European Commission

Opinion paper on FAIR data productivity

By the EOSC Steering Board expert group (E03756)

Independent
Expert
Report

Research and Innovation

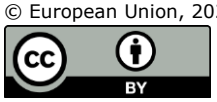
Opinion paper on FAIR data productivity by the EOSC Steering Board expert group (E03756)

European Commission
Directorate-General for Research and Innovation
Directorate A — ERA & Innovation
Unit A.4 — Open Science and Research Infrastructures
Contact Pantelis Tziveloglou
Email rtd-eosc@ec.europa.eu
Pantelis.tziveloglou@*ec.europa.eu*
RTD-PUBLICATIONS@ec.europa.eu
European Commission
B-1049 Brussels

Manuscript completed in May 2024
First edition

PDF ISBN 978-92-68-16211-8 doi:10.2777/49194 KI-02-24-584-EN-N

Luxembourg: Publications Office of the European Union, 2024



EUROPEAN COMMISSION

Opinion paper on FAIR data productivity

By the EOSC Steering Board expert group (E03756)

2024 Directorate-General for Research and Innovation EN

OPINION PAPER ON FAIR DATA PRODUCTIVITY BY THE EOSC STEERING BOARD EXPERT GROUP (E03756)

The EOSC Steering Board expert group (EOSC-SB) has invited the Policy Sub-Group to continue the overarching analysis of the main disruptive concepts and practices connected to the construction and future operation of the EOSC by addressing the issue of FAIR[1] data productivity. Productivity is defined as the ratio of new FAIR datasets over the full new data being collected by research: it is a relative value that represents how quickly the new information is made available in the web-of-FAIR-data. The challenge at hand is how to efficiently populate the FAIR data space with high quality contents that may enable reliable Machine Learning / Artificial Intelligence tools for assisting further research efforts.

The foreseen benefits of EOSC for European research and innovation require transformative practices with respect to FAIRification of data. FAIRification cannot rely on a-posteriori time and energy expensive methods and technologies as these will not follow the pace of new data production. This should be limited to latency data when time series or intrinsic non-reproducible data are concerned, or alternative on-demand by funded user projects.

1. Introduction

FAIR digital objects from research (data, software, technical solutions, workflows, algorithms) are the assets that need to be accessible, under transparent rules, in order to increase competitiveness, reproducibility, robustness and security of European research.

The Open Science approach is fostering the fast development of the EOSC and the building of a federated research environment upon the most successful practices already in place in several thematic communities as well as in institutional or territorial organisations.

The challenge of combining information and methods from different sources and areas imposes an overall coherence of action across different levels, from researchers to the upper governance. When successful, this approach will create a substantial competitive advantage for Europe, whilst developing all technology and methodology to protect intellectual property at the proper scale and secure the whole research data system.

At the level of the productivity of FAIR digital research objects, the goal is to implement a transparent set of common standards, adapted to different domains, for acquiring all new data and generating all new research objects that ***will follow the pace of research*** in the years to come. The productivity here is not in absolute terms, as this is dictated by the dynamics of research, but rather in relative terms indicating ***how efficiently and effectively the new information is made available as FAIR data and metadata sets***. 100% productivity would mean that 100% of the new data is automatically FAIR-ready and therefore shareable, according to the commonly agreed rules, with the whole research community.

A fundamental aspect is that of ***quality of the data*** that will be shared and possibly reused for new analysis and interpretation across thematic domains, aiming to address the global

[1] Findable, Accessible, Interoperable, Reusable.

challenges with the full integrated knowledge available from the many thematic streams of research.

Quality assessment is therefore a key element of FAIR-data/object productivity as the only useful product of research is the robustness of new information and consequent modelling and interpretation feeding into new advanced research and innovation.

Quality assessment involves technical and scientific aspects and can take advantage of advanced digital methods like Machine Learning and Artificial Intelligence under proper conditions of transparency and understandability. Accordingly, one can specify the FAIR character of datasets for ***direct use by researchers***, as well as indirectly through ***machine operability within the EOSC services***.

Assessing and protecting quality of FAIR data/objects throughout the whole pipeline from data acquisition to interpretation and reuse, or across the whole lifecycle of research results requires developing technology, policy and advanced protocols that will implement FAIR data/object sovereignty at the proper level and enable effective protection of the research assets of Europe from misuse, intrusion of unreliable data or malicious algorithms and cyber-attacks. Long-term preservation of valuable FAIR datasets shall also imply adequate policy and resource allocations to support FAIR-archiving, dataset curation and maintenance.

Alignment of practices fostering communication, collaboration and integration of knowledge requires an effort from all research operators in all disciplines. With robust policy and support, this can lead an effective transformation in science production and knowledge consolidation in a transparent and sustainable manner. This transformation is made possible and urgent by the new approach and by the digital advantage, guaranteeing independence, security, and progress.

In this opinion paper, the EOSC Steering Board identifies key recommendations developed upon consultation of the research community. Above all, it draws from insights from internationally operated Research Infrastructures (RIs) which have already reached significant results on data production, curation and sharing of rules. Working both at thematic community level and across communities, these institutions can inspire federation and generalisation at the full scale of research and innovation activities.

2. Main recommendations on FAIR data productivity:

2.1. FAIR datasets must be generated as directly as possible by the very actors conducting the research, who are also responsible for the scientific quality assessment of the datasets

To reach such objective, the research performing organisations (RIs, national or institutional laboratories) must implement, where suitable, FAIR by-design methodologies and technologies for automatically acquiring all relevant metadata that will enable usage of the datasets/objects. The metadata standards must be agreed and validated at community level as the first users of the datasets/objects will be those working in the specific field.

Reaching such goal implies dedicating specific resources, including training and employment of qualified data specialists (curators, stewards). Data specialists ought to be recognized in their professional contribution to the research process.

Currently, the main RIs and research performing organisations (RPOs) have "locally" developed solutions that are endorsed by their respective user communities and percolate to the users' institutions. Nevertheless, the agreed upon solutions need a diffuse support at the home institution level (universities, research laboratories, innovation organisations) to become a synergic system. Training researchers and dedicated data-professionals on advanced data management and curation should be organised in close collaboration with the specific domains, but within a coherent approach at all levels and broad, international, concurrence.

EOSC shall stimulate and lead collaboration towards this goal by sharing competences and resources, with a dedicated support to be harmonized at European and national level.

2.2. FAIR datasets and FAIR research objects must be generated at a rate that must follow the actual production of new information and knowledge as closely as possible

Sharing of the most effective proven solutions shall be supported as a goal of the federation, promoting the generalisation of practices deemed beneficial for new communities. Examples of "exportable" technical solutions across neighbouring domains do exist and are being further developed at the ***science-cluster level***[2]. The FAIR-by-design approach should be supported in all domains where the data are generated by detectors or computers, making technique-specific metadata fully automated and automatically built in the FAIR datasets. Advanced, perhaps AI-assisted, FAIRification of other kind of data, should be supported in the relevant domains to increase productivity.

A policy for FAIRification of latency data should be developed, supported by a well-justified "on demand" approach and appropriate resources.

2.3. Tools and methods supporting the production of FAIR datasets suitable to all domains and specific kind of data (observational, experimental, computational, statistical, correlational, analytical) must become part of the research life-cycle

The most challenging aspects of the FAIR method are interoperability and reusability in the broad sense (across disciplines/research methods). This is the case because they require an effort that is outside the current remit of research production organisations (RIs, RPOs, universities). Dedicated resources by EOSC must be put in place for developing both policy and methodology levels. For instance, thematic metadata may not be adapted for transdisciplinary reuse. This requires a meta-metadata level of information that shall create the proper interoperability level across archives and data lakes populated by thematic communities. Quality assessment of FAIR data includes a ***technical level of interoperability*** (data fit for purpose once the purpose emerges) that is additional to the ***intrinsic scientific quality*** assessed at disciplinary level. Coherence/compatibility of Data

[2] https://science-clusters.eu/

Management Plans and coherent/compatible authorisation access protocols must be enforced to effectively enable interoperability.

Providing adequate access to memory (***FAIR data archives***) and ***computing resources*** is crucial for removing barriers stemming from limited "local" resources (e.g. at RIs or RPOs) that are typically tailored to a restricted user community.

2.4. Quality assessment and permanent control of the FAIRness of the openly shared information must be assured with proper, transparent and robust protocols agreed among researchers and users

Quality Assessed FAIR Data (QAFAIRD) are the premium class of shareable and reusable data, as well as the only ***reliable class of data to be used for training Generative Artificial Intelligence***. Quality assessment has to do with different aspects: quality according to "fit for purpose" criteria, and science quality according to field-specific recognition of soundness and integrity of a FAIR dataset.

The former can be addressed with automatised checklists of technical requirements, to be tailored for the proposed usage of the data. The legal aspects must be enforced within the rules of participation to ESOC.

FAIR-by-design data should be ready for specialist use and reuse within the field but might not fit the purpose of a transdisciplinary usage. In such case the available metadata could be not entirely useful, and other metadata, less relevant for the specialists, might be missing. So, the technical quality shall be defined with respect to the ***reuse purpose***.

The science quality, on the other hand, shall remain under the responsibility of the owner of the data (the research group and, e.g. the RI where the data were acquired or generated) who should retain the right to withdraw the dataset upon new evidence of flaws or systematic errors (FAIR Data Sovereignty).

Automated follow-up of the FAIR dataset lifetime should be developed to guarantee that the ***data quality does not degrade through reuse*** of the data set. This is also a key element for ***ensuring research security***, enabling fast recognition of unreliable or malicious datasets.

FAIR data/objects need structured collaboration among research actors across the whole value-chain.

2.5. Sustainability of commons and protocols must be adequately supported as a key investment of science and innovation activities

It is recommended that suitable support actions are established as structural elements of the EOSC to enforce the commons at increasingly complex levels, i.e. from RI and well-structured research communities, to the ***long tail of science***, to the ***innovation sectors and to the citizens***.

EOSC commons should also become a reference for the other ***Common European Data Spaces*** whose intersection with the research sector must be guaranteed for mutual and general advantage.

2.6. Overall goals of FAIR data productivity

It is recommended that a ***suitable metrology*** on the ***QAFAIRD productivity*** is established and that realistic goals of QAFAIRD fraction of new datasets data and research objects are set, monitored, and constantly updated.

The QAFAIRD population of EOSC shall reach the critical mass for AI training, progressively enabling specific AI applications to become trustworthy and potentially useful as tools for orienting and accelerating the research workflows.

It is recommended that ***good practices on AI training with FAIR data*** are established as commons and enforced at the level of EOSC.

GETTING IN TOUCH WITH THE EU

In person

All over the European Union there are hundreds of Europe Direct centres. You can find the address of the centre nearest you online (european-union.europa.eu/contact-eu/meet-us_en).

On the phone or in writing

Europe Direct is a service that answers your questions about the European Union.
You can contact this service:

- by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),
- at the following standard number: +32 22999696,
- via the following form: european-union.europa.eu/contact-eu/write-us_en.

FINDING INFORMATION ABOUT THE EU

Online

Information about the European Union in all the official languages of the EU is available on the Europa website (european-union.europa.eu).

EU publications

You can view or order EU publications at op.europa.eu/en/publications. Multiple copies of free publications can be obtained by contacting Europe Direct or your local documentation centre (european-union.europa.eu/contact-eu/meet-us_en).

EU law and related documents

For access to legal information from the EU, including all EU law since 1951 in all the official language versions, go to EUR-Lex (eur-lex.europa.eu).

EU open data

The portal data.europa.eu provides access to open datasets from the EU institutions, bodies and agencies. These can be downloaded and reused for free, for both commercial and non-commercial purposes. The portal also provides access to a wealth of datasets from European countries.

In this opinion paper, the European Open Science Cloud (EOSC) Steering Board identifies key recommendations developed upon consultation of the research community. Above all, it draws from insights from internationally operated Research Infrastructures (RIs) which have already reached significant results on data production, curation and sharing of rules. Working both at thematic community level and across communities, these institutions can inspire federation and generalisation at the full scale of research and innovation activities.

Studies and reports

Publications Office
of the European Union