European Commission

Opinion paper on advanced digitalisation of research

By the EOSC Steering Board expert group (E03756)

Independent
Expert
Report

Research and Innovation

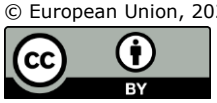
Opinion paper on advanced digitalisation of research

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 978-92-68-16212-5 doi:10.2777/932733 KI-02-24-585-EN-N

Luxembourg: Publications Office of the European Union, 2024



EUROPEAN COMMISSION

Opinion paper on advanced digitalisation of research

By the EOSC Steering Board expert group (E03756)

2024 Directorate-General for Research and Innovation EN

OPINION PAPER BY THE EOSC STEERING BOARD EXPERT GROUP (E03756) ON ADVANCED DIGITALISATION OF RESEARCH

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 advanced digitalisation of research. The challenge at hand is how to meet the advanced digitalisation needs of research from remotisation of collaboration (e.g. remote access to Research Infrastructures) to the development of robust and verifiable Artificial Intelligence instruments (deep learning of FAIR datasets, Automated Research Workflows, real-time Digital Twins…).

The foreseen benefits of advanced digitalisation of research require transformative practices to govern the value chain from data quality to reliability checks of AI-aided results, including rules of engagement with the commercial sector.

AI as a service for the development and operation of EOSC will be of importance for assisting FAIRification of data and technical Quality Assessment of FAIR research objects (datasets, software, workflows).

AI has a potential as research infrastructure on its own, provided that algorithms are developed within the research community, and that the transparency of code, the understandability of results, and the use of quality datasets for AI-training are ensured by proper means and common rules.

1. Introduction

Research reproducibility and data usability will be enhanced if the advanced digitalisation of data collection, validation, analysis, and simulation will be developed and become commons of the research community enforcing the Open Science principles and policies, creating a critical mass of Quality Assessed FAIR Data (QAFAIRD) and research objects enabling reliable and secure Artificial Intelligence, Machine Learning and Virtual Research Environments.

This opinion paper identifies the state of the art of the advanced digitalisation of research as well as the bottlenecks to be addressed to comply with the above objectives and to contribute to make a fully operational EOSC.

2. Advanced digitalisation of access to research tools and resources

Advanced digitalisation of access to research resources has undergone a big leap ahead also in the occurrence of the COVID-19 extended lockdown that imposed developing effective remote access protocols to Research Infrastructures and other laboratories to mitigate the impact of reduced mobility of researchers. The merits of remote access and remote collaboration are of general value even in standard operation of research resources as it represents an alternative, or better a complementary facility, to the physical displacement of users to the research infrastructures, laboratories, and observatories.

Remote access can in some cases enable the remote control of instruments. It is predominantly realised by implementing research workflows, where remote collaboration is facilitated by real time dialogue between researchers and facility operators. Such setup allows the researcher to access the stream of data being acquired and perform data analysis, enabling him/her to feedback to the actual conduction of observations and experiments. This process potentially optimizes access time and reproducibility of the acquired data, which in turn produces an optimal usage of the physical resources of the infrastructure.

Video communication, remote sensing sample/instrument manipulation facilities, real time data transmission, real time computing facilities, and robotisation of some basic but complex activities (such as sample conditioning in the measurement apparatuses and instrument calibration) are requirements for advanced remote collaboration. These technologies bear a high potential for research productivity and reproducibility and need further development.

Domain-specific protocols exist, for example, in protein crystallography, where standardisation and automation of sample alignment, sample quality testing, and acquisition of structural data have been implemented, with the development of community accepted standards. In other domains, e.g. materials science, one can expect similar developments to become available as important resources for research.

Virtual Research Environments (VREs), alias 'collaboratories', and their expansion to virtual research communities are medium-term developments addressing the above, with some notable realisations, for instance, by the Science-Clusters of thematic Research Infrastructures[1] (e.g. CERN and ESO). VREs require access to HPC resources and common services of EOSC. Composable, portable and integrated VREs are needed to support interdisciplinary research that also critically depend on the abundant availability of Quality Assessed FAIR datasets and objects. VREs have the potential to expand to cross disciplinary research projects, to advanced training in research, and to citizen science.

Artificial Intelligence (AI) and Machine Learning (ML) are a second highly relevant aspect of advanced digitalisation of research potentially benefiting the whole research community. These facilities are essential for the data mining or data visualisation techniques that are ubiquitous across research fields, particularly to address the major challenges they face today. These will be key elements to operationalise the concept of EOSC Science Exchange enabling understanding and interoperability of FAIR data.

AI is quality-FAIR-data-hungry and is a driver for increasing research and FAIR data productivity. Large projects to realise Digital-Twins, like the Ocean one, are examples of AI-based FAIR data integration workflows, assisted or automated, which may become a new tool for enhancing science productivity and universality.

AI algorithms need to be **developed first at research community level**. In fact, general-purpose algorithms are unreliable at present, and only a close collaboration with scientific expertise can drive the development of **transparent, understandable, and robust AI tools**. These in turn must be trained on Quality Assessed FAIR datasets for them to become potential drivers of new research results.

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

AI-assisted metadata schemas and technical data quality-checks 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. AI must rely on community-approved software stacks, FAIR datasets, and workflows to be reliable and transparent in terms of provenance of the inputs and to build reputation (e.g. results of AI4EOSC).

In the Life Sciences domain, a paradigmatic example is provided by the implementation of AlphaFold[2], which is being based on expertise in physical and biological knowledge about protein structure, leveraging multi-sequence alignments, to produce the design of the deep learning algorithm. Another case is the controlled generation of synthetic data, as implemented e.g. by BBMRI, opening an avenue to research finalised to clinical practice.

AI can in turn become a key approach to automated data curation, assisted metadata generation and enrichment, quality assessment and augmentation of FAIR datasets as training platforms, improving findability, accessibility, and interoperability. All of this has the potential to add robustness to the EOSC-Exchange of Science contents.

3. AI tools impacting FAIR data management and science exchange

This point is strictly linked with the FAIR productivity issue and regards the fact that AI-tools, shall be "trained" on high quality, well curated, FAIR datasets to yield understandable, verifiable, and overall useful results. This point crosses issues of ***FAIR data Sovereignty***[3] and of the relationship with the ***Commercial sector***[4]. The foreseeable tsunami of AI generated results shall be continuously addressed in the framework of the collaboration with ESFRI RIs and Clusters, the ERICs, the EOSC-A and e-IRG.

- AI assisted metadata schemas and technical data quality shall be developed to become key **EOSC services** (EOSC Science Exchange) 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;

[2] https://alphafold.ebi.ac.uk/

[3] https://op.europa.eu/en/publication-detail/-/publication/c016d11f-7f52-11ed-9887-01aa75ed71a1/language-en

[4] https://op.europa.eu/en/publication-detail/-/publication/37d444ab-7f51-11ed-9887-01aa75ed71a1/language-en

- Fitness-for-purpose assessment 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, central, for instance, in environmental research and assessment of climate change. This concept will further 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 the communities closer and foster commonalities.

- AI frameworks, Data Management Services, workload management systems, virtual service orchestration, monitoring and ticketing need to be developed.

4. AI based research protocols impacting research

The combination of EOSC AI resources and ESFRI (Science-Clusters 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 purely-commercial sources.

- AI must first develop in specific research fields, and then generalise 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 algorithms 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 anonymisation and synthesis) can become effective instruments to guide therapy without time-consuming GDPR barriers;

- AI and ML are at present perceived as fundamental activities, albeit requiring understanding of algorithms to boost hardware exploitation. Having the community aligned and sharing expertise is a key requirement.

- The interplay of FAIR and AI can generate new models from quality data;

- The RIs (ESFRI and national) will consolidate with observation, measurements, statistics and computation the valuable AI-generated research targets;

- The EOSC Strategic Research and Innovation Agenda (SRIA) must fully address these issues, identifying AI common tools as a new service for research;

- The new generation of researchers will fully exploit what we can develop now with an organized effort of active experts and the full research community. It requires specific training;

- 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;

- Data analysis facilities, platforms, and Virtual Research Environments must be developed.

5. Main recommendations

- Advanced digitalisation can enable optimal use of unique research resources as Research Infrastructures and foster novel research workflows with variable degrees of automation.

  - Support of these developments must be assured, aiming at improving the research outcome from the overall investment.

  - Enable access to digital facilities supporting the execution of observations, experiments, computation, and data analysis. It includes supporting the digital networks and services for preparing the background for new projects which crucially depend on the strengthening of Open Science practices;

  - Support the access to research instruments, or whole Research Infrastructures, in remote and virtual manner;

  - Support the access to data-analysis software and simulation;

  - Support the access to computing, communication, storage, and data curation services.

- Artificial Intelligence can become a key service and motor for EOSC-Science Exchange.

  - Support of this development must be assured as it requires coherent efforts by the Research Performing Organisations (e.g. Research Infrastructures) and EOSC. Research Performing Organisations and EOSC must federate the best practices and feedback by providing means aimed at the overall interoperability of research data, objects, workflows and results.

- Artificial Intelligence can itself become a research infrastructure of specific (thematic) and general (mission-oriented) value.

  - Support of this development is a must as it has implications on reliability, understandability, quality, and research security (Digital FAIR Objects Sovereignty) that do imply policy, rules of usage, computing, data transfer and archiving means

for long-term preservation of data and knowledge that exceed the remit of the FAIR data producers;

- Support the access to robust, reliable, and secure Artificial Intelligence algorithms and automated research workflows trained on Quality Assessed FAIR datasets (QAFAIRD);

- Develop suitable rules of access and usage of the advanced digital services as common practices to realise the “common good” space of knowledge production and exchange.

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.

Research reproducibility and data usability will be enhanced if the advanced digitalisation of data collection, validation, analysis, and simulation will be developed and become commons of the research community enforcing the Open Science principles and policies, creating a critical mass of Quality Assessed FAIR Data (QAFAIRD) and research-objects enabling reliable and secure Artificial Intelligence, Machine Learning and Virtual Research Environments.

This opinion paper identifies the state of the art of the advanced digitalisation of research as well as the bottlenecks to be addressed to comply with the above objectives and to contribute to make a fully operational EOSC.

Studies and reports

Publications Office
of the European Union