EOSC Task Force FAIR Metrics and Data Quality

EOSC Symposium

Slides by Carlo Lacagnina & Mark Wilkinson, Task Force co-Chairs
Presented by Mari Kleemola, Tampere University/FSD and CESSDA ERIC, TF member

EOSC Task Force “FAIR Metrics and Data Quality”

- A **multidisciplinary advisory group** of 26 experts in biology, metrology, climatology, data science and management, philosophy, computer sciences, etc. Experts come from 17 different European countries

- Kick-off in December 2021. Two co-chairs coordinate this EOSC TF: Mark Wilkinson and Carlo Lacagnina

- Bi-weekly meetings over two years in a mixed method approach including virtual discussions workshops organization and participation, use cases collection, and survey dissemination

Goals of this Task Force

- Explores issues related to the **governance of FAIR** evaluations
- Examine the problem of **inconsistency between** FAIR evaluation **tools**
- Evaluate the applicability and uptake of FAIR Metrics across research communities
- Undertake a state of the art to generate mutual understanding about data quality
- Conduct several case studies to identify common features and dimensions to **define a data quality approach for EOSC.**

EOSC Task Force FAIR Metrics and Data Quality

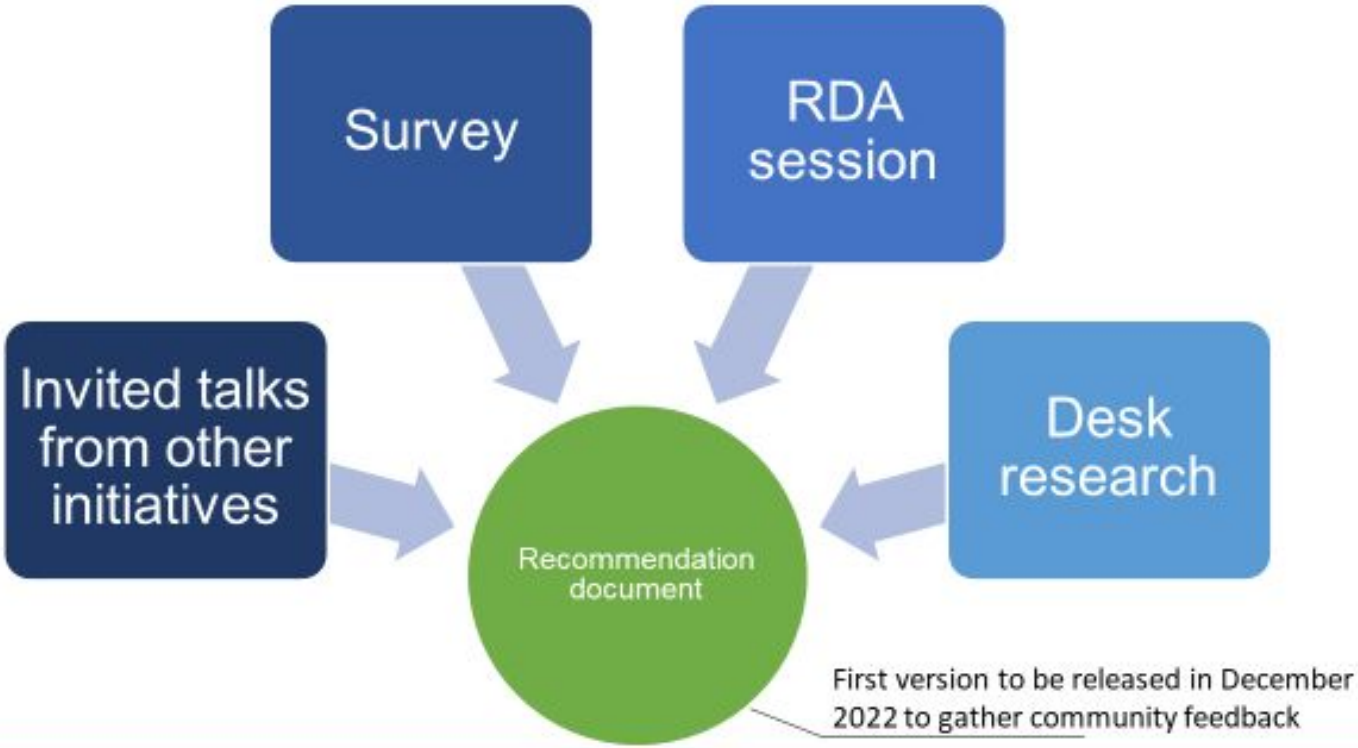
Data Quality group

Current status

Data Quality Group: What has been done so far

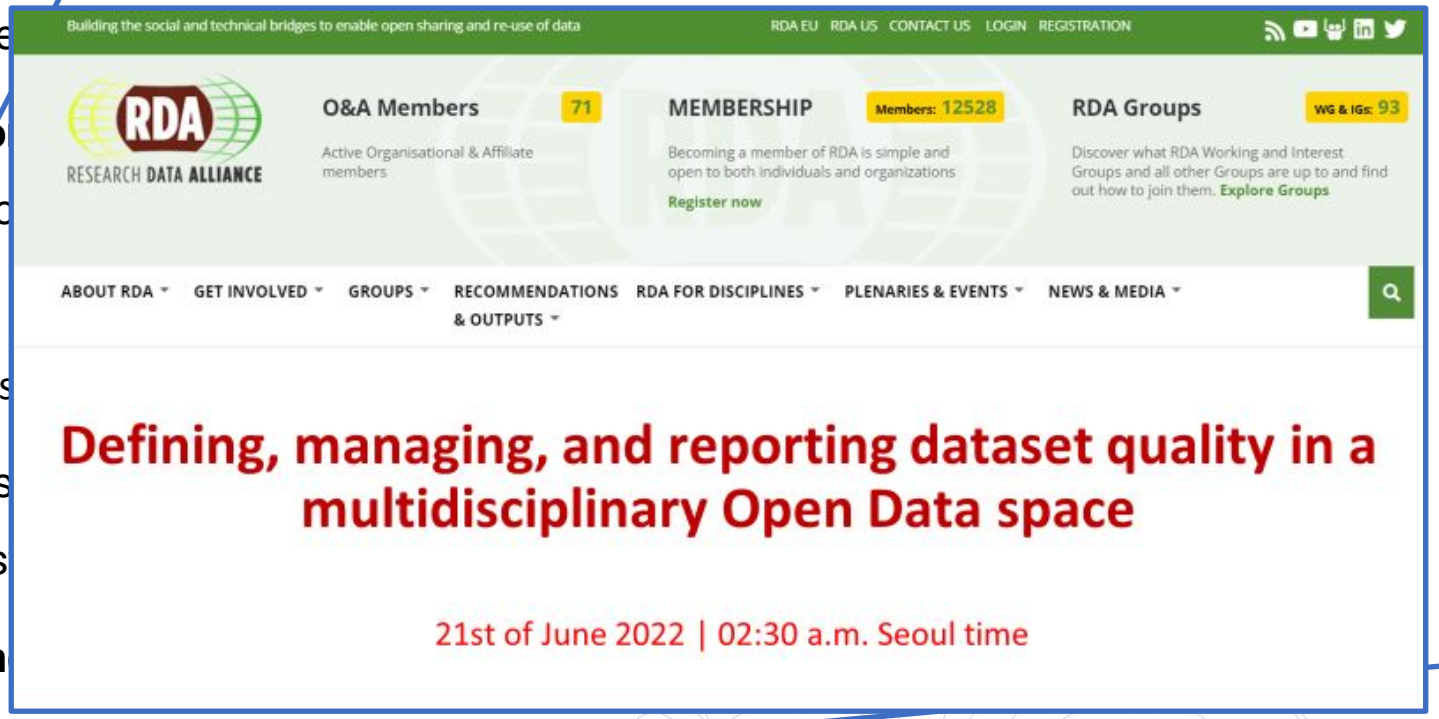
- Pinning down **common ground understanding** about quality approaches, what quality means, dataset lifecycle, actors involved, benefits of quality, workflow for managing quality, data types, certification, etc.
- **Desk research** of ISOs, literature, vocabulary
- Gathering inputs, lessons learned, agreed practices from **various initiatives** (e.g. RDA, INSPIRE, bioimaging, CoreTrustSeal, energy sector)
- Drafting a **recommendation document** – 1st version in December 2022
- **RDA session** organized in June
- Drafted a **survey** released in April: >700 views

Multidisciplinary understanding about data quality



Data Quality Group: What has been done so far

- Kick off, bi-weekly me
- Pinned down **commo** dataset lifecycle, acto certification, etc.
- **Desk research** of ISOs
- Gathering inputs, less bioimaging, CoreTrus
- Drafting a **recommen**
- **RDA session** organized in June
- Drafted a **survey** released in April: >700 views



eosc Data Quality Group: W

- Kick off, bi-weekly meetings and [agenda] set
- Pinned down **common ground understanding** a dataset lifecycle, actors involved, benefits of d certification, etc.
- **Desk research** of ISOs, literature, vocabulary
- Gathering inputs, lessons learned, agreed prac bioimaging, CoreTrustSeal, energy sector)
- Drafting a **recommendation document** – 1st v
- **RDA session** organized in June
- Drafted a **survey** released in April: >700 views

What information do you consider most important to properly use or select a dataset?

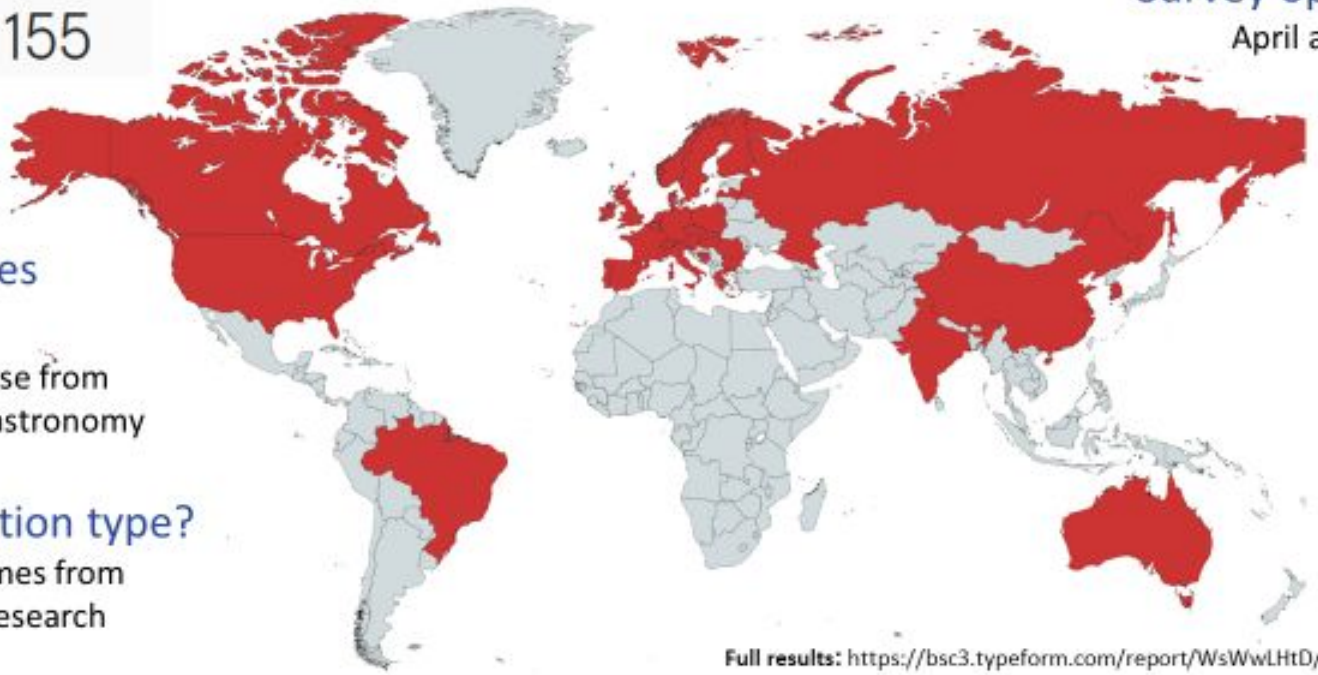
134 out of 134 people answered this question

	Mandatory	Very relevant	Somewhat relevant	I don't know
User guide (including a description ...	49.6%	42.9%	7.5%	0%
Scientifically accurate (e.g. validated...	40.2%	45.5%	13.6%	0.8%
License of use, including terms of use	60.4%	29.1%	10.4%	0%
Version	36.1%	36.8%	23.3%	3.8%
Data dictionary	19.5%	36.1%	34.6%	9.8%
Clarity about how to cite the dataset...	46.3%	35.8%	17.2%	0.7%
Archiving policy	15.8%	34.6%	45.9%	3.8%
Compliance				

eosc Survey: respondents

Views	Starts	Submissions
778	418	155

Survey open during
April and May 2022



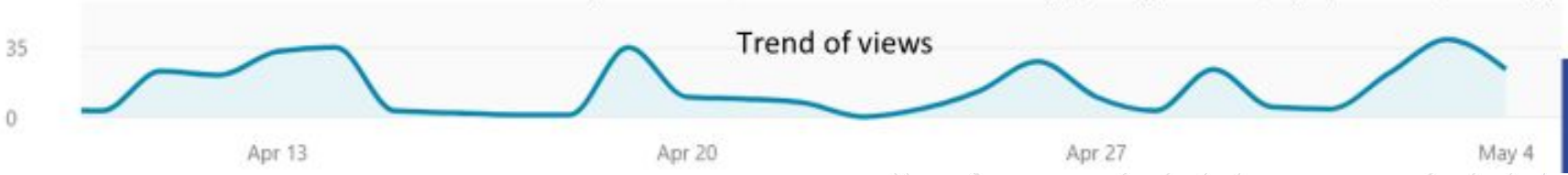
Which communities participated?

All but law, little response from agriculture, chemistry, astronomy

Organization type?

All, 70% comes from academia/research

Full results: https://bsc3.typeform.com/report/WsWwLHtD/ZJe8Zib4pqMz01Hb



eosc Survey: some insights

- Biggest concern/barrier to provide quality assessed data:

- Which practices should a discipline have to **gauge** its **maturity in quality management**?

Metadata standards, agreed definitions, standard quality management framework, metrics to quantify quality, quality assessments are operational routine and funded

- What level of data quality management do you expect from EOSC?

Basic curation: e.g., data content sanity checks, control availability of basic metadata or documentation, basic metadata compliance checks. Allow (re)users to rate or leave comments on data quality

- Some conclusions
  - It must be crystal clear and well advertised that **quality does not refer to data content quality only**, a.k.a. scientific quality. The survey demonstrated that several respondents see quality assessments as dangerous when done by external organizations like EOSC because the respondents see quality usually associated with the assessment of the data content.
  - Striking **preference for no ranking**. If a ranking has to be applied, then priority should be placed on **showing the FAIRness level** of the datasets. **No data content assessment** is expected from EOSC, but check of documentation availability for data understanding.
  - The future quality assessments should be shown first to the **data provider**, to give a chance to **improve the data**, and then to the users. The methodology has to be the same for similar datasets.
  - Create a catalogue of community tests/methods to apply in quality analyses.
  - EOSC users expect tools and services being designed according to a user-centric model.

Survey: some insights

Biggest concern/barrier to provide quality assessed data:

Which practices should a discipline have to gauge its maturity in quality management?

- Metadata standards, agreed definitions, standard quality management framework, metrics to quantify quality, quality assessments are operational routine and funded

What level of data quality management do you expect from EOSC?

- Basic curation: e.g., data content sanity checks, control availability of basic metadata or documentation, basic metadata compliance checks. Allow (re)users to rate or leave comments on data quality

- **Some conclusions**
  - It must be crystal clear and well advertised that **quality does not refer to data content quality only**, a.k.a. scientific quality. The survey demonstrated that several respondents see quality assessments as dangerous when done by external organizations like EOSC because the respondents see quality usually associated with the assessment of the data content.
  - Striking **preference for no ranking**. If a ranking has to be applied, then priority should be placed on **showing the FAIRness level** of the datasets. **No data content assessment** is expected from EOSC, but check of documentation availability for data understanding.
  - The future quality assessments should be shown first to the **data provider**, to give a chance to **improve the data**, and then to the users. The methodology has to be the same for similar datasets.
  - Create a catalogue of community tests/methods to apply in quality analyses.
  - EOSC users expect tools and services being designed according to a user-centric model.

Recommendation document

Recommendations are a set of principles and guidelines for both EOSC and the next TF:

- Datasets have to come with enough **contextualization** information to understand and correctly interpret them
- EOSC is not in charge of **data content** assessments
- Set clear **criteria** to prevent researchers concerns about how professionally their data will be managed, concerns are barriers to data sharing
- Develop a **pre-operational quality function** tailored to the EOSC stakeholders' requirements
- EOSC should support and push each community to agree on **community standards**, which form the basis for any quality assessment and FAIR sharing of research datasets
- We have already identified **minimum requirements**; the next TF will need to identify the exact standards forming the baselines for these requirements assessment

EOSC Task Force FAIR Metrics and Data Quality

FAIR Metrics group

Current status

FAIR Metrics Group: Three key objectives

- Explore issues related to the **governance** of FAIR evaluations
  - Who has the authority to decide what should be tested, how, and what is a successful result? There are (at least) 17 different FAIR evaluation systems, and nobody knows which one to trust
  - This is extremely problematic, when agencies and publishers are beginning to demand FAIRness
- Examine the problem of **inconsistency** between FAIR evaluation tools
  - Evaluators are generating dramatically different results
- Evaluate the applicability and **uptake** of FAIR Metrics (specifically RDA Maturity Indicators)
  - Ongoing… Measuring the effect that a well-governed and consistent FAIR assessment ecosystem will have on stakeholders' perceived trust in FAIRness evaluations, and their willingness to be evaluated using these tools.

FAIR Metrics Group: Three key objectives

- Explore issues related to the **governance** of FAIR evaluations
  - Who has the authority to decide what should be tested, how, and what is a successful result? There are (at least) 17 different FAIR evaluation systems, and nobody knows which one to trust
  - This is extremely problematic, when agencies and publishers are beginning to demand FAIRness
- Examine the problem of **inconsiste**
  - Evaluators are generati
- Evaluate the applicability and **uptal**
  - Ongoing… Measuring
    ecosystem will have on
    their willingness to be

Outcomes:

Whitepaper on Governance *for peer review and to initiate a discussion around governance models for FAIR metrics and testing*

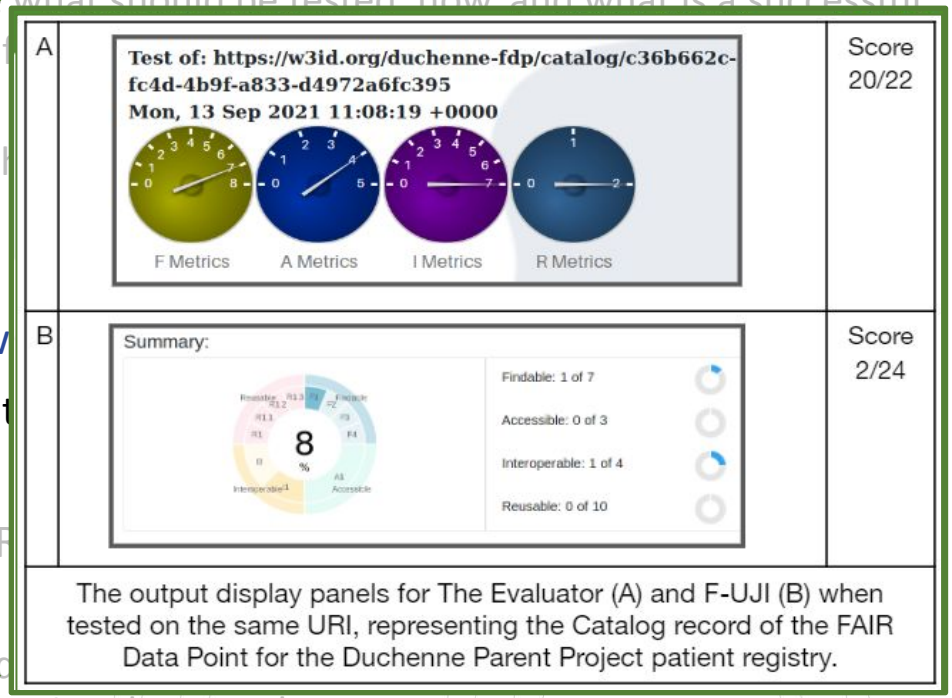
Objective: a self-sustaining, peer-reviewed mechanism for approving FAIR metrics and tests (including domain-specific!) that is ***trusted by the broad community*** *of stakeholders*

FAIR Metrics Group: Three key objectives

- Explore issues related to the **governance** of FAIR evaluations
  - Who has the authority to decide what should be tested, how, and what is a successful result? There are (at least) 17 different FAIR evaluation systems, and nobody knows which one to trust
  - This is extremely problematic, when agencies and publishers are beginning to demand FAIRness
- Examine the problem of **inconsistency** between FAIR evaluation tools
  - Evaluators are generating dramatically different results
- Evaluate the applicability and **uptake** of FAIR Metrics (specifically RDA Maturity Indicators)
  - Ongoing… Measuring the effect that a well-governed and consistent FAIR assessment ecosystem will have on stakeholders' perceived trust in FAIRness evaluations, and their willingness to be evaluated using these tools.

FAIR Metrics Group: Three key objectives

- Explore issues related to the **governance** of FAIR evaluations
  - Who has the authority to decide what should be tested, how, and what is a successful result? There are (at least) 17 dif… which one to trust
  - This is extremely problematic, wh… FAIRness
- Examine the problem of **inconsistency** betw…
  - Evaluators are generating dramat…
- Evaluate the applicability and **uptake** of FAIR…
  - Ongoing… Measuring the effect … ecosystem will have on stakehold… their willingness to be evaluated using these tools.



The output display panels for The Evaluator (A) and F-UJI (B) when tested on the same URI, representing the Catalog record of the FAIR Data Point for the Duchenne Parent Project patient registry.

Inconsistency between FAIR evaluation tools

Evaluator harmonization: find a common workflow

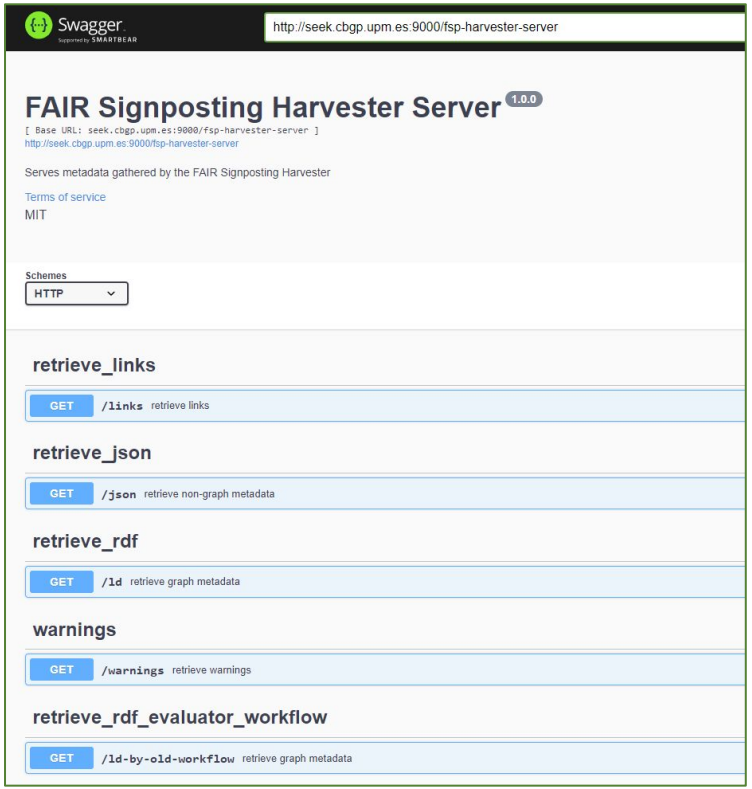
FAIR Signposting: a no-guesswork, unambiguous specification for pointing between a canonical identifier, the data it represents, and the metadata about that data

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.

Four TF-hosted Hackathons → specification and reference environment for checking that all evaluators are behaving identically when faced with a FAIR Signposting-compliant site

Inconsistency between FAIR evaluation tools

Evaluator harmonization: find a common workflow



A FAIR Signposting-compliant metadata harvesting engine has now been published @ UPM that can be used by all Evaluator systems.

FAIR Metrics Group: Three key objectives

- Explore issues related to the **governance** of FAIR evaluations
  - Who has the authority to decide what should be tested, how, and what is a successful result? There are (at least) 17 different FAIR evaluation systems, and nobody knows which one to trust
  - This is extremely problematic, when agencies and publishers are beginning to demand FAIRness
- Examine the problem of **inconsistency** between FAIR evaluation tools
  - Evaluators are generating dramatically different results
- Evaluate the applicability and **uptake** of FAIR Metrics (specifically RDA Maturity Indicators)
  - Ongoing… Measuring the effect that a well-governed and consistent FAIR assessment ecosystem will have on stakeholders' perceived trust in FAIRness evaluations, and their willingness to be evaluated using these tools.
  - https://ec.europa.eu/eusurvey/runner/EOSC-A_FAIR-Metrics-TF_Survey open until 2.12.

Thank you!