EOSC Task Force FAIR Metrics and Data Quality Charter

1. Main aims

The Task Force (TF) FAIR Metrics and Data Quality oversees the implementation of FAIR metrics for the European Open Science Cloud (EOSC), testing them with research communities to ensure they are fit for purpose. More general aspects of data quality are also considered so the data available within EOSC is robust and highly regarded by communities.

1.1 FAIR Metrics

For FAIR metrics, the focus is on implementation and testing, not on (re)definition of the metrics. The aim is to extensively test existing FAIR metrics for EOSC[1] and recently developed FAIR data assessment tools[2] in a variety of contexts with broad consultation, and make recommendations about the improvement and applicability of the existing metrics in different disciplines. Success of the FAIR metrics will be strongly dependent on uptake by communities. Providing feedback to the proposed criteria by the RDA FAIR Data Maturity Model Working Group and participating in discussions on possible updates of those criteria should be done through the RDA Maintenance Group.

The feedback will focus on checking the implementation of current FAIR metrics in terms of:

- established quantitative criteria, measurement tools, and measurement scales to assess how far datasets comply with the FAIR principles;
- clarity of the qualitative interpretations of the realized quantitative FAIR assessment;

[1] Recommendations for FAIR metrics in EOSC (https://doi.org/10.2777/70791) based on the criteria defined by the RDA FAIR Data Maturity Model Working Group. FAIR Data Maturity Model. Specification and Guidelines. (http://doi.org/10.15497/rda00050).

[2] such as F-UJI by the FAIRsFAIR project, the FAIR Evaluator the EOSC-Synergy project and more.

- effective understandability of communication to researchers and citizens about why the FAIR metric parameters are necessary and why these specific criteria are meaningful;
- Highlight the importance of FAIR awareness and constructive communication between data producers and service providers for effective evaluation.

1.2 Data Quality

As far as data quality is concerned, this topic has been less explored in EOSC. Quality attributes and dimensions (e.g. accuracy, completeness, conformity) vary strongly within and across disciplines. Exploring what are the most relevant quality dimensions considered in different communities is one of the goals of this TF. Since datasets are often crafted for specific designated communities with their specific requirements, assessing dataset quality is a multi-dimensional problem. Despite the multi-dimensionality of quality, there are aspects within the quality assessment process that are common across disciplines. For instance, identification of the quality characteristics and baseline, execution and dissemination are practices that are typical during any quality assessment process. The TF will work on specific case studies to give an overview about how different communities assess data quality and to identify features common across disciplines when working on data quality. This will result in preliminary recommendations for quality assessments that reflect the reliability and usability of EOSC data, while pointing to the most relevant opportunities (risks) deriving from improved (poor) data quality.

2. Core activities

This TF hopes to promote high-quality resources in EOSC by focussing on 1) FAIR metrics for EOSC and on 2) aspects of data quality. Both elements might require different activities given the different level of activity and uptake in EOSC.[3] Still, both elements require a tour of different communities in various disciplines. After being informed by the communities, the TF will provide recommendations on how different communities implement FAIR metrics and assess data quality, and what can be usefully taken from these approaches.

Key output of the TF is a report on the implementation of FAIR Metrics and Data Quality in the EOSC with ample opportunity to gather community inputs. Proposed timeline, that might be subject to change tailoring to community needs:

- Drawing in community inputs from ongoing activities and projects related to FAIR Metrics and Data Quality (month 01-month 09)
- EOSC TF FAIR Metrics and Data Quality review of community inputs (month 09-12)
- Initial paper on FAIR Metrics and Data Quality V1 for community feedback (month 12)
- FAIR Metrics and Data Quality report V2 for community feedback (month 18)
- FAIR Metrics and Data Quality report final version (month 24)

The outcomes of this TF are targeted to support a wide range of stakeholders in and around EOSC, though EOSC implementers might benefit most, for example those running EOSC services or defining rules of participation.

[3] For FAIR metrics in EOSC there is already a significant level of activity in this space (FAIR principles > FAIR indicators > FAIR metrics > tests against the metrics), while data quality has been less explored in the context of EOSC.

3. Planned duration

24 months

4. Working methodology

The TF will engage with stakeholders extensively. Use cases are a pragmatic way to develop cross-domain uptake. For both FAIR metrics and Data Quality, the TF will work with a use case driven approach. The use cases below have been proposed in the charter writing phase by the various contributors (see below). These use cases should not be seen as final, and the list might be changed and/or extended based on community inputs. The TF will focus on commonalities, lessons learned and challenges between use case studies as well. Workshops or other forms of interaction could also be used to solicit feedback and test implementation.

4.1 FAIR Metrics

While FAIR metrics are well defined, we still need to know whether they work in the various disciplines, how they are implemented and encouraged, or even enforced. Therefore, initial use cases have been identified:

- Test existing FAIR metrics and data assessment tools with community stakeholders from different disciplines. For example, the Graz University of Technology will evaluate and implement FAIR assessment tools in their institutional repository InvenioRDM, and DANS will test F-UJI to perform systematic assessments of the FAIRness of their data holdings. Such initiatives can share experiences/pilot projects on how infrastructures enable FAIR data and FAIR data management.
- In the field of Health, an Implementation Guide (IG) of the HL7 FHIR standard for the FAIRification of health data collections support the inclusion of metadata with sufficient richness to measure the FAIR data maturity level according to the RDA maturity model (cf. FHIR4FAIR). The IG proposes an evaluation methodology/checklist, exploring both machine-readable and manual evaluation methods.
- In the social sciences community a case study related to the standard of the Data Documentation Initiative (DDI) is pointed out. This standard, developed for documenting survey data (recently extended to other data types) is recommended by the EcoSco FAIR implementation Network.[4] Another one related to FAIR Implementation Profiles (FIP)[5] for data centers and/or repositories can be added: Completing FIPs allows for monitoring the state of implementation of metadata standards, generic technical elements of FAIR data infrastructures and connected to Wikidata resources, it draws a comprehensive map of convergence in the use of standards/ontologies and technologies within and across domains.

[4] Betancort Cabrera, N., Bongartz, E. C., Dörrenbächer, N., Goebel, J., Kaluza, H., & Siegers, P. (2020). *White Paper on implementing the FAIR principles for data in the Social, Behavioural, and Economic Sciences* (No. 274). RatSWD Working Paper. http://dx.doi.org/10.17620/02671.60

[5] Magagna, B., Schultes, E. A., Pergl, R., Hettne, K. M., Kuhn, T., & Suchánek, M. (2020, September 21). Reusable FAIR Implementation Profiles as Accelerators of FAIR Convergence. https://doi.org/10.31219/osf.io/2p85g

- FAIR terminology is the first initiative aiming at providing a paradigm for the optimal organization of terminological data compliant with the FAIR principles of the European Open Science Cloud (EOSC) Association. This initiative is carried out at the University of Padua, Italy. The FAIRified data produced by this initiative can be used for the evaluation of the current FAIR metrics and the feedback for their improvements recommended in this TF.

4.2 Data Quality

For Data Quality the approach of work will start with consulting the ISO standards (e.g. 19158:2012, 19115-1:2014, 19157-2:2016, 9001, 14090/1, 2000, 9126, 25010/1) to identify the best practices associated with quality and the ways quality information is disseminated. Another approach of work is to prepare a survey about data to identify which requirements and dimensions define quality for the different communities. There should also be liaison (by attending conferences, contacting by e-mail) with ongoing activities performing data quality assessments *routinely* to identify what type of assessments are carried out and what metrics are in place to measure quality. Liaising with the activities will also inform about the benefits (risks) about improved (poor) data quality and related stakeholders.

Initial data quality use cases have also been identified:

- A first case at the Faculty of Social and Behavioural Sciences of the University of Amsterdam is linked to the verification of data with full provenance, at the time of archiving and publication. The checking satisfaction concerns the availability of a scientific article or other document that has a full description of the research methods, a codebook of the dataset(s) with all variable names and possible values and their meanings, a description of the steps that were taken to obtain the processed data and the licensing terms.
- This other activity in connection with INetQI (International Network on Quality Infrastructure) provides a focus on value and acceptance of the quality infrastructure and in particular on concepts related to quality, such as "accreditation", "standards baseline" and "conformity assessment". INetQI brings together organisations such as ISO, WTO, BIPM.
- In the healthcare domain, standardised methods of quality assessment have been identified. The challenge is now to measure the scientific and technical quality of data and metadata, such as Electronic Health record (EHR) data. This study case focuses on the concept of data quality such as completeness, plausibility, timeliness, etc. and refers also to the dimension of data reusability.
- Quantifying quality, usually with indicators, offers the possibility of fast and simple insights into the levels of quality and their patterns over time. Quality indicators are objective means for implementation of corrective measures and continuous quality improvement (Vuk et al. 2012[6]). There are several quality indicators, the maturity models are an example. In Earth science, a multitude of maturity models is available [e.g. Product System Maturity Matrix (EUMETSAT 2013[7]), Data Stewardship Maturity

[6] https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1751-2824.2012.01584.x

[7] EUMETSAT 2013 CORE-CLIMAX Climate Data Record Assessment Instruction Manual. Version 2, 25 November 2013. Available from: https://www.eumetsat.int/search?text=CORE-CLIMAX

Matrix (Peng et al. 2015[8]), DKRZ Quality Maturity Matrix (Höck et al. 2020[9])] and are characterized by discrete steps measuring the dataset maturity in categories, e.g. uncertainty characterization, user documentation.

5. Dependencies

The EOSC TF FAIR Metrics and Data Quality does not operate on its own. Input is required from ongoing community initiatives and projects that implement EOSC related to FAIR Metrics and data quality like RDA, the FAIRsFAIR project, the ESFRI Clusters, the Regional nodes/thematic projects, other global FAIR initiatives and emerging projects and initiatives under the Horizon Europe umbrella. This TF should offer feedback and advice.

This TF will have to work in collaboration with other EOSC TFs working with other sets of topics based on outcomes of previous EOSC governance decisions. First and foremost, this EOSC TF should work closely with the EOSC TF Semantic Interoperability (both under the EOSC Advisory Group Metadata and Data Quality), to check alignment about metadata and semantics. Other liaisons should be sought as well with EOSC TFs such as:

- Infrastructure for quality research software
- Long term Data Preservation, and
- Rules of Participation compliance monitoring

The outputs of this TF may not align entirely with the outcomes and timeframes of the EOSC projects and EOSC Advisory Groups/Task Forces. These outcomes and timeframes are dependencies for the success of this TF. Given the planned duration of this TF, the TF needs to take into account the emerging recommendations around the setup and evolution of the new legal entity called EOSC Association and associated procedures and processes.

Finally, policy making and funding authorities are also critical partners with interest in criteria for FAIR metrics and data quality.

6. Membership

Ideal TF members will represent a large range of European Countries involved in EOSC and a variety of disciplinary fields in order to take into account the specific needs of each community and benefit from their hands-on experience and different approaches. Gender balance and adequate geographical representation shall also be ensured.

The TF is open to a wide range of volunteers of varied positions, professional skills or expertise, (researchers, engineers, implementers, research support staff, software developers, data archivists etc.)It is expected to bring together different organizations such as universities, research organizations, private organizations as well as funding agencies to

[8] Peng, G, Privette, J L, Kearns, E J, Ritchey, N A, and Ansari, S 2015 A unified framework for measuring stewardship practices applied to digital environmental datasets. *Data Science Journal*, 13, 231 - 253. https://doi.org/10.2481/dsj.14-049

[9] Höck, H and Toussaint, F 2019 Quality Maturity Matrix Checklist for Levels 4 and 5 with Protocols. World Data Center for Climate (WDCC) at DKRZ. DOI:https://doi.org/10.2312/WDCC/TR_QMM_Checkl_Levels_4-5_Prots

ensure a broad membership base. The size of the group should not exceed 20/25 members to be manageable, efficient and reactive.

To fully participate and develop recommendations, the TF members must be able to share expertise and experiences, here is a non-exhaustive list of professional skills that are suggested within the group :

- Higher specifications for FAIR metrics since the work within EOSC is already well advanced
- Understanding of FAIR data policies and metrics
- Knowledge of FAIR issues and metadata standards
- Skills in data (and metadata) management [data life cycle]
- Experience with data validation tools and processes, data curation
- Experience in the practice of metadata repositories and standards
- Experience in the preparation and processing of surveys

The main coordinators of this charter:

Stefano Cozzini (Area Science Park, Italy) **Christine Hadrossek** (French National Centre for Scientific Research (CNRS/DDOR) – orcid 0000-0002-2638-6373) **Carlo Lacagnina** (Barcelona Supercomputing Center (BSC) – orcid : 0000-0001-9434-9809), **Ilona von Stein** (Data Archiving and Networked Services (DANS) – orcid : 0000-0003-3179-0773)

would like to thank all contributors for their valuable contributions and input:

Christos Arvanitidis (LifeWatch ERIC), **Andrea Bertino** (Switch, Zurich – orcid 0000-0002-5080-036X), **Sorina Camarasu-Pop** (CNRS, Creatis, France – orcid 0000-0002-7923-5069), **Anita de Waard** (Elsevier – orcid 0000-0002-9034-4119), **Giorgio Maria Di Nunzio** (University of Padova, Italy - orcid 0000-0001-9709-6392), **Sarah di Giorgio** (GAAR, Italy), **Salvatore Distefano** (University of Messina, Italy – orcid 0000-0002-2752-626X), **Anette Ganske** (Technische Informationsbibliothek (TIB), Hannover, Germany – orcid 0000-0003-1043-4964), **Duncan Jarvis** (Euramet), **Camilla Lindelöw** (OpenAIRE/National Library of Sweden), **Frans Oort** (University of Amsterdam – orcid 0000-0003-1823-7105), **Elli Papadopoulou** (ATHENA Research & Innovation Center / OpenAIRE – orcid 0000-0002-0893-8509), **Carlos Luis Parra-Calderón** (European Federation for Medical Informatics (EFMI) – orcid 0000-0003-2609-575X), **Joris van Rossum** (STM – orcid 0000-0001-8952-0421), **Sonja Schimmler** (Weizenbaum Institute, Fraunhofer FOKUS, TU Berlin (Germany) - orcid 0000-0002-8786-7250), **Pascal Siegers** (GESIS Leibniz Institute for Social Sciences – orcid 0000-0001-7899-6045), **Sarah Stryeck** (Graz University of Technology)

Acknowledgements as well to:

Christos Marinos-Kouris (Athena Research Centre) for additional support and **Sarah Jones** (GÉANT) for EOSC Board liaison.