Fair Data productivity and advanced digitalization of research

An opinion paper by the ESFRI-EOSC Task Force and Steering Board expert group (E03756)

March 2024[i]

Prepared by:
ESFRI-EOSC Task Force[1]
[1] Full list of contributing members on page 8

The **FAIR principles** promote the reuse of research objects generated by researchers operating in all fields of science and innovation. This encompasses the enhancement of transparency in the research process, improvement of reproducibility of results and a novel opportunity for reuse of data, software, and analysis of results, for new research and innovation, also of transdisciplinary and interdisciplinary character, by the broad research and innovation community as supported by the European Open Science Cloud.

These objectives require that:

- FAIR datasets are generated and maintained as close as possible to the very same actors conducting the research, who are also responsible for the scientific quality assessment of the datasets;
- FAIR datasets and FAIR research-objects are generated at a rate that follows as closely as possible the actual production of new information and knowledge;
- 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 lifecycle;
- Quality assessment and permanent control of the FAIRness of the shared information, always compliant with clear rules of open or restricted access, must be assured with transparent and robust protocols establishing commons (principles, protocols and policy guidelines for sharing data as digital public goods);
- Sustainability of commons must be adequately supported as a key investment for science and innovation activities;
- Responsibility, or sovereignty, of the FAIR-data/objects are with their producers and extended to the funders of the research, who are securing systemic support of it;
- FAIR-data/objects need collaboration among research actors across the whole value-chain;
- Suitable rules of access and the use of the FAIR research data/objects must be enforced as common practices to realise the “common good” space of knowledge exchange.

This opinion paper identifies the state of the art of the research data and research digital object production within ESFRI Research Infrastructures (RI) as well as the bottlenecks to be addressed to comply with the above objectives.

The starting point is the currently most advanced communities and resources that generate research products, among which the research infrastructures and their thematic clusters. The goal is to federate those as a first necessary action to establish the above identified commons. The use of the FAIR data ecosystem concerns the whole research and innovation community, from the producers of the largest research datasets to the individual, long-tail of scientific users.

The ESFRI research infrastructures (RIs), as well as the full system of research infrastructures identified e.g. by the ESFRI Landscape Analysis that have international and thematic communities, and the main research performing organisations (national, institutional,

academic) provide the starting building blocks of the federated system that shall realise the first implementation of the EOSC programme.

The joint ESFRI-EOSC effort to address the Quality-Assessed FAIR-Data (QAFAIRD) productivity issue is articulated by the following key aspects:

1. The current FAIR data productivity level in the case of RIs / Clusters (only a fraction of the newly acquired data that becomes available are FAIR data sets/FAIR research objects within a given delay)

Most data are already 50% FAIR (Findable and Accessible) at production/postproduction stage for the hard science, environmental and life science RIs. This is the endeavour of RIs that serve broad, albeit specialized, user communities in the well-defined research domains of action.
This result already percolates to the broader community of researchers in the domain who can directly exploit the data, independently of having been involved in its generation.
Domain specific data analytics tools and methods enable e.g. research to be performed also at the long-tail level. This is the case represented by the clusters of analytical facilities (PANOSC), of astronomy and particle physics (ESCAPE), of biological and medical research (LIFE), of environmental science (ENVRI), of marine science (…)

Advanced services have been developed e.g. in ELIXIR, EPOS, LHC and other RIs that have the potential of being exported to other domains. FAIR research objects in the form of images are already suitable for use by “external” actors, including industry, to realise catalogues or advanced educational market products, or to stimulate curiosity and scientific interest in people.

Well-established archive data practices, as in astronomy (e.g. ESO) and marine science (Copernicus Marine, EMODNET, DTO) demonstrate that a large fraction of publications is based on archived data.

2. The goal of ideal FAIR data productivity and quality-control, in the case of RIs / Clusters.

The mission of RIs is to serve their identified user community with advanced metadata, data curation, archiving (at least 10 years), data analytics (software, image reduction, some – limited- computation service).
Metadata commons are to be met first at specialist community level and this engages the RIs that need to contribute by providing all detailed metadata for FAIR dataset reuse by the reference user community; beyond that it is the EOSC that must intervene providing the interoperability and reusability facilities suitable for transdisciplinary reuse of the FAIR datasets. Assessment of data quality has several facets: technical aspects regarding the compliance with the commons (e.g. machine actionability), legal aspects regarding intellectual property and GDPR, and scientific soundness.
RIs and their user communities must play a key role in the scientific assessment and technical reusability by the reference community, whilst the fit-for-purpose quality of the FAIR datasets

when reused for interdisciplinary or transdisciplinary research should be assessed at a different level (w/r RIs).
Overall quality assurance is aimed to foster research based on FAIR data use and reuse and, very importantly, to prevent technically FAIR but untrustworthy or fraudulent data to be added to the EOSC.

Artificial Intelligence can play a dual role with this respect. On one hand, it can be an instrument to help FAIRification and consolidation of new datasets/research objects, to match metadata schemas and standards as needed by different users, and to detect fabricated or falsified data. AI services, described by the EOSC and accepted by the federation are essential to enable effective data/object quality assessments and verifying the persistence of the quality throughout the lifecycle. On the other hand, AI can become a research infrastructure itself, capable of generating useful data provided the accessible algorithms are trained on QAFAIRD and the generated data subject to quality assessment and validation as all FAIR datasets.

Publishing annotated RAW DATA as a complementary route to fully analysed research papers is a possibility, for some communities, to substantially increase FAIR data productivity. This practice, although not applied generally, shall be regulated and controlled, e.g. by making mandatory citation of the Persistent Identifiers (PIDs) of the FAIR dataset. The availability of annotated RAW DATA could stimulate collaborations as the user would relate directly with the original data producer to develop and agree on a data analysis protocol and scientific discussion.

The FAIR data preparation for analysis and reuse amounts to substantial local work at the RIs and RPOs and requires dedicated staff with specific competences that today are scarce and insufficient. Resources and training are key elements that must be addressed for the needs of the FAIR-data/research objects curation and archiving enabling openness. This exceeds the limits of the focussed budget of the RIs requiring therefore a generalised support by EOSC specific resources. Interoperability for the broad user community is necessary for reuse and this is the EOSC's key contribution and role: connecting researchers across thematic and national domains and establishing priorities.

- A support policy and connected resources addressing the whole supply-chain of FAIR data/objects must become a common involving all relevant actors in order to build EOSC as a common good.

  Computing and networking services are provided primarily by national e-infrastructures (HPC, cloud and edge computing, high speed data transfer, archives and massive memories) and this aspect cannot be considered a granted service in the future, as the volume of EOSC activity will increase potentially absorbing a significant fraction of the national resources.

- Review and identify possible convergence of Data Management Plans (DMPs) of RIs, RPOs, individual instruments, and the effective enforcements of those.

- Update the European charter of access for research infrastructures to include the role of AI, Machine Learning and virtual laboratory resources.

- Establishing recognition criteria and value attribution to curator / data steward professional's work acknowledging their contributions at local and community level.

3. The bottlenecks that limit FAIR data productivity in RIs / Clusters.

General bottlenecks are: i) the overall insufficient and non-coordinated training of data curators and data-stewards; ii) insufficient staffing of these professionals across the whole FAIR value-chain; iii) insufficient local computing power at many RIs or resources to access external service providers ; iv) resistance to the generic concept of openness by some researchers; v) cost of archiving (albeit decreasing as novel technologies take over); vi) interconnectivity barriers; vii) substantial absorption of data transmission network capacity by heavy data traffic. All national e-infrastructures may have to face large international requests of service: large investments connected with EOSC are not limited to creation and operation of the federation, but involve the national e-infrastructures (networks, archives, computing resources) and the necessary inter-operation agreements.

The above imposes an analysis of the status of capacity and operation of the national e-infrastructures and a prevision of the developments of data traffic and data analytics requests generated by EOSC. The supporting infrastructure needs (data transfer, memory, computing) shall be addressed by national and EC collaboration, defining a clear business plan making policy and governance sustainable.

Intermediate objectives can be:

- Make all quality assessed NEW data FAIR but afford the FAIRification of latency data only on demand, with appropriate proposal scheme and funding mechanisms;
- Enable EOSC users, at short and long tail level, to combine data in a "robust way" to retrieve new information from FAIR datasets;
- Address and reduce computing models diversity, usage of different technologies and scarce dedicated expertise. This is in general problematic for experiments and computing centres;
- Global file transferring machinery connecting different storage systems.

4. The needed EOSC services to improve FAIR data productivity.

The EU node shall provide core services to enable the FA (Findable and Accessible) data/objects from the production side to become progressively IR (interoperable and reusable) for general users. Thematic and national resources (perhaps nodes) shall find in the EOSC the complementary services to give full value to the data produced, curated, and archived also progressively aligning to good practices and reference standards. Trusted and commonly accepted AI and ML services to feedback on technical and legal quality and clear

rules of participation with built-in the flexibility to match in real time the overall international advances.

In particular:

- Establishing/enforcing standards for interoperability and support to adopt and use them;
- Establishing coherent training actions and facilities (schools, contents) at European level;
- Support RIs to make the Findable-Accessible data sets more and more compliant with the needs of general Interoperability and Reusability;
- Support for those researchers/organisations engaging in the FAIRification of datasets;
- Monitor the efficacy of OSCARS and EVERSE and learn for suitable developments in the next Framework programme to facilitate collaboration and development of commons.

5. AI tools/solutions impacting FAIR data management.

AI is quality-FAIR-data hungry and is a driver for FAIR data productivity. The large projects to realize Digital-Twins, like the Ocean one, are examples of FAIR data integration. The Virtual Research Environment (VRE) developed by CERN is another example of a methodology that EOSC should disseminate making it available for other communities and cross-community research. Workflows, assisted or automated, may become a new tool for enhancing science productivity and universality. FAIR in the health-data domain (EU Heath Data Space-EHDS) is a frontier to be visited and the Life RI cluster is addressing it (BBMRI).

- AI assisted metadata schemas and technical data quality shall be developed to become key EOSC services assisting the effort of metadata matching and semantic crosswalk;
- AI has also potential in supporting data curation and categorisation of large datasets;
- AI experts must work in close contact with the research communities to advance in the above and optimise efficacy and uptake;
- “Fitness for the purpose” of FAIR datasets must be assessed as a key quality aspect;
- Data lake approach for data of different origins could be an effective way to facilitate the development of original combinations of FAIR data by more users;
- Multi-messenger research has been pioneered in astronomy along with the observation of gravitational waves, introducing the time dimension (simultaneous/time correlated observations), but it is a broader concept that will percolate in all research domains where diverse complementary observations on a phenomenon will constrain analysis and theory, accelerating discovery;
- Alignment of technologies and pursing joint efforts tools and services will bring closer the communities and foster commonalities.
- AI frameworks, Data Management Services, workload management systems, virtual service orchestration, monitoring and ticketing are examples of digital services that must be developed to become common facilities of the EOSC.

6. AI based research protocols are expected to impact research by RIs/Clusters and general science production.

The FAIR-AI alliance is already a fact that needs to be described and understood. One early example is the success of Alpha Fold I and II algorithms that, building on real data, have identified a range of targets for novel measurements, guiding full studies. Examples also exist in material sciences.
The combination of EOSC AI resources and ESFRI (and all RIs) experimental, observational, and computing capabilities (e.g. EGI) bears an extremely high potential for accelerating science in Europe, making it at the same time more open and more robust, even against unwanted intrusions or cyber-attacks. This can be achieved by fostering controlled data generation methods (both at the level of FAIRification and of generative-AI) and preventing the risks of intrusion of unreliable data generated by untraceable or commercial sources.

- AI must develop first in a field of specific endeavour, and then generalise the successful algorithms and practices to the broader transdisciplinary and interdisciplinary research goals.
- AI can generate data augmentation and create larger datasets for training of robust algorithms;
- AI models must be explainable to foster genuine new research, traceable and therefore usable;
- Generation of synthetic data (e.g. in the health domain using quality imaging data certified by BBMRI, by anonymization and synthesis help to guide therapy in a timely manner) can become effective instruments;
- AI and ML is at present perceived as a fundamental activity but at the same time requiring understanding of algorithms to boost hardware exploitation. Having the community aligned and sharing expertise is key.
- The interplay of FAIR and AI can generate new models based on quality data, whose validity may be established by performing novel observations, measurements and analysis;
- The RIs (ESFRI and national) will consolidate with observation, measurements, statistics and computation the valuable AI generated research targets;
- Analysis facilities, platforms, and Virtual Research Environments to be able to connect to content delivery.
- The new generation of researchers will fully exploit what we can develop now with an organised effort of active experts and the full research community;
- Policy (in the framework of the EU AI Act), development of commons, training and staffing of the research operators are strategic assets that will enable a new paradigm for science production and broad collaboration to address new knowledge in all fields and strengthen European competitiveness;
- The EOSC SRIA must fully address these issues.

ESFRI-EOSC Task Force Contributing Members

Please be advised that the following list is not exhaustive in terms of Task Force Member/Observer organizations. It encompasses all entities that have contributed to the Opinion Paper herein.

Chief editor

Giorgio Rossi, EOSC Steering Board

Members

Javier Lopez Albacete, European Commission
Volker Beckmann, EOSC Steering Board (Chair)
Martyn Chamberlain, European Commission
Elena Hoffert, ESFRI
Jan Hrušák , ESFRI

Task Force Support

Fotis Karayannis, StR-ESFRI 3

Observers

Jens Habermann, ERIC Forum
Stefan Hanslik, e-IRG
Mark Johnson (ILL), EIRO Forum
Bob Jones, EOSC Association

Observers Rotating on a 6-month basis

Tiziana Ferrari, ERF
Andy Goetz, ESRF

StR-ESFRI 3 project

Anna Bertelli
Georgia Kalozoumi
Tasos Patrikakos

[i] *Fourth draft based on results of ESFRI-EOSC Policy Workshop on "FAIR Data Productivity and Advanced Digitalization" (23-24 January 2024, Milan, Italy) and comments/integration to the previous three drafts circulated among the contributors and ESFRI-EOSC Task Force members.*