European Open Science Cloud and Research Data Alliance (Europe)

Peter Wittenburg (RDA Europe)

Key Statements

From an RDA perspective a few key statements can be made and EOSC needs to be designed according to the spirit of these statements:

- Data sharing is a reality and despite all current hurdles it will make its way and revolutionise science.
- The domain of data and data science is undergoing enormous changes and there is an increasing pressure to identify infrastructure components that have a common usability and stability.
- Since the success of Internet, self-regulation at global and cross-disciplinary level became the norm to identify such components. Infrastructure complexity is so high that mainly speeding up the virtuous circle of creating ideas and testing them will guarantee competitiveness.
- Imposing too bureaucratic solutions will not work, constrain creativity and hamper European innovation.
- A balance between the different stakeholders and a leading role for those driving innovation in science needs to define governance.
- Essential is to create a climate that encourages EOSC to become an active and respected global player in supporting science, in advancing infrastructure work and in supporting innovation in particular by start-ups.

Background

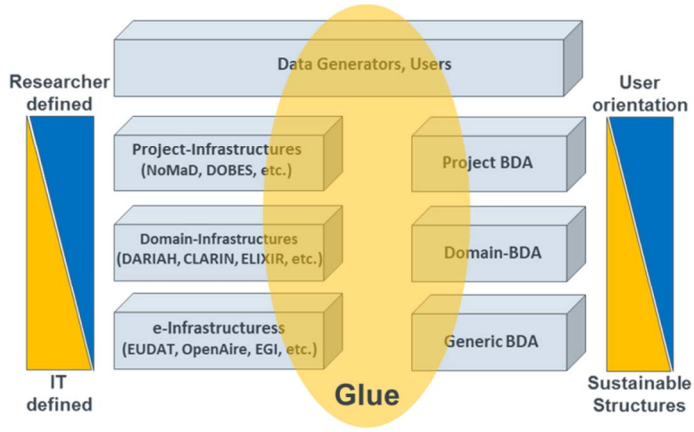
Open Science and Open Data is a priority issue to allow scientists to tackle the Grand and many Small Challenges of societies and to inspire new types of scientific insights and new types of businesses creating completely new types of jobs for data professionals and knowledge workers. Open Science and Open Data (as indicated by the G8[1]/FAIR[2] etc. principles) require the availability of a user centric and highly integrated and interoperable landscape of sustainable services offered by powerful and persistent centres around some identified and interoperable core components. Yet we are still far away from such a user driven ecosystem of infrastructure with services, although the situation has improved during the last few years due to an increased focus towards user driven services and a better understanding of shared core common components. This is also strongly demonstrated through the recent OECD report "Making Open Science a reality"[3].

G8/FAIR/etc. on data:
searchable/findable
accessable
interpretable/re-usable
manageable/persistent

The following diagram which is an extension of the well-known diagram from the High Level Expert Group on Scientific Data (see Riding the Wave Report[4]) characterizes the landscape made up by variety of information infrastructures (II) driven by clear research needs, by community expectations or by the insight that many services or components can successfully be shared across disciplines. Currently a large number of such Information Infrastructure initiatives are being funded and developed globally, on the European level and also at the national level. In parallel we see much

[1] https://www.gov.uk/government/publications/open-data-charter/g8-open-data-charter-and-technical-annex
[2] https://www.force11.org/group/fairgroup/fairprinciples
[3] OECD (2015), "Making Open Science a Reality", OECD Science, Technology and Industry Policy Papers, No. 25, OECD Publishing, Paris. http://dx.doi.org/10.1787/5jrs2f963zs1-en
[4] http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf

work being invested into the direction of more or less generic big data analytics software packages such as for machine learning, stochastic modelling, etc. The current funding practices follow the principle of "let 1000 flowers blossom", which might after all be the only way to deeply understand and master the social and technical challenges in the rapid digital change to prepare and engage the researchers and data centres for all the data challenges ahead. Earlier attempts to use a generic template from one science discipline and drop it on to other sciences disciplines have failed and any top-down dominated initiative will also be doomed to fail. Taking the early Internet initiative experiences as an example, which was also confronted with a large fragmentation, it is widely accepted and agreed that a broad dominantly bottom-up, rough consensus driven initiative can pave the way for the agreements necessary to identify and specify recommendations for common components, preventing major fragmentation.



RDA has proved its role in building elementary parts of the glue for collaboration with other major initiatives and stakeholders at the practitioner level (W3C, IETF, etc.) and policy level (CODATA, G8, etc.). However, RDA will not build this technical ecosystem of infrastructures. We need to rely on strong consensus driving initiatives such as the European Open Science Cloud (EOSC) as another elementary step towards this ecosystem made up of user centric services and to function as the glue at the implementation level.

Challenges for EOSC

Therefore a few characteristics can be mentioned for the EOSC that seem to be essential from the point of view of RDA:

- *The major dimensions that need to be looked at to enable efficiently work with data have been identified by the well-known data principles (G8, FAIR, Nairobi, etc.). Along all dimensions a large fragmentation can be observed. EOSC must give a substantial impulse to overcome this fragmentation across disciplines, countries and projects by supporting bridge building embedded in a global dialogue.*
- *The EOSC should give a substantial impulse towards implementing the domain of registered*[5]*, metadata described and safely stored and accessible digital objects*[6] *which is compliant with the model worked out in the Data Foundation & Terminology group of RDA*[7] *and with the basic FAIR principles. We call this domain the* ***Internet of Data (IoD)***[8]*.*
- *The EOSC should also give a substantial impulse to make knowledge (schemas and semantics) explicit and to extract assertions where possible to come to a rich and machine processable domain of knowledge.*

We need to accept that an exploratory phase was absolutely essential in infrastructure building, but that now a change in strategy towards consolidation and interoperability is required.

[5] Data should be *associated with a unique and persistent identifier that is worldwide resolvable.*
[6] *Data to be stored in trustful and certified repositories.*
[7] https://rd-alliance.org/groups/data-foundation-and-terminology-wg.html
[8] The term Internet of Data emerged in discussions with(in) the FAIR community.

- *A key for overcoming fragmentation is to identify and implement common components and common services in a way that efficient integration and interoperation can be guaranteed. Together with global initiatives EOSC should take an active role in accelerating the process to come and test such components and thus to come to global recommendations.*
- *Common services of EOSC need to go beyond basic offers such as storage capacity, CPU cycles, network capacity and include all virtualization layers facilitating data to be searchable, accessible, re-usable and manageable. EOSC therefore needs to act on these two layers: basic infrastructure layer of IoT and virtualisation layer of IoT. Each of these layers requires different approaches, technologies and governance.*
- *The EOSC should not be built from scratch, but evaluate components and services that have been deployed by the various infrastructure initiatives and be built on top of the useful ones that can be maintained and do not lead to technology lock-ins.*

EOSC needs to tackle and combine two layers of IoD each being guided by different principles, knowledge and experts:

- basic infrastructure layer
- virtualisation layer

- *The basic infrastructure layer of IoD should be based on already existing powerful and persistent centres that have shown their willingness to offer services at a pan-European level to researchers in an efficient way, i.e. EOSC needs to be a federated system of centres offering relevant services.*
- *The virtualisation layer of IoD should be driven by domain experts that have shown their capability, insight and openness, i.e. EOSC needs to engage experts that can drive advancements.*

- *The EOSC should strengthen the activities to train next generations of professionals who can build advanced infrastructures and who can work efficiently making use of infrastructures.*
- *The EOSC should be governed by bodies that combine deep insights about scientific and societal needs and technological possibilities guaranteeing that science is driving decisions.*
- *Infrastructure funding must develop in a way that focusing on the testing and adoption of common components and sharing common services will become an essential part, since otherwise sustainability cannot be achieved.*

RDA can offer collaboration opportunities for EOSC in most of the described areas except for building physical infrastructures or implementing hardware components. However, RDA has proved to be a successful platform for bringing researchers, computer and data centers, service providers, research funders and other stakeholders together for fruitful discussions on many levels. In particular the strength being through,

- *RDA's mission is to work on the various social and technical bridges at a cross-disciplinary, cross-country level and in doing so it produces suggestions for globally relevant common building blocks and services, guides the global testing work and thus has the potential to come to global recommendations.*
- *RDA Europe can help fostering and accelerating the global interaction process covering many disciplines and can participate in the evaluation of existing components and services on a fair and neutral ground.*
- *RDA Europe can help in building the social bridges which seem to be crucial to overcome the silo mentality.*