GEDE - EOSC: Views and Wishes from RI Experts - 1st Version 1.1

Editors: Peter Wittenburg, Carlo Maria Zwölf, Zsuzsanna Szeredi

27.6.2019

## 1. Preremarks

This first report version is based on interactions with 22 experts[[1]](#footnote-1) from different research infrastructures (RI) involved in GEDE. After the virtual GEDE meeting at 28.6[[2]](#footnote-2) we will make this version open for comments and contributions as a Google Doc (see <https://docs.google.com/document/d/1kVJa3K7x4wlv3NnyRnJC1YwnAWeK_NnkMSBOCUIBpmA/edit?usp=sharing>).

GEDE is a loose collaboration of 47+ research infrastructure initiatives (mostly ESFRI/ERICs) and is based on finding maximal agreements on views based on rough consensus, but not neglecting views that indicate disagreement. GEDE relies on open discussions between experts and it is up to the participating initiatives to appoint the experts for these discussions. Therefore, the statements made do not per se represent official statements from the individual initiatives.

## 2. Views on EOSC

### General Points

It is commonly agreed by the GEDE experts that any support and funding for fostering sustainable Open Science and FAIR Data and Services is very much welcome, since the tasks ahead of us are huge. In this view EOSC needs to be seen as a great opportunity for European research.

EOSC and similar initiatives in member states take a lot of attention and are consuming substantial effort and time of key people; however, experts feel that their time is not spent efficiently and that their efforts don't have effects.

The political timeline and that of RI practitioners offer a huge gap. RIs need often solutions within short time frames, while the EOSC processes due to its complexity and political involvement take months if not years.

It will be a continuous point of discussion where exactly the boundaries are between the responsibility of EOSC and that of the individual RIs are, in particular, since these boundaries will change over time. As an example it was mentioned that RIs produce the metadata relevant for their researchers. Adding additional information (context, provenance) for others will costs funds and is not trivial. Will there be additional funds, will there be special institutions doing enrichments/brokering such as performed by ANDS (Australia)? It is expected that EOSC will help in clarifying the borders and responsibilities in interaction with the RIs.

### EOSC Image

EOSC wants to be an open process, but factually for the experts involved in most RIs it is a closed process. There are many interaction channels[[3]](#footnote-3), participating actors, boards and meetings making it impossible to follow and to judge relevance. As a consequence the state of activities such as the formation of EOSC WGs was basically unknown and people see EOSC as a big bubble where some actors being paid for these policy level actions are participating and which is far away from their needs. Only in very selective cases colleagues have a direct link to a member of a board and thus have access to appropriate information. Effectively the voices of the RI experts are not represented in the EOSC process.

Yet, EOSC content is not well defined, it is utterly complex tackling many issues at the same time and it lacks the "down-to-earth" approach. Therefore, EOSC discussions until now are very abstract, mostly of political nature and far away from the matters being tackled by the RIs. It is therefore not obvious for many whether EOSC will be something that will be useful to foster competitive research.

### Expectation & Experience Uptake

The EOSC concept makes especially sense in relation to cross-disciplinary research and thus it should facilitate cross-disciplinary research. EOSC should stimulate (and fund) cross-disciplinary research to create a need for using a cross-disciplinary infrastructure.

It was stressed by colleagues that EOSC should primarily focus on data and not on tools/services. The latter are subject of high competitive dynamics and partly discipline specifics and technological innovation will push new solutions. Generic type of components which need to stay for longer periods need to be regarded differently and are not subject of this competition. With respect to generic solutions, however, one should not hesitate to use commercial solutions as long as certain conditions are met. The ownership about data should remain with the researchers and when using commercial services there should be exit options.

Many colleagues have the impression that at political level the efforts that are needed to put abstract concepts and principles effectively in place is underestimated. FAIR principles are great and FAIR Maturity Indicators are required, but the gaps to the data practices in the labs are huge. The experiences made by the RI and the use cases from RDA are underused.

Several colleagues expressed their feeling that they cannot see how the existing solutions and knowledge about building infrastructures and components of relevance is being included in the discussion processes. It seems that we lack a neutral forum where existing solutions are being evaluated. However, EOSC needs to build up on existing solutions to be efficient. Current RI components should be analysed to determine cross-disciplinary possibilities based on abstraction processes.

One colleague looked at the published funding model for EOSC. Expecting funds from the various RIs and projects is not seen as feasible.

### Role of Clusters

Clusters have been built amongst others to channel views from RIs and all cluster initiatives organise the discussions differently. Colleagues do not feel well represented by "their" clusters in all cases. Cluster projects are seen as useful instruments to try out new technologies and methods, sometimes they seem to be used in political ways which can be problematic.

## 3. Comments on Major Issues

### Service Shell

Basically RI experts see EOSC as a federated shell to combine the data and services the various disciplines are creating and needing in addition to some centrally offered facilities. Basically, they all expect to make their data, tools, services and repositories visible to others via the EOSC. There is the expectation that other RIs could benefit from their work and knowledge.

It is also expected that EOSC will develop "frameworks" that would facilitate the setup of repositories, the building of workflows, the execution of distributed workflows, etc. In the interaction it was not easy to define exactly how these frameworks would look like.

The idea of a "market place" for data, tools and services with massive contributions from the RI is widely agreed upon and it is agreed that EOSC should play a role in organising such a "market place", but a number of considerations were mentioned:

* terminology needs to be discipline oriented to be understood and accepted
* the "market place" concept needs a proper definition, since it could easily be misleading
* features such as semantic tagging, pro-active advertisement etc. should be in place
* bringing metadata of data, tools and services from all RIs would create an enormous haystack, leading to the question how this needs to be organised to make it useful
* in this respect the "EOSC Portal" was not seen as a feasible solution
* the concept of "catalogue of catalogues" was mentioned, but just seen as one possible idea
* one needs to carefully check which type of discovery solution makes sense and has sufficient add-on value to what Google already delivers
* if one would like to include data and tools in workflows for example, simple metadata would not be sufficient, one would have to extend the metadata descriptions which is not a trivial task
* assigning software maturity levels is seen as problematic for scientific software, since often newly created software makes the scientific difference
* tools/methods to facilitate formal licensing specifications are needed to facilitate automatic processing
* options to track usage would be excellent since they would increase trust, but they are not easy to realise
* in some communities "reference collections" have a great relevance, EOSC could have a register of such collections which can be used for various purposes (calibration, training, etc.)

### Central Services

Urgently needed is a functioning eduGain service which it can be used by everyone. This seems to be the most hampering bottleneck for many RIs.

Another urgent Europe-wide services would be a Handle Service available for everyone which EOSC should push forward. Handle Systems and DOIs (also being Handles) are widely adopted in many communities.

There are already centrally offered services (B2DROP (EUDAT) was mentioned as one example) but they are not yet known by most researchers. Sometimes they lack a few features that would make them simple to use or they define small restrictions hampering usefulness. Therefore, some decided for example for Google services since they are professional and "you only need a credit card". No one is expecting long-term provisioning from GAFA, but their offers are even better than the 2-year lifespans academic services can in general promiss.

For some communities centrally provided stores and VMs would make sense - this does not mean that additional hardware resources at EC level need to be bought. Rather a flexible platform with defined access points allowing easy actions would be required. A lean logical centralisation combined with physically distributed solutions is aimed at. Smart and easy to access/use solutions seem to be optimal for the purposes of the RI compared to monolithic and inflexible services. There is no need for an expensive administrational layer. But it was made clear by several that the main task of EOSC is NOT providing basic services for storage, cycles, etc. The huge problems are to be found in the lack of interoperability and automation. Achieving interoperability which necessarily includes provenance and context information at different levels should be core in the EOSC funding plans. This includes the need to have well-maintained solutions to register schemas and vocabularies. Platforms such as RDA need to be used to define such solutions.

EC made efforts to offer useful data services (EUDAT, OpenAIRE), but there is a lack of offering the next logical step: carrying out work on these data. EOSC is expected to overcome this gap.

### Sustainability and Long-Term Preservation

The other big bottleneck is the lack of a "long-term" perspective for what has partly already been started in the RIs. There is no trust from researchers to use some of the already existing services since their funding is often only guaranteed for short periods. Yet, there is no convincing funding model for long-term sustainability of data and services.

In addition, cross-country procurement rules are missing which is inadequate given that research is crossing national boundaries. Europe risks of lying behind due to administrational inflexibility. EOSC could play an enormous role to improve the situation even if EC will not have funds to guarantee sustainability, but it could at least put this on the agenda of the MS.

EOSC key components need to be designed with long life times (20 years and more) in mind. Where possible such key components need to be self-supporting elements in the emerging infrastructure eco-system.

Costs for the curation of digital data are high and underestimated and it is completely unclear who is going to pay for it.

It was doubted whether continuity could and should be achieved at EC level. National funds should be used for this.

### Repositories as Pillars

Repositories are key pillars in many RI and a number of points in this respect were mentioned:

* RIs already frequently form federations with data, compute and service centres as their backbone network.
* Guidelines about data organisation, service quality, legal issues, etc. are required to help making repositories FAIR compliant.
* CoreTrustSeal seems to be widely accepted as rule set to assess quality and to increase trust; it should gradually being extended and improved.
* re3data should be used to register repositories and it should become a pillar in EOSC. But it was also mentioned that re3data would need an uplifting to cope with the increasing requirements
* A more generic description of repositories characteristics would be needed to improve efficiency, since increasingly often repositories need to provide information to various service providers.

In these RI federations it is important to take care of the visibility of the individual repositories and service centres. They often need to offer special services for sub-disciplines where the support of specific terminologies are crucial for success.

### FAIR, DO & Automation

FAIRNess of data and services is important and EOSC should be the caretaker in this respect. However, FAIR principles alone wil not be enough to achieve interoperability. Semantic mapping will remain an era where handwork will be required. Therefore, EOSC should fund the development of frameworks to make, for example, semantic mappings much easier.

EOSC should adopt the FAIR DO model since it is offering cross-disciplinary interoperability at data management level and fund the development and testing of yet missing components and testbeds. The FAIR-DO model is stressing metadata, but it needs to be extended to explicitly include provenance and contextual information to be sufficient.

Improving automation is urgently needed but not easy to achieve in practice. Flexible, handy frameworks such as Jupyter are ways to go, but more support needs to be available to make the creation of workflows easier by reusing available components.

### Culture, Trust & Sensitive Data

EOSC needs to give impulses to promote a data sharing culture. Since trust with its many facets is key to come to such a component of open science, EOSC should push trust building measures. Measures of traceability would also help since it would improve the visibility of the creators.

In this context also blockchain technology might be worth considering. The special issues of sensitive data and the GDPR requirements are often neglected in EOSC discussions until now.

### Training & Support

EOSC is expected to give a lot of support on training and fellowship programs and help in producing useful manuals/guides. There seems to be a willingness in the RIs to actively contribute to such education activities. EOSC would have to organise this in a smart way and measures need to be concrete such as being achieved with Hackathons using real data etc. Since the preparation of such practical courses are expensive, more funds need to be available.

In addition, helpdesk functionality to be funded by EOSC should be offered for aspects such as FAIRness, the use of agreed components and services, etc.

Some colleagues even would expect the availability of a developer team that takes care of filling eminent software gaps on request, but it is admitted to be difficult in organising such a team. Should they be associated with ERICs with the risk of silo solutions or as a separate unit with the risk of having a team following their own interests?

### Other Topics

RDA work is seen as important but often it lacks funds to participate actively. EOSC should reserve some funds to support the active participation of experts in RDA WGs.

1. It should be noted here that "RI experts" cover a mix of data scientists, infrastructure specialists, data managers, data stewards, etc. [↑](#footnote-ref-1)
2. At the recent GEDE video meeting some EOSC actions were explained and with respect to some comments made actions are already on the way. It was, for example, welcomed that a separate EOSC WG on "sustainability" was established. [↑](#footnote-ref-2)
3. Although most experts were positive about GEDE's involvement, we also heard a few critical voices, since finally GEDE is yet another channel. [↑](#footnote-ref-3)