Akerman, Joy Davidson, Herve L'Hours.

- FAIRsFAIR Data Object Assessment Metrics Specification v0.3
- **FAIR-Aware**





FAIR-Aware

Let's assume you have research data almost ready for uploading to a repository: do you already know how you and the repository can work together to make the data as findable, accessible, interoperable and reusable (FAIR) as possible? By guiding you through the assessment process, the FAIR-Aware tool can help you to better understand the FAIR Principles and how making data FAIR can increase the potential value and impact of your data.

FAIR-Aware is an online tool developed by the FAIRsFAIR project. The tool is not meant to give you a score for the FAIRness of a specific dataset. You should, however, have a target dataset in mind to be able to answer the questions and complete the assessment.

The assessment starts with a few questions 'about you' followed by 10 questions about FAIR. After you answer each question additional information and guidance will be displayed. The majority of the questions will help you assess your current level of awareness about what actions are needed to make data FAIR. At the end, Your feedback will help us improve FAIR-Aware and make it as user-friendly as possible. You will need between 10 and 30 minutes to complete the assessment depending on your familiarity with the subject and issues covered.

The FAIRsFAIR Team (DANS, DCC, UniHB)

Find out more about FAIRsFAIR on the project's website . If you have any questions, drop us an e-mail.

Α	bo	ut	yo	u
, ,	\sim	uс	y U	ч

work in?	Domain▼
Which of the following describes your role? Please select all that apply.	☐ Researcher ☐ Policy maker ☐ Research support (e.g. data steward, curator, data manager, librarian, information technology professional)
Which of the following types of organisations best describe your employer? Please select all that	Research Infrastructure/eInfrastructure (e.g. data repository, service provider,

library)

☐ University or Research Performing Organisation	

☐ Research Pe	erforming	Organisat	ı
☐ Government			

elnfrastructure (e.g. repository
scientific data provider)

FAIR questions

FΙ			

Are you aware that a dataset should be assigned a globally unique and persistent identifier when deposited with a data repository?	O Ye
Are you aware that when you deposit a dataset with a repository, you will need to provide some details (known as discovery metadata) in order to make the data findable, understandable and reusable to	O Ye



☐ Funder ☐ Publisher ☐ Other

☐ Funding Body ☐ Publisher

☐ Industry

☐ Other





It's quiz time;)

Go to www.menti.com and use the code 8888200



source: cruzysuzy