TOWARDS A EUROPEAN COLLABORATIVE DATA INFRASTRUCTURE

Damien Lecarpentier a, , Mark van de Sanden b, Peter Wittenburg c

a CSC — IT Center for Science Ltd, FI-02101 Espoo, Finland - Damien.Lecarpentier@csc.fi
b SARA, Science Park 140, 1098 XG Amsterdam, The Netherlands - sanden@sara.nl
c Max Planck Institute for Psycholinguistics, PO Box 310, 6500 AH Nijmegen, The Netherlands - Peter.Wittenburg@mpi.nl

KEY WORDS: Data infrastructures, data management, High Performance Computing, Persistent Identifier, Metadata, Authentication and Authorisation Infrastructure

ABSTRACT:

The EUDAT project is a pan-European data initiative that started in October 2011. The project brings together a unique consortium of 25 partners – including research communities, national data and high performance computing (HPC) centres, technology providers, and funding agencies – from 13 countries. EUDAT aims to build 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. This short paper presents the achievements of the project during its first year and describes the services that have been chosen to meet the requirements of the initial research communities involved in the project.

1. INTRODUCTION

In recent years significant investments have been made by the European Commission and European member states to create a pan-European e-Infrastructure supporting multiple research communities. As a result, a European e-Infrastructure ecosystem is currently taking shape, with communication networks, distributed grids and HPC facilities providing European researchers from all fields with state-of-the-art instruments and services that support the deployment of new research facilities on a pan-European level. However, the accelerated proliferation of data – newly available from powerful new scientific instruments, simulations and the digitization of existing resources – has created a new impetus for increasing efforts and investments in order to tackle the specific challenges of data management, and to ensure a coherent approach to research data access and preservation.

Although some solid experience exists in Europe in dealing with data infrastructures, the current data landscape is still fragmented, with most initiatives addressing the needs of a specific discipline or community. This has resulted in increasing diversity with respect to data architectures, organizations, formats and semantics. Issues related to integration and the interoperability of existing data infrastructures are a growing concern. Rising costs due to the explosion of data are also threatening the financial viability of those infrastructures.

2. SHARED SOLUTIONS: THE CASE FOR CROSS-DISCIPLINARY DATA SERVICES

The way data is organized differs from one research community to the next; we must acknowledge this heterogeneity as a starting point, while at the same time looking for some degree of integration through common solutions and services where possible. 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 a Collaborative Data Infrastructure (CDI).

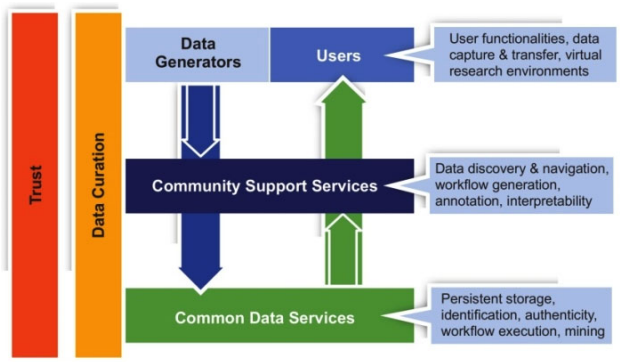

Figure 1: The Collaborative Data Infrastructure: A framework for the future © HLEG on Scientific Data, 2010

Figure 1 is taken from the *Riding the Wave* report by the High Level Group (HLEG) on Scientific Data (High Level Expert Group on Scientific Data, 2010). It illustrates the kind of collaboration required between the different parties involved in the future CDI and proposes a particular framework whereby centres offering community-specific support services to their users could rely on a set of common data services shared between different research communities.

The benefits associated with creating such a collaborative framework are many and will result in better exploitation of synergies. 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

Corresponding author.

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.

It is vitally important that large e-infrastructures meet the concrete needs of research communities, and that they are designed and set up in accordance with professional IT principles. To achieve this, there must be a close interaction between various stakeholders throughout the development process. Building the CDI requires active collaboration in particular between the communities involved in designing specific services and the data centres willing to provide generic solutions. To this end EUDAT has formed a unique consortium that brings together 25 partners, including research communities, national data and high performance computing (HPC) centres, technology providers, and funding agencies from 13 countries.

3. THE RESEARCH COMMUNITIES IN EUDAT AND THEIR DATA

Five research communities joined the EUDAT initiative at the start. They are acting as partners in the project, and have clear tasks and commitments. These initial communities come from different research areas:

- LifeWatch (Environmental Sciences – Biodiversity)
- ENES (Climate Modelling)
- EPOS (Earth Sciences)
- CLARIN (Linguistics)
- VPH (Biological and Medical Sciences)

Since EUDAT started on the 1st of October 2011, we have been reviewing the approaches and requirements of these five communities regarding the deployment and use of a cross-disciplinary and persistent data e-Infrastructure. This analysis was conducted through interviews and frequent interactions with representatives of the communities and the preliminary results are presented in this paper.

It is important to note that not only does the actual data organization vary between these communities, but there are also differences in how far individual communities have come in discussions about their data, and in the terminology that the communities use to describe their own data. Therefore we chose to use the "Digital Object Architecture" as introduced by Kahn and Wilensky (Kahn, R., Wilensky, R., 1995) as a kind of reference model and a basis from which to study the communities. For each community we looked at their general data landscapes and architectures, the types of data objects being handled, and the data flows describing how their data is manipulated. We begin here by presenting some general characteristics of the general data landscapes in each of the five communities, and then describe some of the common service requirements that were identified.

CLARIN (Common Language Resources and Technology Infrastructure) is a large-scale European initiative aiming at improving the use and availability of language resources and language technology for linguists and also other researchers from the European humanities and social sciences community. CLARIN centres form the backbone of the CLARIN research infrastructure and work with various types of data ranging from unstructured book and newspaper data to structured data, such as complex annotations, lexica and ontologies. Common types of streaming data (for example, audio and video data), along with other types of time series data (such as eye or gesture tracking and brain imaging data) are also used by language researchers. There are about 25 to 30 CLARIN centre candidates, but some heterogeneity exists between these centres in terms of data organisation. Minimal requirements (related to repositories, formats, metadata, and persistent identifiers) are being set forth for organizing the data within CLARIN centres.

ENES (European Network for Earth System Modelling) gathers together about 20 institutions working on climate modelling research. Climate change models need to account for detailed processes occurring in the atmosphere, in the ocean and on the continents. These models need to capture complex nonlinear interactions between different components of the Earth system and assess how these interactions can be perturbed as a result of human activities or natural variability. ENES works with large volumes of data generated from modelling, or collected from observation points all over the world or from satellite observations. ENES climate modelling centres use the CIM data model with an architecture separating metadata from data and using persistent identifiers. However, this model is still in the prototype phase, and the centres continue to use file systems where directory and file names include essential information about the relationships.

EPOS (European Plate Observing System) is an infrastructure for researchers in the solid Earth Sciences – studying, for example, the physical processes controlling earthquakes, volcanic eruptions, and tsunamis, as well as those driving tectonics and Earth surface dynamics. EPOS researchers work a lot with raw data streams originating from different types of sensors. Many data sensor stations used by EPOS ingest data in real time in such a way that each stream is sent to several data centres. Sensor station data is produced as a never-ending sequence of packets, while, at the data centres, data streams must be divided into files. Every centre has its own system, which means that the stored data objects are not forcibly identical. Although some work has been made to integrate the many centers, in particular within the seismology community (where there are agreements for the formats and the manner the data are federated among archives), further integration across sub-communities is needed. EPOS's intent is to virtually integrate the various data streams to offer a complete overview of the available data to users.

LifeWatch is a European initiative aiming to provide tools and services enabling researchers in biodiversity (who come from diverse disciplines) to share expertise and information remotely, through "virtual labs". Data formats vary according to the community that the data originates from. A large amount of LifeWatch data is geospatial – for example, remote sensing data from satellite imagery or real time sensor data. Other data comes from environmental and life sciences, and also from national biodiversity collections.

The VPH (Virtual Physiological Human) project aims to provide digital representations of the entire human body,

http://cmip-pcmdi.llnl.gov/cmip5/

including biological, imaging, clinical and genomic data that can be used by academic, clinical and industrial researchers to improve their understanding of human physiology and pathology, and thus find better ways of treating individual patients. Data generated and used by VPH researchers includes imaging data, and genetic data, along with simulation model data and output data.

Thus, there is considerable variation between the data landscapes in these communities, and also in the ways that researchers in these communities make use of their data. All communities rely on an infrastructure and sets of services (either existing or being developed) to support their needs. However, some of these needs are currently only partially fulfilled while at the same time some generic requirements are shared across these communities.

After several months of discussion and interaction with representatives from these communities, we have shortlisted six types of generic services that have been identified by these communities as priorities. These six services are being built jointly within the EUDAT project through multi-disciplinary task forces involving representatives from communities and data centres. The services are data replication from site to site, data staging to compute facilities, metadata, easy storage, persistent identifiers and authentication and authorization.

4. EUDAT SERVICES AND TECHNOLOGIES

4.1 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.

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.

Several pilot studies involving three of the five communities (EPOS, ENES, and CLARIN) and five data centres (JUELICH, SARA, RZG, CSC, and CINECA) have been launched and consist of replicating data sets between community and data centre sites. The first phase involves different "islands" in which a particular community is working closely together with one or several data centres to implement, test and evaluate the service. The next phase will consist of merging the islands into a single EUDAT space where communities are able to replicate digital objects (DO) to all data centres.

After investigating several technologies, EUDAT chose to use iRODS as an initial replication middleware. For the management of the persistent identifiers – which are automatically assigned to the digital objects to make it possible to keep track of all the replicas – EUDAT chose to use the handle system through the services provided by the European Persistent Identifiers Consortium (EPIC) .

http://www.pidconsortium.eu

Once users have their data replicated on the EUDAT infrastructure, we anticipate that they will want to be able to use neighbouring computing facilities to analyse this data. In particular, this is required by VPH, ENES, and EPOS as they all need to perform statistical model analysis on stored data.
Another series of pilots involving VPH, EPOS, CINECA, SARA and CSC is currently under implementation to build such a "data staging" service. Similar processes to those used in the safe replication service (involving communities and data centres working initially in separate islands) have been adopted.

Several technologies and techniques are being evaluated for staging data such as basic iRODS tools, Globus On-line, XSEDE file manage, UNICORE FTP, and Parrot. The input data sets can range from tens of gigabytes to a few terabytes in the case of special events, such as big earthquakes for EPOS. The results of the computations, which need to be ingested back into the EUDAT storage facility, are usually larger than the input data by a factor of two.

The areas of safe data replication and dynamic data replication are obviously closely connected. Figure 2 shows the different steps to be considered in a scenario where data coming from a research community (in this case EPOS) is staged from the EUDAT store to three HPC facilities (CINECA, SARA, and PRACE).

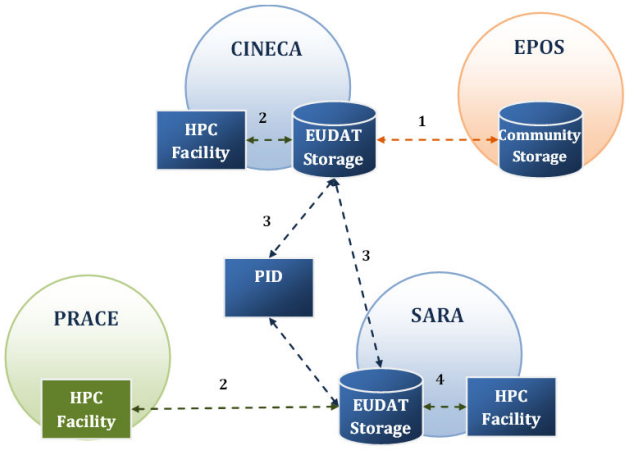

Figure 2: Utilization scenario steps for replicating and staging data from one site to another

In this scenario, data is first replicated from a community storage facility to one of the EUDAT nodes using "safe replication" solutions (1). The data is then staged to an HPC facility, either close to the EUDAT node or available outside, for example, within the PRACE infrastructure (2). The data can be replicated between two EUDAT nodes to target the required HPC facility. The corresponding PID record contains all relevant URLs of the copies (3). The replicated data is then staged to the local HPC facility and the analysis results are staged out to the original source (4). The results can then be copied back to the community storage facility.

4.2 Making Data Visible and Reusable

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. To achieve these goals, data stored on the EUDAT infrastructure must be visible, readable, understandable, and easily accessible by all. 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 metadata solutions that allow metadata from different communities to be integrated into easily searchable catalogues. To this end, an EUDAT metadata task force has been set up and is currently investigating the best way to develop a joint metadata catalogue. Using the OAI-PMH protocol and embedding domain specific metadata (as an extra available metadata record) within the OAI-PMH record is currently seen as the best option for harvesting metadata from communities and developing a joint catalogue.

The EUDAT metadata service should offer basic metadata search and browsing services to researchers looking for, or exploring, the resources from other disciplines, and could also include a "commenting" function allowing researchers to comment on the usability and/or quality of the data sets found in the catalogue. The metadata service could also be used by emerging communities that do not (yet) have their own metadata service or that are too small to provide one. Although EUDAT is in favour of open data in the scientific environment, granting access to data should ultimately remain a matter for the communities.

Thus, EUDAT's prime objectives are to build services that are shared across disciplines, and that can support cross-disciplinary data-intensive science. Despite this emphasis on commonality, some services can be tailored to a smaller subset of communities or even to individual researchers. EUDAT will host "community services", allowing user communities to use EUDAT resources to deploy and run specific services on the EUDAT infrastructure. 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.

4.3 Federated AAI and Access with SSO

In order to achieve these objectives we must work to facilitate easy access to the infrastructure and its services, while at the same time ensuring that the data is well preserved and that access rights are correctly managed. A federated authentication and authorization infrastructure (AAI) supporting single identity and single sign-on (SSO) is required.

Many communities already have AA infrastructures or rely on others provided by universities, national (academic) identity federations or other e-infrastructures (such as EGI and PRACE). The approach taken in EUDAT is to make as much use as possible of existing infrastructure. In this way EUDAT will make it possible for users to identify themselves to services in the way that they are familiar with, instead of introducing additional methods or requiring new credentials for specific EUDAT services.

Because of the many different technologies and methods available for authentication and authorization, as well as the different national legislations to be taken into account when implementing AAI solutions, this task is one of the most complex tasks involved in the project.

5. REACHING OUT TO OTHER COMMUNITIES

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.

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. 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.

Communities that are active in the field of digital cultural heritage and that are eager to take full advantage of the recent e-Infrastructure developments could also be interested in the EUDAT initiative. A recent document published by the DC-NET project (DC-NET Working Group 3: New Services Priorities, 2012) listed the priorities of such communities in terms of services. Areas, such as long term preservation, persistent identification, advanced search, user authentication and access control, are all services that could potentially be addressed by EUDAT.

6. CONCLUSIONS

After only one year of activity, significant progress has been made by EUDAT to lay out the foundations of the CDI. Yet there is still much to achieve before the CDI becomes reality and can be effectively used to support the needs of the many research communities that are facing the challenges associated with the so-called "data deluge" today.

Another important strand of activity in EUDAT focuses on the operation of the collaborative data 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 challenges are technical, but also social and organizational. Successful collaboration must be built on trust between service providers and users, and also between the researchers and disciplines themselves.

We must also plan, from the very beginning, for 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

http://www.openarchives.org/OAI/openarchivesprotocol.html

services developed for, and partly operated by, the different research communities.

REFERENCES

DC-NET Working Group 3: New Services Priorities, 2012. "Service Priorities and Best Practices for Digital Cultural Heritage". http://www.dc-net.org (accessed 19 August 2012)

High Level Expert Group on Scientific Data, 2010. Final report submitted to the European Commission "Riding the wave: How Europe can gain from the rising tide of scientific data". http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf (accessed 19 August 2012)

Kahn, R., Wilensky, R., 1995. "A Framework for Distributed Digital Object Services", Virginia, U.S.A. http://www.cnri.reston.va.us/k-w.html (accessed 19 August 2012)

ACKNOWLEDGEMENTS

This work has been supported by the European Community's Seventh Framework Programme (FP7/2007-2013) under the EUDAT Project (http://www.eudat.eu), grant agreement n° 283304.