Charter of the National Data Infrastructure Architecture Working Group

Version 2.0 of 9 January 2024, by David Antoš (david.antos@cesnet.cz) et al.

1 Introduction

One of the main pillars of EOSC implementation in the Czech Republic is the establishment of the National Repository Platform for scientific data. It is to serve as one of the essential elements of the infrastructure for scientific data in the Czech Republic, together with an environment for storing unstructured data, a computing environment, and an environment for collaboration.

Background: the e-infrastructure in the Czech Republic operates several data repositories with a total volume of tens of PB. This includes mainly file-based or object-based repositories with minimal support for metadata storage. Part of the storage capacity is intended for data directly processed in the computing environment and is strongly interlinked with it (e.g., geographical proximity, direct access), and part is for data of a more archival and longer-term nature. A significant amount of scientific data is simultaneously stored in a number of disciplinary repositories at national and international levels or held by scientific teams or even their institutions with limited visibility and accessibility beyond the immediate authors and users.

One of the key objectives of EOSC implementation in the Czech Republic is to build a national environment for scientific data storage and processing that will significantly reduce the current fragmentation, be integrated into a single infrastructure, and provide flexible platforms to cover the specific needs of scientific groups and user institutions.

The National Data Infrastructure Architecture Working Group (ANDI WG) aims at the design and architectural oversight of the implementation of the national data infrastructure, specifically the platform for data repositories and immediately downstream services such as the technical implementation of persistent identifier handling, metadata collection and indexing, metadata search, binary reliable dataset storage and the integration of these functionalities into the national e-infrastructure as a whole. The functional and quality requirements for the national data infrastructure will be mainly formulated in and with the other WGs. The ANDI WG is developing these requirements into a system implementation and architecture design concerning scalability and performance of the solution for data volumes in the order of tens of petabytes, ensuring integrity and availability of data and metadata, integration of individual components at the technical level, and considering the economics of the solution.

2 Objectives

- **The National Repository Platform (NRP)** should provide the functionality of a general national repository (i.e., storage of data with metadata and especially persistent identifiers). The NRP will be closely linked to the national e-infrastructure, the development of which must be coordinated. The NRP will enable the creation of specific repository instances for specific purposes, e.g., covering metadata and search specificities of individual institutions or expert communities (up to the level of discipline-specific clusters). In doing so, these communities will move away from having to run their own infrastructure, but it is expected that expert communities will put the effort required to create specific repository instances specifically to support specific

metadata and specific workflows. We anticipate that one of the NRP instances will be a generic national repository for data without specific disciplinary or institutional affiliation.

- **The metadata directory for scientific data** will aggregate metadata from all repositories that will form the national data infrastructure (in particular, all instances in the NRP and disciplinary and other repositories). It will thus serve as an access point for user searches (thus ensuring data discoverability).
    - At the same time, it will also be suitably connected to international systems such as OpenAIRE, international disciplinary repositories, etc.
    - Machine-usable interfaces (APIs) will also be provided for data access.
    - Interoperability with the IS VaVaI will be set up.

WG will create/develop:

- Technical architecture of a scalable multi-tenant repository platform for scientific data and publications (tens to hundreds of PB).
- Standards for interoperability of repositories at the technical level.
- Methods to ensure the required quality of service.
- Technical treatment of working with sensitive data.
- Integration of repositories into the national data infrastructure so that metadata about finished live data (e.g., data directly from instruments) and the datasets themselves can be efficiently transferred to the repository(s), with metadata being automatically generated as far as possible (all while maintaining complete control of the data owner over its access). Very close cooperation with the metadata directory WG is expected here.
- A strategy for the long-term preservation of binary data regarding operational economics.
- In the medium and long term, tools for LTP (Long-Term Preservation)
- The technical architecture of the Metadata Directory of Scientific Data, again in close collaboration with the relevant WG.
- A set of recommendations and standards for the implementation of discipline-specific cluster repositories
    - as NRP instances,
    - for integrating existing repositories into a metadata collection system,
    - for integrating existing repositories into a proprietary data transfer system (technical interoperability).
- PID allocation system architecture.
- Linkage to other data services of the national e-infrastructure.
- Linkage to the AAI architecture of the national e-Infrastructure.
- Methods of technical implementation of principles such as 'data is under the full control of the data originator'.
- Suggests priorities for implementation work of individual components, considering available development capacity.

The WG does not aim to create new standards and practices unless absolutely necessary (an important principle is "what can be adopted, let it be adopted"). Therefore, The WG will follow European and global trends in repositories and cooperate with relevant communities, such as European e-Infrastructures, Zenodo, OpenAIRE, DataCite, CrossRef, and others. The group outputs must align with EOSC policies (such as PID Policy and Architecture and compatibility with AAI principles).

By the time parts of the infrastructure are operational, the WG should be transformed into a technical advisory body for the operation and development of the infrastructure.

European Open Science Cloud
Czech Republic (EOSC-CZ)
Masaryk University, Institute of Computer Science
Mailing address: Sumavská 33, 602 00 Brno
info@eosc.cz; www.eosc.cz

3 Outputs and their applications

The architecture designs will be an important advisory input for implementing systems within the e-infrastructure. The agreed standards will be binding for the infrastructure once the steering committee approves.

4 Membership and expected members

The WG is open to all interested parties, and the number of members is not limited. However, applicants are expected to have experience in building infrastructure services, at least at the level of a scientific institution or a larger project, optimally in the role of system architects. It would be advisable to have representation from institutions, discipline-specific clusters, and projects or institutions that already operate significant repositories.

The organization of the WG, including the election of the leader(s) and his/her deputy(s), is subject to the Statutes and Rules of Procedure of the EOSC WG in the Czech Republic.

The group is expected to cooperate with relevant partners in the Czech Republic and abroad. In particular, other project WGs will be the primary national partners for intensive cooperation.

European Open Science Cloud
Czech Republic (EOSC-CZ)
Masaryk University, Institute of Computer Science
Mailing address: Sumavská 33, 602 00 Brno
info@eosc.cz; www.eosc.cz