Big Data in Science and the EUDAT Project

Wolfgang Gentzsch[1] and Damien Lecarpentier[2]

[1] EUDAT project advisor and independent consultant for HPC and Cloud Computing,
Regensburg, Germany
[2] EUDAT project manager, CSC – IT Center for Science Ltd, Espoo, Finland
emails: Gentzsch@rzg.mpg.de, Damien.Lecarpentier@csc.fi

1. Introduction

Started in October 2011, the pan-European data initiative EUDAT brings together a unique consortium of 25 European partners – including research communities, national data and high performance computing (HPC) centres, technology providers, and funding agencies. EUDAT aims at building a sustainable cross-disciplinary and cross-national data infrastructure that provides a set of shared services for accessing and preserving research data. The design and deployment of these services is being coordinated by multi-disciplinary task forces comprising representatives from research communities and data centres.

2. EUDAT building a generic multi-disciplinary infrastructure

EUDAT aims at laying out the foundations of a Collaborative Data Infrastructure (CDI) in which centres offering community-specific support services to their users could rely on a set of common data services shared between different research communities. The main focus of EUDAT will be on building this common layer of generic cross-disciplinary data services. Although research communities from different disciplines have different ambitions and approaches – particularly with respect to data organization and content – they also share many basic service requirements. This commonality makes it possible for EUDAT to establish common data services, designed to support multiple research communities, as part of this CDI.

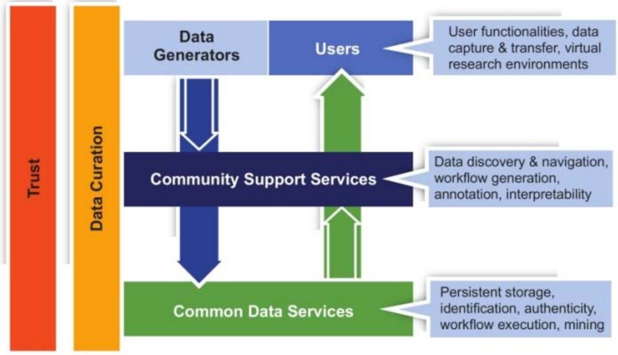
Trust
Data Curation
Data Generators
Users
User functionalities, data capture & transfer, virtual research environments
Community Support Services
Data discovery & navigation, workflow generation, annotation, interpretability
Common Data Services
Persistent storage, identification, authenticity, workflow execution, mining

Fig. 1. The Collaborative Data Infrastructure - A framework for the future © HLEG on Scientific Data, 2010.

By supporting the infrastructures that existing scientific communities have for their generic data services, the CDI will enable the communities to focus a greater part of their effort and investment on services that are discipline-specific. The CDI will also provide individual researchers, smaller communities, and projects lacking tailored data management solutions with access to sophisticated shared services, thus removing the need for large-scale capital investment in infrastructure development. Lastly, by providing opportunities for disciplines from across the spectrum to share data and cross-fertilize ideas, the CDI will encourage progress towards the vision of open and participatory data-intensive science.

3. The EUDAT data services

EUDAT has been reviewing the approaches and requirements of a first subset of communities from linguistics (CLARIN), earth sciences (EPOS), climate sciences (ENES), environmental sciences (LIFEWATCH), and biological and medical sciences (VPH). Several service cases have been identified by these communities as priorities and pilot services are being built jointly within the EUDAT project through multi-disciplinary task forces involving representatives from communities and data centres.

3.1 Data Replication and HPC Access

There is strong demand among the research communities involved in EUDAT for data replication services associated with better access to computing power. This demand underpins two of EUDAT's common data services – safe data replication, and the ability to move data to and from HPC facilities. When combined, we expect that these services will constitute a fundamental component of the CDI.

The "safe replication" service team is working on developing a service that will make it possible to replicate data from one site to another, for example, from a scientifically-oriented community centre to a data centre. This service is required across all five research communities, in particular it is needed to facilitate better data access and data preservation.

One of the strengths of the EUDAT consortium is the massive amount of computing power available at European HPC centers involved, most of which are members of PRACE and among the most advanced supercomputing centers in the world. EUDAT will leverage the experience gained in DEISA and PRACE to build an infrastructure that can provide access to this computing power. Once users have their data replicated on the EUDAT infrastructure, we indeed anticipate that they will want to be able to use neighbouring computing facilities to analyse this data.

3.2 Joint Metadata Catalogue

Complex problems or "grand challenges" increasingly require a trans-disciplinary approach relying on data coming from multiple research fields. In this context, making data from various disciplines available in one collaborative infrastructure can be extremely beneficial. This requirement is shared across the five research communities, not only to allow them to make their data more visible, but also to make it possible to work with data coming from other disciplines.

Part of the challenge resides in finding good solutions that allow metadata from different communities to be integrated into easily searchable catalogues. EUDAT is currently investigating the best way to develop a joint metadata catalogue, using as a starting point the OAI-PMH protocol for harvesting metadata from communities.

3.3 Simple Store

In addition to providing services to large research communities, individual researchers and small projects will also be catered for, with a "simple store" service that allows the storage and sharing of the vast quantity of ”small” data, that is, data that is not part of official data sets or collections, but that is equally important for the advancement of research. EUDAT will provide a means to easily store, search, view, and retrieve this data.

This Simple Store service complements other EUDAT services that will manage large volumes of official community data. Ultimately the Simple Store will enable and encourage users to access and share data that would otherwise be unavailable to them.

3.4 Training as a service

Training plays an important role in EUDAT. In particular, we must ensure that potential users and providers of the infrastructure are fully trained in how to optimally use, operate and extend the platform of technologies, tools and services provided by the project. Users of common EUDAT data services appear at two layers: the researchers as end-users and experts from community-specific centers.

In its first year, EUDAT has been undertaking a needs analysis involving the EUDAT user communities in order to develop appropriate training programmes. EUDAT has already identified some important core technologies and associated best practices. Our first training event took place in June 2012, in Amsterdam, and focused on three important areas: Policy/Rule-based Data Management, The Use of Handles for Persistent Identification, and Distributed Authentication and Authorization.

One of the key aspects of our training activities in the near future will be to develop partnerships with existing training providers, especially those addressing specific disciplines, or with projects working on common solutions for a cluster of research communities, such as ESFRI.

4. What’s coming next?

After now one year of activity, significant progress has been made by EUDAT to lay out the foundations of the CDI. We expect that the services being designed in EUDAT will be of interest to a broad range of communities that lack their own robust data infrastructures, or that are simply looking for additional storage and/or computing capacities to better access, use, re-use, and preserve their data. The first pilots will be completed in 2012 and the services will be available to all communities in a production environment by 2014.

In the coming months, increasing effort will be put on two strands of activities: the first consists in planning for the operation of the infrastructure, particularly providing secure, reliable (generic) services in a production environment, with interfaces for cross-site and cross-community operation. The operation of the infrastructure should provide full life cycle data management services, ensuring the authenticity, integrity, retention and preservation of data, especially data marked for long-term archiving. The second strand is to plan the evolution and sustainability of the infrastructure. Among other things, this implies early definition of future partnership and business models for adopting, supporting and sustaining common services developed for, and partly operated by, the different research communities.

Although EUDAT has initially focused on a subset of research communities, it aims to engage with other communities interested in adapting their solutions or contributing to the design of the infrastructure. Sharing our plans and gathering new input from other communities is crucial if we are to develop services that can be broadly adopted across different disciplines. Discussions with other research communities – belonging to the fields of environmental sciences, biomedical science, physics, social sciences and humanities – have already begun and are following a pattern similar to the one we adopted with the initial communities. The next step will consist of integrating representatives from these communities into the existing pilots and task forces so as to include them in the process of designing the services and, ultimately, shaping the future CDI.