EARTO Background Note:
The US Open Science Data Cloud

7 July 2017

At a time when the policy debate in the EU has been very much focused on the open access to research data, EARTO has been actively involved in the discussions with for instance the publications of the EARTO paper on Open X and the EARTO Background Note on the US Federal Agencies Data Sharing Policies. As a follow-up, particularly in the frame of the discussions on the creation of a European Open Science Cloud, this short note aims at giving an example of a specific Open Science Cloud focused on a few scientific fields in the US. In conjunction with the analysis of the US federal agencies' data sharing policies, this note contributes to provide a global understanding of the US policy for Open Science. This could bring an interesting perspective into the EU debate, also given the fact that Research and Innovation is now very globalised and that competition takes place at an international level.

1. Rationale: A storage & computing infrastructure to store & analyse large datasets

The US Open Science Data Cloud (OSDC) is a cloud-based storage system and computing infrastructure which provides the research community with resources for storing, sharing and analysing very large scale of scientific datasets[1] in very specific fields like genomics or satellite maps. The rationale behind such cloud is to say that with the increasing size of datasets, the bottleneck to scientific discovery in fields where big data is an issue, is no longer the lack of data, but rather an inability to manage, analyse, and share large datasets"[2]. Created in 2010, the OSDC was built to remove this bottleneck. It is especially useful for medium to large datasets.

2. **Management: A non-profit entity responsible for maintaining the infrastructure**

The OSDC is a hosted platform managed by a non-for-profit entity, the Open Commons Consortium (OCC). OCC members include over 30 universities, companies, Research & Technology Organisations (RTOs), national laboratories, and government agencies. One of the OCC's goal is to maintain the infrastructure and ensure the integrity and long-term access to data. On top of the OSDC, the OCC manages several other similar data cloud platforms.

3. Approach: bottom-up, with focus on a few data-intensive fields

The OSDC is based on a bottom-up approach. It is mostly used if really needed to advance science in specific data-intensive research fields where big data plays an important role, such as biology (genomics), earth sciences (satellite maps) or social sciences. For instance, a large amount of computing resources are made available to research projects through a selection process so that interested projects can use the OSDC to manage and analyse their data.

4. Data & Openness: Focus on public input data required for research projects

The OSDC is much more focussed on storing public freely available input data (raw data) to be analysed during research programmes and projects – DNA (mostly provided by citizens, patients) or satellite maps (raw data mostly provided by NASA) for instance, rather than on output data (data outputs of the research programs created by the researchers and/or analysed data) stemming from research projects, which often has restricted access. Indeed, the OSDC was designed to provide an infrastructure for researchers in data-intensive fields to analyse large datasets and develop and test new types of data-intensive algorithms. In addition to storing large public datasets, the OSDC allows its users to share private datasets with specific users or groups of their choosing.

5. Sustainability Model: computing infrastructure and access to input data available at a price

Computing infrastructure and access to Input Data are charged on a cost-recovery basis. The billing and accounting system has proved very efficient to limit unpleasant behaviour and maintain a good quality of shared resources.

Besides, one of the OCC's responsibility is to find other means of ensuring the sustainability and maintenance of the OSDC, for instance through partnerships with R&I organisations to gain research funding, raise funding from donors and not-for-profits, or work on reducing the costs of operations.

[1] Currently, the OSDC consists of more than 3700 cores and 6.6 petabytes (PB) of storage distributed across four inter-connected data centres.

[2] https://www.opensciencedatacloud.org/

Sources:

- The Open Science Data Cloud's website
- The Open Commons Consortium's website
- The Design of a Community Science Cloud: The Open Science Data Cloud Perspective
- An Overview of the Open Science Data Cloud

EARTO - European Association of Research and Technology Organisations
Founded in 1999, EARTO promotes Research and Technology Organisations and represents their interest in Europe. EARTO network counts over 350 RTOs in more than 20 countries. EARTO members represent 150.000 highly-skilled researchers and engineers managing a wide range of innovation infrastructures.

RTOs - Research and Technology Organisations
From the lab to your everyday life. RTOs innovate to improve your health and well-being, your safety and security, your mobility and connectivity. RTOs' technologies cover all scientific fields. Their work range from basic research to new products and services development. RTOs are non-profit organisations with public missions to support society. To do so, they closely cooperate with industries, large and small, as well as a wide array of public actors.

EARTO Working Group Legal Experts*: is composed of 25 corporate legal advisers working within our membership. Established in autumn 2013, this Working Group has also worked on the revision of the State-Aid Rules & the GBER. Our experts also contributed to the setting-up of the DESCA Consortium Agreement model for Horizon 2020. More recently they were at the origin of the EARTO Paper on Open X, the EARTO Background Note on the US Federal Agencies Data Sharing Policies, and the EARTO voting recommendation for Globally Competitive Standardisation in the Digital Single Market.*