IJDC | *General Article*

Quality and Trust in the European Open Science Cloud

Juan Bicarregui
Scientific Computing Department
Science and Technology Facilities Council

Abstract

The European Open Science Cloud (EOSC) has the objective to provide a virtual environment offering open and seamless services for the re-use of research data across borders and scientific disciplines. This ambitious vision sets significant challenges that the research community must meet if the benefits of EOSC are to be realised. One of those challenges, which has both technical and cultural aspects, is to determine the "*Rules of Participation*" that enable users to assess the quality of the data and services provided through EOSC and thereby enable them to trust the data and services they access. This paper discusses some issues relevant to determining the Rules of Participation that will enable EOSC to meet these objectives.

Received 15 December 2019 ~ *Accepted* 19 February 2020

Correspondence should be addressed to Juan Bicarregui, Rutherford Appleton Laboratory, Oxfordshire, OX14 1DS. Email: juan.bicarregui@stfc.ac.uk

An earlier version of this paper was presented at the 15th International Digital Curation Conference.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

Introduction

The European Open Science Cloud (EOSC) has the objective to "*ensure that European scientists reap the full benefits of data-driven science, by offering 1.7 million European researchers and 70 million professionals in science and technology a virtual environment with free at the point of use, open and seamless services for storage, management, analysis and re-use of research data, across borders and scientific disciplines"* [1]. This ambitious vision sets significant challenges that the research community must meet if the research benefits of EOSC are to be realised. One of those challenges, that has both technical and cultural aspects, is to enable users to assess the quality of the data and services provided through EOSC and thereby enable them to trust the data and services they access. For example, data that is distributed though EOSC should have sufficient provenance information associated with it to enable users to assess its origin and quality, so that EOSC resources become not only be *accessible*, but also *assessable*.

If data and services provided through EOSC are to be assessable by users, resource providers will need to present sufficient information about their resources for users to make this assessment. Providers will have to undertake to abide by certain constraints, or *Rules of Participation*. Similarly, the fact that services are open and free at the point of use, does not mean that they are necessarily anonymously accessible without constraints. Service users will also need to comply with certain conditions of use, for example, in order for usage to be monitored and accounted.

The EOSC implementation roadmap [1] describes EOSC as comprised of data infrastructures that "*enter the federation on a voluntary basis based on the commitment of resources and on the capacity to comply with its rules"* under a *"compliance framework including notably the Rules of Participation.*" However, it is clear that different forms of participation will require different rules, so the Rules of Participation (RoP) may vary across research domains due to differing domain requirement and different functionalities of current data infrastructures. They may vary according to different locations, for example, where resources are provided for researchers from specific geographic regions. They may also depend on the nature the use, in particular when the EOSC user base is broadened to include both non-profit and commercial private sector stakeholders on the supply side as well as on the demand side. Furthermore, the RoP may need to evolve as needs and practices develop in response to compliance with existing and emerging legal frameworks, such as GDPR and the free flow of data.

The Implementation Roadmap further envisages "*a pan-European federation of research data infrastructures built around a federating core including a compliance framework which outlines the Rules of Participation*." However, compliance constraints will be different for the different stakeholders participating in EOSC in different roles. For example, rules for data providers will be different from those for services providers and from rules concerning users of data and services.

Rules of Participation for Data Providers

Openness

The underlying motivation is to open up all aspects of research, so that research processes and outputs can be reviewed and reused. The principle of openness should therefore permeate all aspects of EOSC. In particular, data made available through EOSC should be available for use by anyone for any legitimate purpose. Furthermore, if these data are to be usable by others, then sufficient metadata to enable that reuse must also be universally available.

On the other hand, some exceptions to openness may be required in particular cases, for example, in order to comply with legal constraints such as the GDPR and the Copyright Directive. Furthermore, openness does not necessarily mean that data are accessible anonymously; prospective users of data may be required to authenticate themselves with personal or organisational credentials, for example, so that data usage can be monitored and accounted.

Transparent Subsidiarity

While participating as a data provider in EOSC implies commitment to the principles of openness described above, custodianship of the data remains with the data provider. Thus, individual data providers determine the precise conditions under which the data they expose though EOSC may be accessed and used provided that these do not contradict the underlying principle of openness. Such resource-specific Terms of Use may, for example, require users to inform the data providers of the purpose for which the data will be used.

In line with the principle of transparency, data providers will clearly define and publish any such Terms of Use for the data they provide. These will include any licensing information, whether access requires authentication and/or authorisation, and any conditions regarding how data can be processed, changed and redistributed by users.

Free at the Point of Use

Although the aim is that EOSC data and services should be freely available, clearly there are costs involved that need to be covered, and providers need to be fairly compensated for their efforts. Given an appropriate recognition and reward system, researchers who create or analyse data should be motivated to prepare and make their data available for others just as they are motivated to write and publish papers. In a culture where research outputs are valued appropriately for their contribution to knowledge, whatever form the outputs take, researchers would be as eager to open their data as they are to publish papers. However, the effective opening of data requires infrastructure that is best provided by dedicated data infrastructure providers for whom academic recognition may not be the primary driver. Many such infrastructure providers already exist and are sustained through a variety of funding mechanisms, including structural funding, institutional funding, project-based funding, data deposit fees and data access charges (OECD Global Science Forum, 2017). Given the widely agreed principle that publicly funded research data should be a public good, it is natural for research funders to

support the means to maximise the use of that data through support for data infrastructures, thus enabling data to be made available freely at the point of access.

Registration and Discoverability

The EOSC will be primarily a federation of existing data and services where data remain in their current repositories and EOSC provides a means to make those data more broadly discoverable and interoperable. To enable this federation, ESOC must recognise resources or collections of resources through registration of those resources in an EOSC catalogue. Participation in EOSC is therefore defined by registration of resources as EOSC resources or in an EOSC-recognised collection of resources. Although somewhat tautological, this definition embodies the fact that participation works on a voluntary basis, if and when a provider chooses to register a resource with ESOC, it becomes discoverable and accessible through it. A digital resource is therefore considered to be an EOSC resource if and only if it is registered in an EOSC-recognised catalogue of resources. Registration of resources also indicates compliance with the EOSC Rules of Participation and use of EOSC branding is available only to registered resources.

Rules of Participation for Service Providers

Regarding service providers, three types of provision can be distinguished: the federat*ed* services that are brought together by EOSC, the federat*ing* services that enable EOSC to operate as an integrated whole, and EOSC *compliant* services that are external to ESOC but are useful as part of research workflows.

Rules of Participation for Federated Services

As for data, in order to be available to EOSC users, services that are federated in EOSC need to be registered in a service catalogue that is itself registered with EOSC. This is not to say that users will necessarily access these services though a generic EOSC gateway, rather that researchers may continue to access resources through their existing field-specific portal with these portals being enhanced through access to wider range resources, mediated and adapted by the providers of the domain-specific resource. As with many forms of infrastructure, providers of existing portals may be able to hide the technical details of how services are delivered and seamlessly present new functionality in a way that is tailored to communities in their specific fields.

For such an invisible infrastructure to be achievable and maintainable, service descriptions and protocols will need to be provided in both human[1] and machine readable forms. The metadata supporting this may include: parameters related to terms of use including any accessibility constraints and/or quotas; the means of accounting and monitoring; measures concerning assessability and quality of service including any service levels; definitions for technical interoperability such as API descriptions; and declarations related to liability[2].

1 Human readable form does not necessarily mean that humans can easily read the raw metadata. Machine readability against standardised schema is sufficient to imply human readability, as software can be employed to allow humans to easily interrogate machine readable metadata.

2 See EOSCpilot D2.5: Recommendations for a minimal set of Rules of Participation for details (Kahlem, et al., 2018).

For these metadata to be machine processable without the need for software to be hard-coded to particular schemes requires the definition and agreement of the metadata schema and vocabularies to be used. While it is unrealistic, in the short term, to expect all communities to agree on a single, universal metadata scheme, it is feasible to envisage adoption of a registration service for schemata with the individual schema being agreed within specific communities through global consensus building activities such as those supported by the Research Data Alliance (RDA).

Rules of Participation for Federating Services

The EOSC federating services are those core services that are required to support the functioning of EOSC itself, enabling it to function as a federation. Such federating services include those concerning: authentication, authorisation and accounting; registration of users, organisations and projects; monitoring and accounting of usage; and service and data catalogues. Central to this suite of services, and also underpinning findability and accessibly, are the persistent identifier services that can provide some necessary stability and provenance in an otherwise highly dynamic and flexible environment.

These federating services will necessarily be subject to more stringent requirements in order to support the levels of availability and reliability that users will expect from a functioning research infrastructure. Unlike the federated services, each of which will have their own independent community-focused funding mechanisms and metrics for success, the federating services are generic in nature and will therefore be more directly linked to the EOSC governance framework through qualitative and quantitative Service Level Agreements.

Rules of Participation for EOSC Compliant External Services

It should be recognised, however, that EOSC will never provide, nor should it attempt to provide, all the services, resources and tools that will be used by researchers. Many tools such as Internet search engines, social media communication channels and office systems tools are currently provided, and will continue to be provided, by suppliers external to EOSC. An important consideration for EOSC will be how to accommodate use of such external tools into research workflows, and whether a notion of EOSC compliance needs to be developed for such external tools and services.

It is also crucial that EOSC interoperates with other open research support environments outside of Europe. Research is global, therefore research infrastructures need to support global communication and collaborations. Here global reciprocity agreements and discussions, such as those provided by the RDA Working Group on Global Open Research Commons, are an essential component for establishing common principles.

Rules of Participation for Users

As mentioned above, although data and services provided through EOSC will be open to all and free at the point of use, this does not necessarily mean that all resources will be accessible without constraints. It may be necessary, for example, for academic users to identify themselves, along with their affiliation and project, and agree to some terms and

condition for use of a specific service, whereas researchers from outside the academic sector, perhaps from a not-for-profit or commercial enterprise, may require specific registration and authorisation. For resources that are not non-rivalrous, such as compute, storage and perhaps even bandwidth, quotas or merit-based authorisation may be required.

As part of registration, users of EOSC data and services agree to adhere to the RoP and not to wilfully violate the Terms of Use for specific resources as determined by the resource provider. For example, users of data may need to agree to acknowledge the source of the data they use in every communication where they make use of, or refer to, the data resource. If required to do so, data consumers may also be required to acknowledge the intellectual work of the original creator(s) of the data. Where a resource stipulates a standard form for this acknowledgment, this form will be the form used. Where a persistent identifier is provided for the resource, this will be quoted in the acknowledgement.

Quality and Trust

The above Rules of Participation define an operating framework for EOSC providing a level of assurance regarding the quality of resource sharing mediated by EOSC. However, a governance framework is also required if these if these RoP are to engender trust in these resources.

Proper Research Conduct

Considering the aim to enable the review and reuse of data, it is essential that data and services be shared in a way that supports this. Therefore, resource providers and users must act in accordance with commonly agreed principles regarding the conduct of research[3] that assure, for example, the quality of research underlying the data and do not wilfully misrepresent or provide false data.

Quality of Data is Different from Value of Data

Data resources can include material from all stages of research, including raw measurements, observations or simulations, processed data, analytical parameters, and summary data. Data can be static or time dependent and it can be persistent or streamed in real time. In every case data are entirely valueless without information that enables their interpretation. These metadata might include descriptive information and provenance information, as well as information regarding curation and sustainability mechanisms.

It is important to distinguish quality of data from value of data. Data of low quality can be of high value. For example, a researcher may be interested in novel data even if it is of uncertain quality. Conversely, data of high quality can be of low value; an instrument can produce high quality data about an uninteresting subject which is therefore of low value.

3 For example, The European Code of Conduct for Research Integrity: https://ec.europa.eu/research/participants/data/ref/h2020/other/hi/h2020-ethics_code-of-conduct_en.pdf

In any case, researchers must be able to ascertain the quality of the data being used if they are to draw valid conclusions from it. Whilst EOSC has no role in the assessment of data value, it can have some role in enabling judgements of particular aspects of data quality. By encouraging data providers to indicate the quality of their data through supporting metadata descriptions, and enabling community feedback on these assertions, EOSC can provide a basis for community-specific indication of data quality.

Setting and Maintaining the Rules of Participation

To oversee its implementation and operation, EOSC requires a governance structure with the authority not only to determine the strategic direction of EOSC but also to assume ownership of the RoP, to monitor compliance, and to oversee their evolution. For the initial phase of development, until the end of 2020, an EOSC Executive Board has been established that will work within the context of relevant policy documents[4] to provide advice regarding the implementation of EOSC and to make recommendations on possible forms of future EOSC governance. To undertake this work, the Executive Board has established a number of working groups covering Landscape, Sustainability, Architecture, FAIR, and Rules of Participation. Among these, the RoP Working Group[5] (WG) is charged with setting out the standards governing the rights, obligations and accountability of the providers and users of data and services within EOSC.

The de facto starting point for EOSC are the data infrastructures underpinning the current European Research Infrastructures and e-Infrastructures, so the EOSC must, be compatible with practices in these systems. The RoP WG will therefore engage with relevant initiatives to understand the motivations of parties participating in EOSC. By understanding current practice and setting a relatively low barrier to entry, the on-boarding of existing research services into EOSC should be straightforward. Furthermore, the possibility to agree and adopt optional quality certification standards, or "badges", that require higher standards, alongside the minimal RoP should enable quality assessment and trust to cross domains without disruption of existing practice.

The RoP WG is therefore focusing on recommending a minimal set of Rules of Participation that will define the rights, obligations and accountability governing transactions between the various EOSC users, providers and operators. In order to ensure a low barrier to inclusion of existing services in EOSC, emphasis is being put on requirements that are common across the heterogeneous European landscape of research infrastructures and services. The RoP will therefore facilitate the federation of entities into EOSC, balancing the usefulness to their "own" community and other communities, and considering community-driven development versus interdisciplinary-driven evolution. The RoP will aim to guarantee an open, secure, cost-effective and pan-European EOSC that is compatible with other international initiatives.

The RoP WG has recently published an initial set of RoP for community discussion and will be receiving feedback during 2020 leading up to a "Version 1" set of rules for use in EOSC from 2021 onwards. After this, the ROP will continue to evolve on an on-going basis as EOSC matures to meet its ambitious vision of universality.

4 For example, the EOSC Implementation Roadmap (March 2018) SWD(2018) 83 (European Commission, 2018a); the ECI Communication (COM(2016) 178) (European Commission, 2016); the workprogramme 2018-2020 (C(2018)7238 (European Commission, 2020); and the Recommendation on Access and Preservation of Scientific Information (L134/12 2018) (European Commission, 2018b).

5 See the EOSC RoP WG web site at: https://www.eoscsecretariat.eu/working-groups/rules-participation-working-group

Acknowledgements

The author would like to thank the EOSC Executive Board and RoP Working Group for their essential contributions in formulating the ideas discussed in this paper.

References

European Commission. (2016). Communication: European Cloud Initiative – Building a Competitive Data and Knowledge Economy in Europe. COM(2016) 178. Retrieved from https://ec.europa.eu/digital-single-market/en/news/communication-european-cloud-initiative-building-competitive-data-and-knowledge-economy-europe

Europoean Commission. (2018a). EOSC Implementation Roadmap (March 2018). SWD(2018) 83. Retrieved from https://ec.europa.eu/research/openscience/pdf/swd_2018_83_f1_staff_working_paper_en.pdf

European Commission. (2018b). Recommendation on access to and preservation of scientific information. Retrieved from https://ec.europa.eu/digital-single-market/en/news/recommendation-access-and-preservation-scientific-information

European Commission. (2020). The Work Programme 2018-2020 – European Research Infrastructure (including e-Infrastructures. Retrieved from http://ec.europa.eu/research/participants/data/ref/h2020/wp/2018-2020/main/h2020-wp1820-infrastructures_en.pdf

Kahlem, P., Jimenez, R., Smith, A., Lecarpentier, D., Castelli, D., Zoppi, F. (2018). EOSCpilot D2.5: Recommendations for a minimal set of Rules of Participation for further details. Retrieved from https://eoscpilot.eu/content/d25-recommendations-minimal-set-rules-participation

OECD Global Science Forum. (2017). Business models for sustainable research data repositories. Retrieved from http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=DSTI/STP/GSF(2017)1/FINAL&docLanguage=En