Towards a collaborative data infrastructure for science

FEATURE | APRIL 4, 2012 | BY DAMIEN LECARPENTIER

EUDAT is a pan-European data project, bringing together a unique consortium of research communities and national data and high performance computing centers, aiming to contribute to the production of a collaborative data infrastructure to support Europe's scientific and research data requirements.

In Barcelona from 7-8 March 2012, EUDAT held its first User Forum, providing an opportunity for 18 research communities across Europe to discuss their specific data requirements and expectations. At this forum, EUDAT unveiled a set of cross-disciplinary data services, designed to service all European research communities. The deployment of each service is being coordinated by multi-disciplinary taskforces comprising representatives from user communities and data centers. EUDAT aims to deliver pilot services in 2012, with full services available to all research communities by the end of 2014.

But what exactly are these services and what benefits can user communities expect from them?

Via Laietana in Barcelona, Spain. Photo CC BY-NC-SA 2.0 MorBCN.

Shared solutions: the case for 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 basic service requirements. This commonality makes it possible for EUDAT to establish shared pan-European data services, designed to support multiple research communities.

"The way data is organized differs from one community to the next," said EUDAT scientific coordinator Peter Wittenburg, from the Max Planck Institute for Psycholinguistics at Nijmegen, the Netherlands. "EUDAT must acknowledge this heterogeneity as a starting point, while looking at the same time for some degree of integration through common solutions and services. For the CDI to succeed, an abstract architecture is required, allowing users' pre-existing data solutions to be integrated with data centers that support common data services."

Data replication and HPC access

There is strong demand among research communities 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, these services will constitute a fundamental component of the CDI:

Participants attend a panel at the EUDAT User Forum at the University of Catalunya in Barcelona, Spain. Image courtesy of Nagham Salman, Barcelona Supercomputing Center.

The 'safe replication' service will enable data replication from one site to another, for example, from a scientifically oriented community center to a data center.

The service will be flexible as well as secure," explained Mark van de Sanden, who supervises this work for EUDAT from the SARA computing center in the Netherlands. "It will allow, for example, users to ask for the creation of M replications of a data set, to be stored at different data centers for N years, with the possibility of excluding centers X to Z from the replication scheme. EUDAT has access to huge data storage facilities, provided by national data centers, and can use these to support research communities who are lacking a robust data infrastructure or who want multiple copies of data sets in geographically dispersed locations."

Another strength of the EUDAT consortium is the massive amount of computing power available at European HPC centers, most of which are members of PRACE and are among the most advanced supercomputing centers in the world.

"Once users have their data replicated on the EUDAT infrastructure, we expect they will also want to use neighboring computing capacities to analyze that data," said van de Sanden. "We are working on ways to move data between the EUDAT infrastructure and the HPC workspace."

These services will be enormously beneficial to research communities, providing a storage solution coupled with access to the most powerful computing machines in Europe. Large-scale research infrastructures (e.g. those arising from the ESFRI roadmap) will be able to use the EUDAT infrastructure to complement their own solutions, and smaller research communities will be able to rely on the EUDAT infrastructure for their data services, removing the need for large-scale capital investment in infrastructure development.

Visible, reusable data

Complex problems or 'grand challenges' increasingly require a trans-disciplinary approach and rely on data from multiple research fields. In this context, making data from various disciplines available in one collaborative infrastructure is extremely beneficial. Thus there is widespread recognition, among communities that use data and those that fund e-Infrastructures, that data federation must be improved. Improved data federation leads to better data preservation, optimized data access, and increased usability, and such improvements facilitate data reuse in new contexts, across different communities and between disciplines.

To achieve these goals, data stored on the EUDAT infrastructure must be visible, readable, understandable, and easily accessible by all, especially those coming from a discipline different than the one which created the original data.

Metadata solutions; persistent identification

Part of the challenge resides in understanding the data sets and finding good metadata solutions that allow data from different communities to be integrated in easily searchable collections. To this end, one of EUDAT's tasks is to create a catalogue that allows users to search stored data. User communities need to be heavily involved with this task, since they are the ultimate providers of metadata.

In collaboration with EPIC, EUDAT will also deploy persistent identification services, providing robust, highly available and high-performing systems that release persistent identifiers (PIDs) that, in turn, can be used within research communities and the EUDAT CDI to regulate data movement and search and query.

Tailored services

EUDAT's prime objectives are to build services that are shared across disciplines, and to 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 will also be catered to, with a 'simple store' service that allows the storage and sharing of 'small' data that are not part of official data sets or collections, but are equally important for the advancement of research.

"If EUDAT is to stimulate cross-disciplinary research, it must become a major portal for scientific data. It must offer state-of–the-art services, not only to research institutions, but also to individual researchers, since they are the ultimate users of the infrastructure," said Kimmo Koski, CSC (IT Center for Science, Finland) managing director and EUDAT project coordinator. "Services developed as part of the CDI must be user-driven, which means intense collaboration with users is absolutely crucial. We know users have high expectations from EUDAT, and we are looking forward to meeting these expectations. There will be challenges along the way, but the path becomes much clearer thanks to these strong links with user communities."

More information on EUDAT is available from: http://www.eudat.eu

Average:

Your rating: None Average: 4.3 (6 votes)

About the Author »

Damien Lecarpentier

EUDAT Project Manager

RELATED TERMS: Europe data management systems computer science

Comments

ADD NEW COMMENT

Post new comment

Subject:

Comment: *

Input format

By submitting this form, you accept the Mollom privacy policy.

SAVE PREVIEW