Future Trends for the Data Fabric

*Data Fabric IG, P11*

**Over the past plenaries, the RDA Data Fabric IG has worked on issues focusing mostly on the role and exploitation of Persistent Identifiers (PIDs), originating in a discussion on essential components that are common across a wide variety of research infrastructures and research data management practices. There has been significant progress in the area of PIDs, resulting among other things in a supporting output document, the GEDE group and the Kernel Information group.**

**It is now time to revisit the original concerns of data fabric components and look at other aspects pertinent to the Data Fabric idea. The Plenary session in Berlin will be used to discuss these points and set the pathway for the future activities of the group. While we need a set of examples to feed the discussion, the list is in principle open-ended, and all contributions before and during the plenary are most welcome.**

## A common object and collection management approach (separate joint session)

*Tobias Weigel, Larry Lannom, Peter Wittenburg and others*

RDA has produced several outputs around digital objects, including in particular the PID Information Types and Research Data Collections specification. The PID Kernel Information WG is preparing another output that complements the Information Types recommendation. In addition, many communities have established and matured metadata schemas and services around them, such as searching and curating.

The RDA recommendations describe common APIs for working with properties of individual objects and aggregations of multiple objects. However, the management of objects in research infrastructures can benefit from a more comprehensive management of objects, no matter whether they are singular or in collections: All of them require agreements on the most common operations, such as create, copy, move and delete, but also second-level operations with small predefined workflows, such as replication, versioning or other derivation actions.

**This is a super topic across the other contributions listed further below. It will have a dedicated joint session at the P11.**

## Object management and provenance in data analytics

*Sandro Fiore, Sofiane Bendoukha, Tobias Weigel*

For the European Open Science Cloud (EOSC), the ENES community provides a comprehensive data analytics service called ECAS (ENES Climate Analytics Service). Originally designed for usage scenarios within the Earth system modelling community, the service is now available for further cross-disciplinary use. It is based on the concepts of datacubes, enabling efficient computation on structured, multidimensional data.

As part of future developments, ECAS can become connected to PID management services on all sides (input, output, processing, …). This opens up a chance to build a small-scale demonstrator that exploits the provenance capabilities laid out via Kernel Information. One of the FAIR aspects is re-usability, and an approach based on this exemplary analytics service, embedded within EOSC, can provide a showcase both within individual communities and across them with real users involved in the evaluation and design.

## ENVRI provenance concerns and their mapping to Data Fabric components

*Abraham Nieva de la Hidalga, Alex Hardisty, Barbara Magagna, Stephan Kindermann*

GDOC virtual layer components serve as a model of the complex ecosystems that support research infrastructures operations. For instance, IS-ENES performs a seven-step data lifecycle from data generation to long-term archiving and publishing. In this lifecycle, generation and management of provenance data is required to facilitate tracking relations between data products generated by IS-ENES entities. The IS-ENES data infrastructure consists of six types of entities that support the research data lifecycle: (1) modelling centers and (2) observation data providers are primary sources for (3) data nodes; (4) index nodes provide cataloguing services to facilitate access to data nodes; (5) compute [processing] nodes facilitate generation of derived data products on demand; and (6) access portals facilitate access to data products.

IS-ENES components map directly to three types of GDOC objects: Repository Registries (observation data providers, data nodes) Ecosystem of Tools [Tool Registries] (modelling centers, compute nodes), and Metadata Registries (index nodes). Additionally, four types of GDOC components are required to complement the IS-ENES model: PID registries to provide data product identifiers, Authorization and License Registries to manage access to collections (through IS-ENES portal), and Type Registries to facilitate access and reuse of data. Similar mappings to GDOC components are expected to occur in other ENVRI Community research infrastructures.

## Metadata components and metadata fabrics

*Rainer Stotzka, Kirsten Elger*

In RDA many disciplinary groups aim to develop or seek help in metadata descriptions within their scientific domains. The term “Metadata Fabric” aims to develop a methodological and technical approach to support a resourceful and technically sustainable research data management by metadata. It will invest in a systematic analysis of current metadata methodologies, technologies and standards within and outside of RDA to derive commonalities. As a long term goal a set of interoperable technical components will be recommended as a basis for a metadata infrastructure harmonizing with the components of the data fabric.

## SDONet：Big earth data infrastructure for CASEarth

*LI Jianhui, CHEN Xin, LIU Jia*

In 1986, Chinese Academy of Sciences initiated a project on scientific databases development and sharing. Since then, significant progresses have been made in scientific data management and sharing in China, forming a technique stack and tool sets for the management and service of distributed scientific data. More than 350 distributed research databases and 1350 datasets by 61 institutes have been established, and over 600TB data available for open access and download.

Last month, CAS has just launched a new project on Big Earth Data called “CASEarth”.This is a CAS Strategic Pioneer Research and Development Programme mainly focusing on building a global big data network to study Earth and support research on climate change, as well as to predict and mitigate natural disasters. Data sharing will be the key means to carry out the Big Earth Data and the interdisciplinary data discovery, and reuse will be a big challenge.

To achieve the vision of CASEarth, a novel data infrastructure SDONet-Scientific Data Object Network, which is compliant with the FAIR principles, has been proposed and funded by CASEarth Programme. Data will be described and wrapped up in a universal format, including PID, metadata, data entity and extended service parameters. At the level of implementation, SDONet consists of five main parts: 1) a set of distributed-deployed data publishing tools to register and publish data in semi-SDO formats; 2) a common repository to provide unified storage and accessing service; 3) a data integrating system to organize and manage multi-disciplinary data in SDOs; 4) a set of registries for PID, data types, dictionaries, etc.; 5) a unified search and service engine for data discovery and sharing. These systems and tools will be developed in the future 2 or 3 years. SDONet will try to open some common components, and engage the GDOC activities.