FAIRSFAIR
Fostering Fair Data Practices in Europe

# An Automated Assessment of the FAIRness of Research Data

Anusuriya Devaraju and Robert Huber
PANGAEA, MARUM, University of Bremen.
{adevaraju, rhuber}@marum.de

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

# Presentation Outline

- FAIR Data Principles
- Fostering FAIR Data Practices in Europe (FAIRsFAIR)
- FAIR Assessment of Research Data
  - Approach
  - Object Assessment Metrics
  - F-UJI : An Automated FAIR Data Assessment Tool
- Conclusions

# FAIR Principles

A set of high-level guidelines to maximize the reuse digital resources.

For more information on the principles, see
- https://www.force11.org/group/fairgroup/fairprinciples
- Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016), doi:10.1038/sdata.2016.18

## FINDABLE:

F1. (meta)data are assigned a globally unique and eternally persistent identifier.

F2. data are described with rich metadata.

F3. metadata specify the data identifier.

F4. (meta)data are registered or indexed in a searchable resource.

## 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.

# The Principles…

- Not strictly defined, led to a wide range of interpretations of FAIRness.
- More is not always better…

?

How can we measure the FAIR aspects (aka. FAIRness) of data systematically in practice?



Devaraju, A. and Huber, R. (2020). An Automated Assessment of the FAIRness of Research Data, Geospatial Sensing 2020.

# Translating FAIR into Practice…

<u>Priority Recommendations</u>

Rec. 8: Facilitate automated processing

Rec. 12: Develop metrics for FAIR Digital Objects

<u>Supporting Recommendations</u>

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

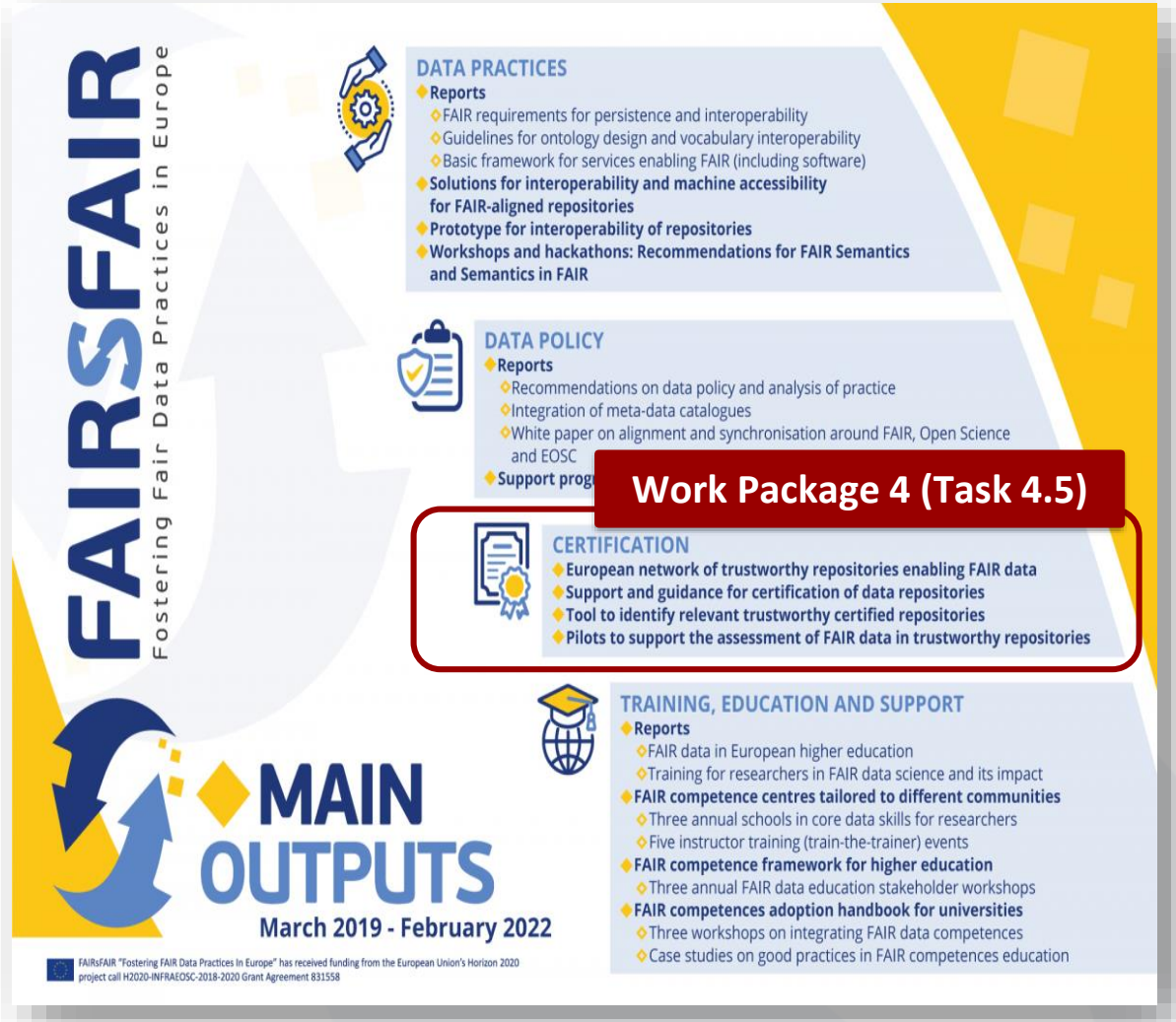
# Presentation Outline

- FAIR Data Principles
- Fostering FAIR Data Practices in Europe (FAIRsFAIR)
- FAIR Assessment of Research Data
  - Approach
  - Object Assessment Metrics
  - F-UJI : An Automated FAIR Data Assessment Tool
- Conclusions

# FAIRsFAIR

https://www.fairsfair.eu

- Aims to supply practical solutions for the use of the FAIR principles throughout the research data life cycle.
- Budget: €10 million
- 22 partners from 8 member states.






Devaraju, A. and Huber, R. (2020). An Automated Assessment of the FAIRness of Research Data, Geospatial Sensing 2020.

# Presentation Outline

- FAIR Data Principles
- Fostering FAIR Data Practices in Europe (FAIRsFAIR)
- FAIR Assessment of Research Data
  - Approach
  - Object Assessment Metrics
  - F-UJI : An Automated FAIR Data Assessment Tool
- Conclusions

# Scope

- The goal of the task is to pilot the FAIR assessment of published research data from five trustworthy data repositories.
- Research data, i.e., (digital) information that has been collected, observed, created to validate research findings.*

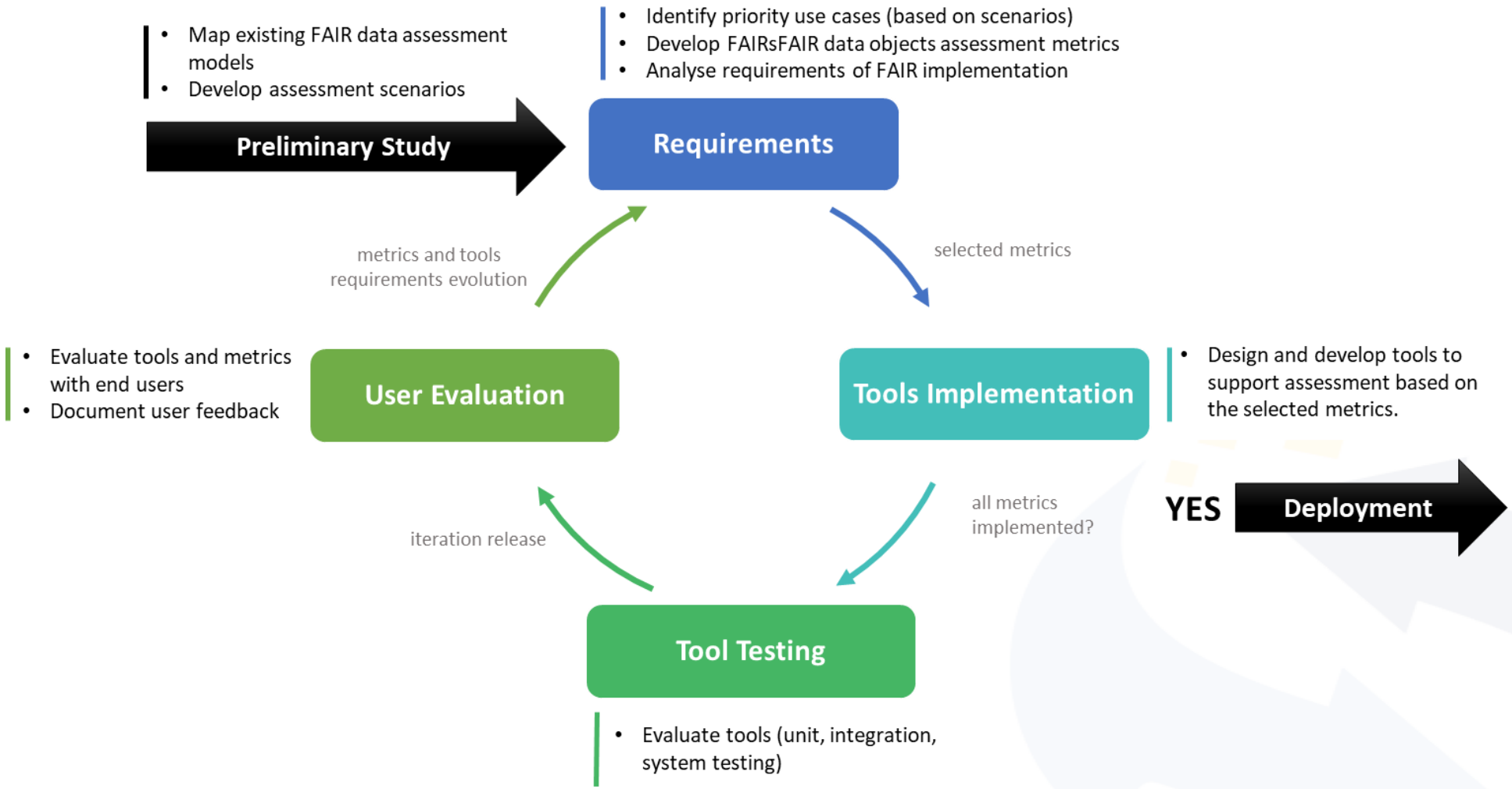
- FAIR assessment implementation comprises the development of two main components - metrics and tool

*Modified from https://library.leeds.ac.uk/info/14062/research_data_management/61/research_data_management_explained

# Presentation Outline

- FAIR Data Principles
- Fostering FAIR Data Practices in Europe (FAIRsFAIR)
- FAIR Assessment of Research Data
  - Approach
  - Object Assessment Metrics
  - F-UJI : An Automated FAIR Data Assessment Tool
- Conclusions

# Iterative Development






Devaraju, A. and Huber, R. (2020). An Automated Assessment of the FAIRness of Research Data, Geospatial Sensing 2020.

# Collaborating Repositories

| Repository | Certification | Subject Areas | URL |
|---|---|---|---|
| PANGAEA | CoreTrustSeal, WDS Regular Member | Earth and Environmental Science | https://www.pangaea.de/ |
| Phaidra-Italy | CoreTrustSeal | Cultural Heritage | https://phaidra.cab.unipd.it/ |
| CSIRO Data Portal | CoreTrustSeal | Multiple disciplines | https://data.csiro.au |
| DKRZ - World Data Centre for Climate (WDCC) | CoreTrustSeal, WDS Regular Member | Climate | https://cera-www.dkrz.de/ |
| Tbd. | | | |


Devaraju, A. and Huber, R. (2020). An Automated Assessment of the FAIRness of Research Data, Geospatial Sensing 2020.

# Presentation Outline

- FAIR Data Principles
- Fostering FAIR Data Practices in Europe (FAIRsFAIR)
- FAIR Assessment of Research Data
  - Approach
  - Object Assessment Metrics
  - F-UJI : An Automated FAIR Data Assessment Tool
- Conclusions

FAIRSFAIR
Fostering Fair Data Practices in Europe

About | FAIR Support | FAIR Landscape | Tools & Software | Events | Project Outputs | Outreach

[1]*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/fairsfair-data-object-assessment-metrics-request-comments



Devaraju, A. and Huber, R. (2020). An Automated Assessment of the FAIRness of Research Data, Geospatial Sensing 2020.

# Object Assessment Metrics

- There are 15 core metrics (v0.3) developed based on:
  - RDA FAIR Data Maturity Model
  - Fairdat/FAIR Enough?
  - WDS/RDA Assessment of Data Fitness
- Correspond to all or part of one or more FAIR principles.
- The specification follows the template modified from (Wilkinson et al., 2018).

| 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).<br>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)<br>• 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.<br>• Structured data embedded in the landing page of the identifier (e.g., Schema.org, Dublin Core and OpenGraph meta tags)<br>• Typed Links in the HTTP Link header; for more information, see https://signposting.org/conventions/<br>• 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** | |

Related Resources, Constraints and Limitations



Devaraju, A. and Huber, R. (2020). An Automated Assessment of the FAIRness of Research Data, Geospatial Sensing 2020.

# Example

**COMMENTS**

**Related Resources**

- Examples of metadata recommendations:
  - EOSC EDMI metadata properties, https://eosc-edmi.github.io/properties
  - W3C Recommendation Data on the Web Best Practices, https://www.w3.org/TR/dwbp/#metadata
  - W3C Data Catalog Vocabulary, https://www.w3.org/TR/vocab-dcat-2/
- Sites that provide a list of metadata standards:
  - FAIRsharing standards, https://fairsharing.org/standards/
  - DCC List of Metadata Standards, http://www.dcc.ac.uk/resources/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 data:
  - 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
  - Schema.org, https://schema.org/Dataset
  - 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.

**FsF-F2-01M**
**Metadata includes descriptive core elements (creator, title, data identifier, publisher, publication date, summary and keywords) to support data findability.**

Related Resources, Limitations and Constraints

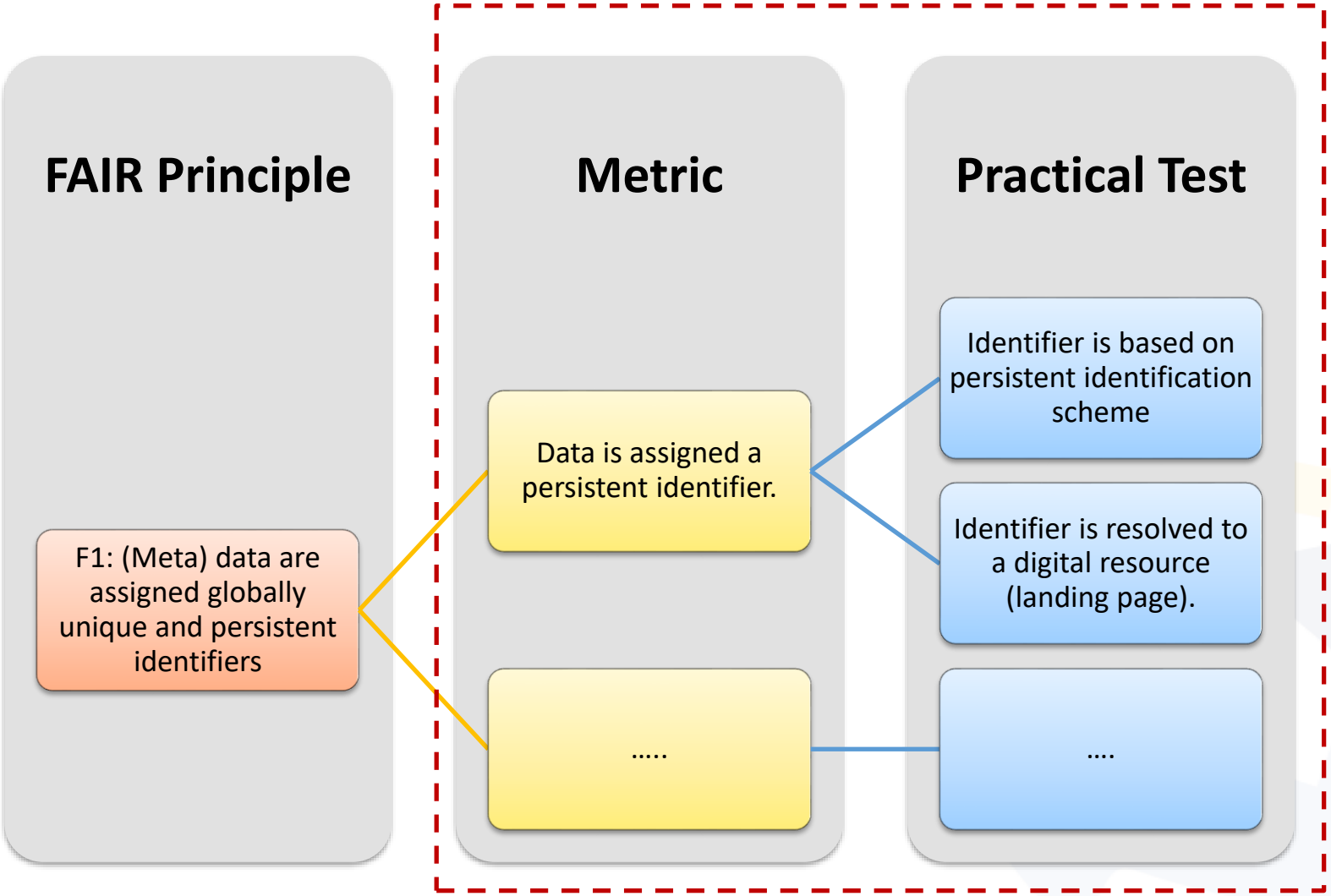


Devaraju, A. and Huber, R. (2020). An Automated Assessment of the FAIRness of Research Data, Geospatial Sensing 2020.

# Presentation Outline

- FAIR Data Principles
- Fostering FAIR Data Practices in Europe (FAIRsFAIR)
- FAIR Assessment of Research Data
  - Approach
  - Object Assessment Metrics
  - F-UJI : An Automated FAIR Data Assessment Tool
- Conclusions

# From Principles to Practical Tests






Devaraju, A. and Huber, R. (2020). An Automated Assessment of the FAIRness of Research Data, Geospatial Sensing 2020.

# F-UJI - Automated FAIR Data Assessment Tool

/fuji/api/v1/openapi.json Explore

## F-UJI 1.0.0 OAS3

/fuji/api/v1/openapi.json

A Service for Evaluating Research Data Objects Based on FAIRsFAIR Metrics.

This work was supported by the FAIRsFAIR project (H2020-INFRAEOSC-2018-2020 Grant Agreement 831558).

Contact the developer
MIT License
Find out more about F-UJI

Servers
/fuji/api/v1

Authorize

### FAIR object FAIRness assessment of a data object

POST /evaluate

### FAIR metric FAIRsFAIR assessment metrics

GET /metrics Return all metrics and their definitions

Code Details

200

Response body


```
    "metric_identifier": "FsF-F1-02D",
    "metric_name": "Persistent identifier",
    "output": {
      "pid": "https://doi.org/10.1594/PANGAEA.902845",
      "pid_scheme": "doi",
      "resolvable_status": true,
      "resolved_url": "https://doi.pangaea.de/10.1594/PANGAEA.902845"
    },
    "passed": true,
    "score": {
      "earned": 1,
      "total": 1
    },
    "test_debug": [
      "INFO: Persistence identifier scheme - doi",
      "INFO: Retrieving page http://doi.org/10.1594/PANGAEA.902845",
      "INFO: Request status code - 200",
      "INFO: Found HTML page"
    ]
  },
  {
    "id": 3,
    "metric_identifier": "FsF-F2-01M",
    "metric_name": "Descriptive (core) metadata",
    "output": {
      "core_metadata_found": {
        "creator": [
          "Bärfuss, Konrad",
```


Download

Response headers

```
content-length: 5116
content-type: application/json
date: Fri, 24 Apr 2020 17:14:06 GMT
server: Werkzeug/1.0.0 Python/3.7.6
```



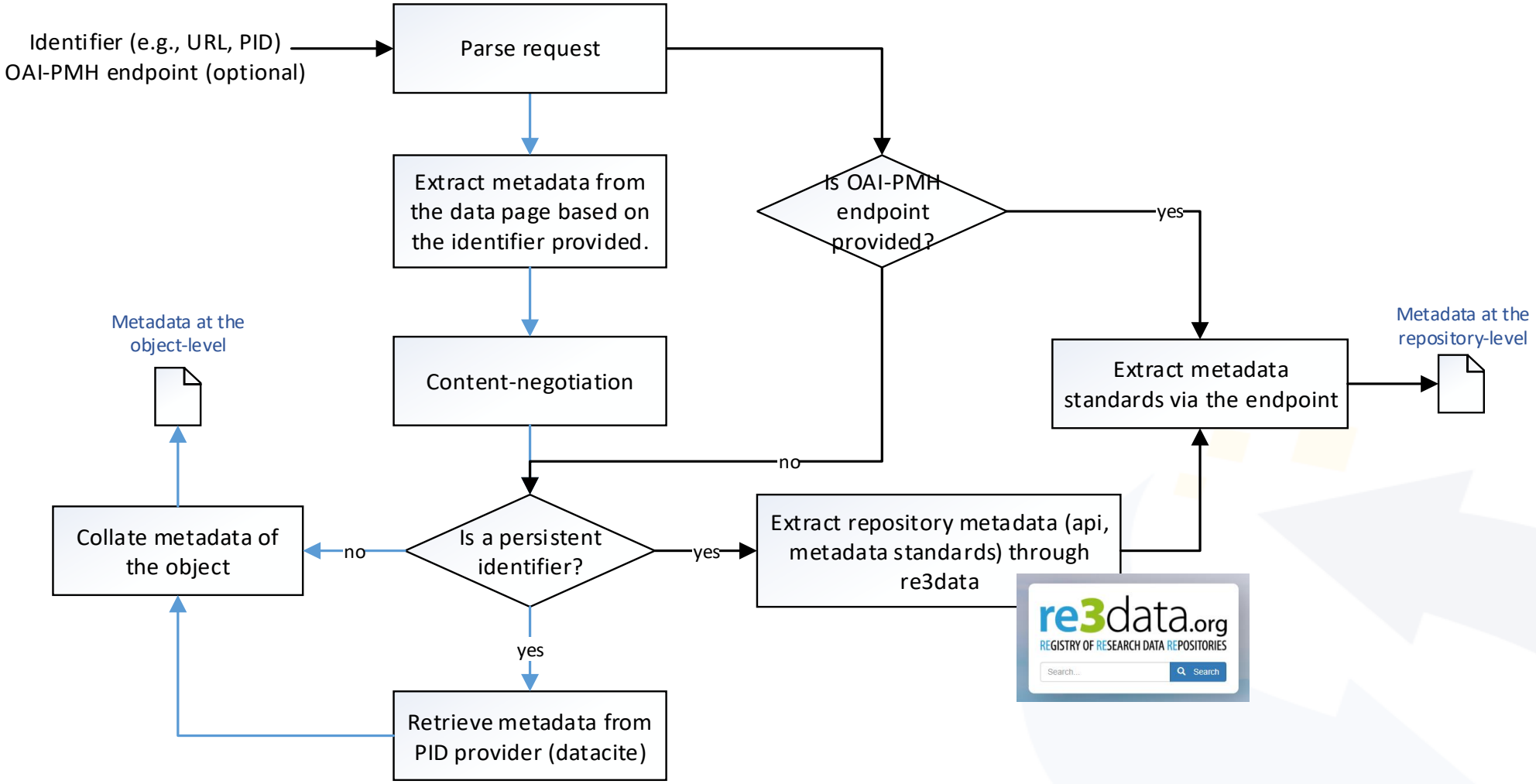
Devaraju, A. and Huber, R. (2020). An Automated Assessment of the FAIRness of Research Data, Geospatial Sensing 2020.

# Resources Used

- Metadata
  - Structured data within the HTML page
  - Content-negotiation
- Data file
- External resources
  - PID provider services (e.g., datacite)
  - r3data.org (repository APIs and metadata standards)
  - SPDX license list
  - RDA Metadata Standards Catalog
  - Linked Open Vocabularies (LOV)
  - ISO/TR 22299 (Digital file format recommendations for long-term storage)
  - Wolfram scientific formats
  - more ….

Devaraju, A. and Huber, R. (2020). An Automated Assessment of the FAIRness of Research Data, Geospatial Sensing 2020.

# Assessment Overview (Data Gathering)






Devaraju, A. and Huber, R. (2020). An Automated Assessment of the FAIRness of Research Data, Geospatial Sensing 2020.

# Request and Response


```
{
    "object_identifier": "https://doi.org/10.1594/PANGAEA.908011",
    "test_debug": true
}
```



```
},
"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 http://doi.org/10.1594/PANGAEA.908011",
  "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', 'summary
],
"test_status": "pass"
```

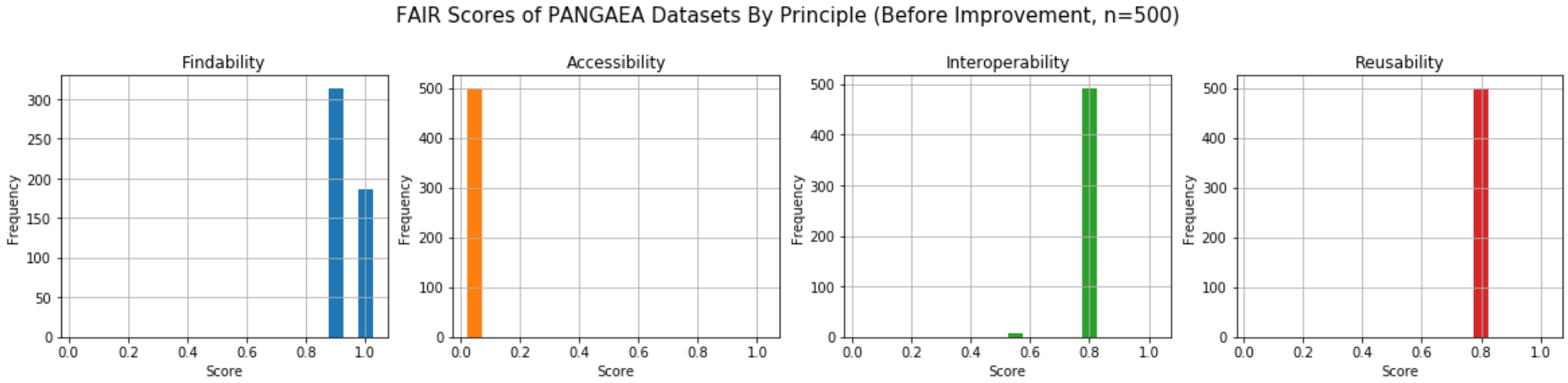


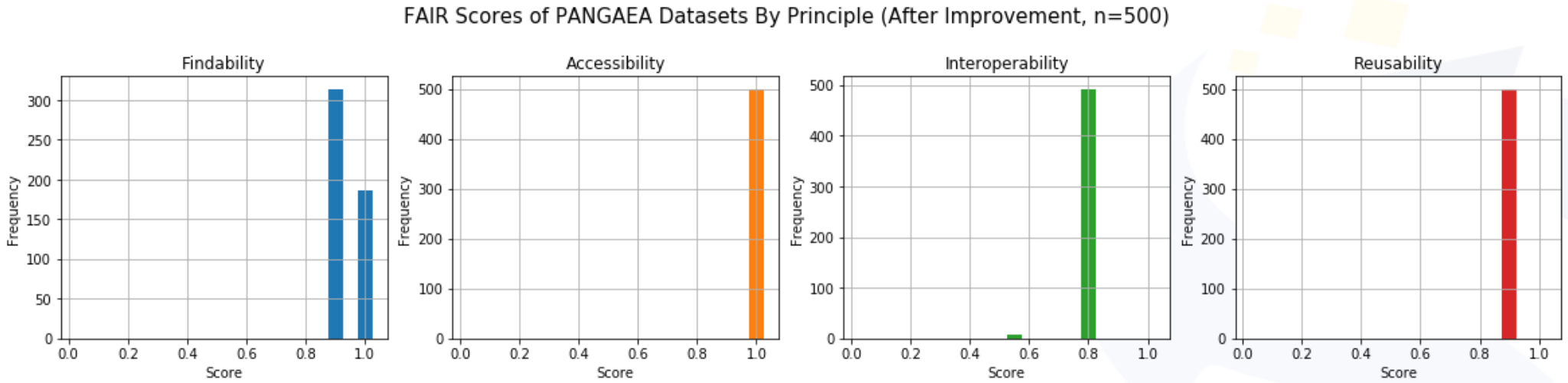
```
{
  "id": 3,
  "metric_identifier": "FsF-F2-01M",
  "metric_name": "Metadata includes descriptive core elements (creator, titl
  keywords) to support data findability.",
  "output": {
    "core_metadata_found": {
      "creator": [
        "Robert Huber"
      ],
      "keywords": "Event label, Neogloboquadrina 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)",
```




Devaraju, A. and Huber, R. (2020). An Automated Assessment of the FAIRness of Research Data, Geospatial Sensing 2020.

# Before and After





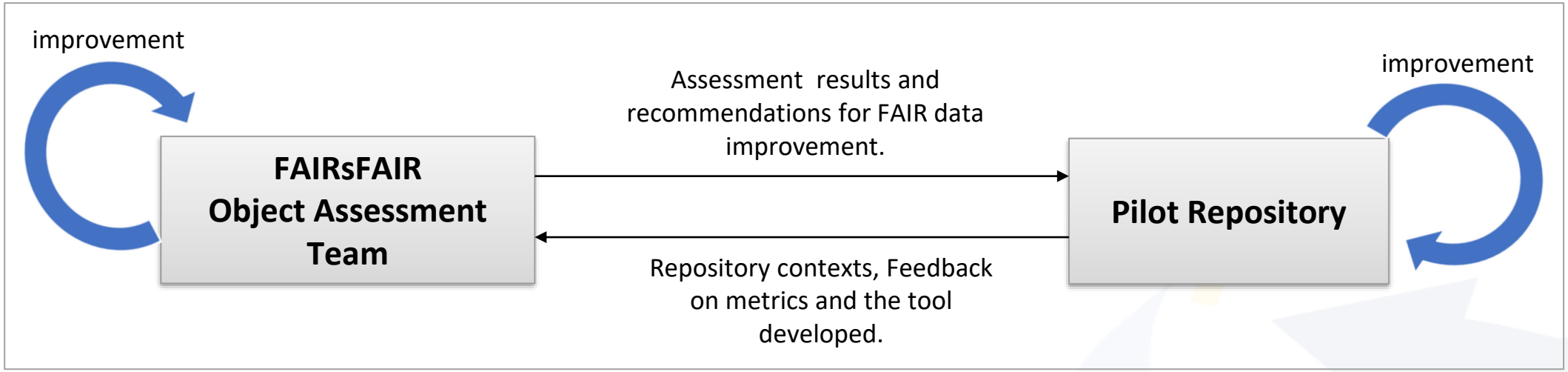

Devaraju, A. and Huber, R. (2020). An Automated Assessment of the FAIRness of Research Data, Geospatial Sensing 2020.

# Presentation Outline

- FAIR Data Principles
- Fostering FAIR Data Practices in Europe (FAIRsFAIR)
- FAIR Assessment of Research Data
  - Approach
  - Object Assessment Metrics
  - F-UJI : An Automated FAIR Data Assessment Tool
- Conclusions

# Conclusions

- Our approach → use-case driven, iterative, reuse & adapt.
- Continuous improvement matters!





Devaraju, A. and Huber, R. (2020). An Automated Assessment of the FAIRness of Research Data, Geospatial Sensing 2020.

# 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 types) and data infrastructure.
- 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 for repositories/services evolves in parallel.



Devaraju, A. and Huber, R. (2020). An Automated Assessment of the FAIRness of Research Data, Geospatial Sensing 2020.

# Related Links

- FAIRsFAIR Metrics Specification v0.3, https://doi.org/10.5281/10.5281/zenodo.3934401
- F-UJI, https://github.com/pangaea-data-publisher/fuji
- FAIR-Aware, https://fairaware.dans.knaw.nl/

Thank You

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.

About you

Which research domain do you 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)
- Funder
- Publisher
- Other

Which of the following types of organisations best describe your employer? Please select all that apply.
- Research Infrastructure/eInfrastructure (e.g. data repository, service provider, library)
- University or Research Performing Organisation
- Research Performing Organisation
- Government
- eInfrastructure (e.g. repository or scientific data provider)
- Funding Body
- Publisher
- Industry
- Other

FAIR questions

FINDABLE

Are you aware that a dataset should be assigned a globally unique and persistent identifier when deposited with a data repository? Yes / No

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 others? Yes / No



Devaraju, A. and Huber, R. (2020). An Automated Assessment of the FAIRness of Research Data, Geospatial Sensing 2020.