Supporting a collaborative data infrastructure — for all kinds of research

FEATURE | NOVEMBER 6, 2013 | BY ANDREW PURCELL

Last week, *iSGTW* was at the EUDAT second conference in Rome, Italy. EUDAT, which is funded under the European Commission's FP7 scheme, seeks to support a collaborative data infrastructure which will allow researchers to share data within and between communities and enable them to carry out their research effectively. Now two years into the project, the EUDAT team is ready to provide solutions that will be affordable, trustworthy, robust, persistent, and easy to use. These solutions are B2SAFE to replicate research data safely, B2SHARE to store and share long-tail research data, B2FIND to find research data, and B2STAGE to get data to computation. These are also to be followed by new services, such as Dynamic Data and Semantic Annotation, which were discussed intensively during the conference.

High-profile speakers at the EUDAT Second Conference included Richard Frackowiak, who spoke about the Human Brain Project, and Ewan Birney, who discussed the role of big data in genomics. Image courtesy Andrés Arce Maldonado/EUDAT. More photos from the event can be found on the EUDAT website, here.

Kimmo Koski, project coordinator of EUDAT, spoke during the opening plenary session at the event: "Science is global and so should be the e-infrastructures and the related services," he says. "What we do at a European level we need to link tightly to national and global activities." Koski also stressed the importance of trust between communities: "Services need to be user-driven; we need to build trust between researchers and e-infrastructure providers."

Data entropy

While giving an overview of the DataONE project in the US, Bill Michener of the University of New Mexico, neatly summed-up the challenges faced by EUDAT and the greater community. "We're working across disciplines and in large teams to deal with the grand challenges in science. However, the data that we need in order to address these challenges has many problems."

During his presentation, Michener cited research which shows that the information content of databases typically decreases over time — "data has entropy," he argues. This may be due to poor metadata that makes the data difficult to interpret, poor archiving strategies, or even data archives simply being difficult to discover. "Over five million repositories exist worldwide, but you need to know which repository is holding the data you're looking for before you can go and search through it."

Michener also cited a paper by Carol Tenopir and colleagues which shows that scientists are generally interested in sharing their data, but that they often don't know how to go about doing so. "They're particularly confused about the issue of metadata," says Michener, who laments the lack of standardized approaches among scientists for documenting their datasets. He reported on a working group within the Research Data Alliance which seeks to develop a collaborative, open directory of metadata standards. "This is a very focused goal, which should easily be accomplished within the 18-month-or-so targeted timeline," says Michener.

Not just natural science

However, discussion about data and data infrastructures at the event was by no means limited to the field of natural science. On Monday, a workshop was held on the subject of big data in the social sciences and humanities. During this workshop, several exciting case studies, demonstrating what can be achieved with big data in these fields, were discussed: from new archaeological methods to understanding historical census data, and from handling audio-visual data to understanding works of art.

Peter Doorn, director at Data Archiving and Networked Services in the Netherlands, stressed the important role played by methods in defining what constitutes big data. "It's not just about volume," he says. "Data from the past may only be quite big, but it is likely to be fuzzy and complex since it was produced without the intention of it being processed by computers." Doorn differentiates between historical data such as this, produced through mass digitization, and data in the humanities and social sciences that's "born digital", such as archives of social media posts, for instance.

Binyam Gebre, from the Max Planck Institute for Psycholinguistics, also in the Netherlands, says that massive crowdsourcing of such digital data has the potential to change how research is done in the humanities and social sciences. "It enables us to collect big data from a large number of people in a small amount of time," says Gebre. "But these methods come with their own challenges, such as data collection, management, processing and dissemination."

Equally, there are issues surrounding digitized historical data, particularly relating to the time and cost of the digitization process itself. "We need to explore the use of robotics for mass production and digitization of difficult items," says Luca Pezzati of the Italian National Institute of Optics, who also cites research by Nick Poole suggesting that it would cost around €100bn to digitize all of Europe's cultural heritage.

Growing awareness

"If you're going to tackle big data in the social sciences and humanities, then you have to deal with these issues," says Doorn. "The use of grid technology and big data in the humanities and social sciences is growing, but in general the acceptance of such technology is still low, with many researchers of the opinion that their laptop simply has enough processing power." He adds: "Grid computing, in particular, is perceived as being too complicated by most humanities and social science scholars."

"Social scientists should focus more on the analytical potential of big social data," Doorn concludes. "Few researchers are even aware of the data management issues that they have, or of the research potential of humanities and social science data. I think that we're still at a stage where we need more big-data demonstration projects."

Digital, open, and collaborative

Kostas Glinos, head of the European Commission's e-infrastructures unit, spoke during the opening plenary session of the event. He discussed the European Commission's vision of making every researcher digital by 2020, as well as the importance of supporting innovation and helping to make science in Europe more open." We believe that tomorrow's science will be increasingly open...this means open access; openness within and between disciplines; and openness between science and society," says Glinos. "All of these things need infrastructural support to make them happen."

"None of this is easy because of the explosion we're seeing in data," Glinos explains."At the start of FP7, we had just a few exabytes of data worldwide, but by the end of Horizon2020, we will probably have around 40,000 exabytes." In addition, Glinos spoke about "global connections" as one of several drivers for change within the e-infrastructures community: "As teams and research efforts become increasingly interdisciplinary, you get ever larger groups of people that need to work together online and this requires bandwidth. Equally, as more bandwidth becomes available, the groups that can work together become bigger and bigger."

A persistent issue

In addition, Glinos stressed the import role of persistent identifiers during his presenentation. The need to cite data sets, as well as just the research articles based on this data, was also discussed by Michener. Read more on this topic in our recent article: *Tracking scientific output across the web*.

United we stand

The third plenary session at the event was dedicated to the Research Data Alliance (RDA). Having been launched earlier this year, the RDA aims to accelerate and facilitate research data sharing and exchange. John Wood, EU co-chair of the RDA, spoke at length during the session about the organization's *raison d'être* : "There are big new data projects coming up," he says, citing the example of the Square Kilometre Array telescope. "We're seeing a whole new way of doing science."

"It's all about complex science and societal challenges… the point is to be able to leave a heritage to our children," explains Wood. "We need to make sure that every region doesn't just do its own thing and that we end up not being able to share between countries. If we're going to solve these societal challenges, we've got to tackle them together. We can't just do it as the US, Europe, or China: these are global problems. Either we work together, which is what the RDA is trying to achieve, or we fall apart."

Average:

Your rating: None Average: 4.7 (6 votes)

About the Author »

Andrew Purcell

Editor

Andrew Purcell is the editor of iSGTW and is based at CERN, near Geneva.

RELATED TERMS: arts and humanities EUDAT Europe grid computing RDA Research Data Alliance Rome scientific software middleware and services data management systems interoperability high-performance computing standards information services workflow management systems portals, science gateways, and hubs physics and astronomy

Comments

ADD NEW COMMENT

Post new comment

Subject:

Comment: *

Input format

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

SAVE PREVIEW