Towards a pan-European collaborative data infrastructure

FEATURE | NOVEMBER 9, 2011 | BY DAMIEN LECARPENTIER

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 digitization of library 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.

On 1 October 2011, the EUDAT project, co-funded by the European Commission's Framework Programme 7, was launched to target a pan-European solution to the challenge of data proliferation in Europe's scientific and research communities. The project aims to contribute to the production of a Collaborative Data Infrastructure (CDI) driven by researchers' needs, and is coordinated by CSC - IT Center for Science, Finland. It comprises 25 European partners, including data centers, technology providers, and research communities and funding agencies from 13 countries.

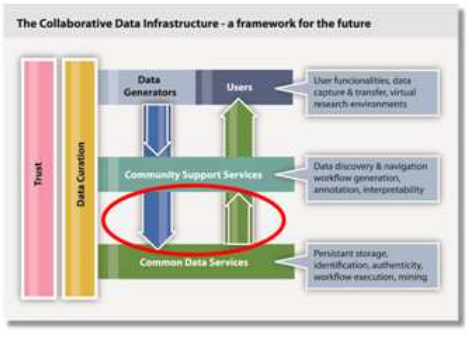
The data deluge is occuring in almost every discipline of research. Image courtesy Wikimedia.

Multi-disciplinary data management

There is already a long history of data management projects and initiatives in Europe, with several existing data infrastructures dealing with established and growing user communities: the European Center for Nuclear Research (CERN), the European Space Agency (ESA), the European Organisation for Astronomical Research (ESO), and the European Bioinformatics Institute (EBI) are examples of large-scale inter-governmental infrastructures generating and managing massive volumes of data. The numerous Research Infrastructures being created as part of the ESFRI roadmap can also be seen as data management initiatives, connecting repositories, aggregating and sharing data across national borders, and developing tools to make these data widely available to their communities and beyond.

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 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.

It is a fact that research communities from different disciplines have different ambitions, particularly with respect to data organization and content. Yet they also share basic service requirements. For example, there is a strong agreement on the importance of long-term data archiving for integrity and authenticity control, and a shared demand for data federation and services enabling discovery, access, data mining, virtual integration and curation. This commonality makes it possible to establish generic pan-European services designed to support multiple communities, as part of a collaborative framework. Building this common layer of generic and cross-disciplinary data services is precisely the focus of EUDAT.



This picture captures the kind of collaboration required between the different actors pertaining to the future Collaborative Data Infrastructure. On the top level, we have data generators and users, who rely on community support services specific to their disciplines; in turn, these community support services rely on a set of common data services that can be used by different disciplines.Image courtesy EUDAT/HLEG report.

The benefits associated with creating a Collaborative Data Infrastructure, in which research communities can rely on a set of common data services, are many and will result in better exploitation of synergies. The Collaborative Data Infrastructure will help to support the infrastructures of existing scientific communities by offering them an infrastructure on which they can rely for their more generic data services, thus allowing them to focus a greater part of their effort and investment on services that are discipline-specific. The Collaborative Data Infrastructure will also provide individual researchers, smaller communities, and projects lacking tailored data management solutions with access to sophisticated shared services, removing the need for large-scale capital investment in infrastructure development.

One of EUDAT's fundamental goals is the facilitation of cross-disciplinary data-intensive science. By providing opportunity for disciplines from across the spectrum to share data and cross-fertilize ideas, the Collaborative Data Infrastructure will encourage progress towards this vision of open and participatory data-intensive science. Increasing the scale of data federations and improving the interoperability of data objects is central to EUDAT's overall approach to the development, deployment and operation of shared services. EUDAT begins from the principle that individual and community-based data infrastructures should be federated using an architecture that fosters integration without requiring massive changes to existing and proven community-based solutions. We believe that establishing a Collaborative Data Infrastructure — using the results of organic discussion on the one hand, and advocated solutions based on concrete experiments on the other — is the best way to handle the scale and complexity of data that will be generated over the next 10 to 20 years.

First steps and later challenges

To build a sustainable data infrastructure upon which common services can be deployed for use by diverse communities, a comprehensive approach is required, including several activity strands. EUDAT is currently investigating user requirements, starting with research communities in linguistics (CLARIN), earth sciences (EPOS), climate sciences (ENES), environmental sciences (LIFEWATCH), and biological and medical sciences (VPH), which have been allocated project resources to help specify their requirements and co-design related services. This investigation will be extended to additional communities in 2012.

A second activity strand concerns the appraisal of technologies and service candidates, which involves identifying, designing and

constructing appropriate services, using existing solutions where possible. The third activity strand involves primarily the data centers and deals with the operation of the collaborative infrastructure, particularly the provisioning of 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 those marked for long-term archiving.

Building the Collaborative Data Infrastructure will not be a trivial task. It will require active collaboration between all actors, and in particular between the communities involved in designing specific services and the data centers willing to provide generic solutions. We must also plan, from the very beginning, 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. To achieve this, we first need to show that our service approach is feasible; therefore the design and deployment of early services will be critical for the success of the project. Data reusage in an open data infrastructure scenario also implies that data creators, managers and users no longer know each other: they are acting anonymously, but nevertheless must rely on each other's quality of work. Thus new mechanisms are also necessary to establish trust between all stakeholders.

EUDAT is calling for the contributions of all stakeholders interested in adapting their solutions or contributing to the design of the Collaborative Data Infrastructure. The EUDAT user forums and the Data Access and Interoperability Task Force (DAITF) already provide some opportunities to join in the discussion.

Average:

Your rating: None Average: 4 (13 votes)

About the Author »

Damien Lecarpentier

EUDAT Project Manager

RELATED TERMS: Europe data management systems interoperability standards

Comments

ADD NEW COMMENT

Post new comment

Subject:

Comment: *

Input format

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

SAVE PREVIEW