AT THE HEART OF THE HIGHER EDUCATION DEBATE

Tracking and joining billions of dots of data

16 January 2014 | By Holly Else (URL=/holly-else/1211.bio)

Work is under way to ensure that big data do not mean big headaches for researchers in the physical sciences

Source: Cern

Supersized: work at facilities such as Cern can generate huge datasets that can ‘take longer to put on a disk than it takes to generate them’

Imagine you are a physicist and you have worked almost non-stop for five days to complete hundreds of experiments at the world’s largest neutron facility. You have reams of data that you cannot wait to get home and analyse. The last few hours have you climbing the walls in frustration over how agonisingly long it takes to copy huge datasets to a portable storage device.

It is a problem that many scientists face when working at big science projects. Often, the quantities of data produced by experiments are so vast that not all the information can be collected, let alone analysed.

Rudolf Dimper is head of the technical infrastructure division at the European Synchrotron Radiation Facility – the world's brightest source of X-rays – in Grenoble, France. He said that scientists who visit the ESRF for a three-day research trip, for example, need the same time again to copy the data they generate. "Very often, big datasets take longer to put on a disk than [it takes] to generate them." This places academics under real pressure, he said.

"We have to get organised so that scientists have the means and tools to interact with massive amounts of data," Mr Dimper added.

Enter the Crisp project (Cluster of Research Infrastructures for Synergies in Physics). One strand of this project is working to help researchers manage the avalanche of data they grapple with at such facilities. It brings together 11 European research centres, such as the European Centre for Nuclear Research (Cern), the ESRF and the Institut Laue-Langevin (ILL) in Grenoble, the world's largest neutron facility. Once complete, it will offer scientists a new way of storing, organising and processing their data.

Two and a half years into its three-year life, Crisp has already provided academics visiting a number of the associated centres with a universal login that works at all the facilities. This allows researchers to merge data collected at different sites more easily than before.

By this summer, the project also hopes to offer an automated research proposal system. This would allow scientists for the first time to send research proposals to multiple facilities with one click. Currently, researchers must re-enter the details of a proposal for each facility.

Once the system is fully developed, any data collected will be stored and linked to an academic's name and publication record. The data archive could be accessed remotely, eliminating time-consuming downloads.

One academic who knows only too well the challenges of managing the mountains of data that come from large experiments is Julian Eastoe, professor of chemistry at the University of Bristol.

About once a month he and his team visit facilities such as the Diamond Light Source in Didcot, near Oxford, the ESRF and ILL. On site, they conduct experiments 24 hours a day for up to seven days at a time, which results in hundreds of data files.

Professor Eastoe can accumulate "thousands and thousands" of data files over a year. "There are only so many that one person can process, so we need a way of being able to automatically process, handle and store the data."

Having the ability to do this would be "very attractive" and would speed research, he said.

Your work stays with you

Crisp will not be the first system to link facilities and offer a universal login, says Bob Jones, head of Cern Openlab, who is involved in the project. A similar set-up exists at Cern, where a global system links about 130 remote data centres to process 210,000 DVDs' worth of experimental data each day, and users have a single account to log in.

But Crisp will take the idea a step further, he said. Researchers would be able to use the same identity throughout their career, so if an academic moves universities, or into the commercial sector, any data collected in previous positions will remain available to them.

The Crisp team is also working with separate projects that are developing facilities to manage, store and process data in other academic disciplines, with the aim of creating a common approach to identity management across science.

Dr Jones said: "Then it becomes very exciting because it can be a basis for multidisciplinary research."

Mr Dimper said that without such initiatives to organise information, dealing with the ever growing deluge of data will consume a larger proportion of researchers' time. "At one point, maybe, it becomes impossible…they will not be able to do science any more; they will be doing housekeeping of data," he said.

With this in mind, the Crisp project is also developing a system to label datasets with information that describes each file's contents. Standardising this system across facilities will allow researchers to draw data from different laboratories and will ensure that scientists can see what a file contains in the years ahead.

Look behind the desk

Mr Dimper added that the quality of data filing systems currently used in laboratories is "pretty poor". This is something that Professor Eastoe can attest to. He admitted that his team have had to redo experiments in the past because they could not find data that they had already collected.

"Maybe an experiment you did eight years ago might be useful for one tomorrow, but you have to get hold of that," Professor Eastoe said. At the moment, researchers store information about past experiments in notebooks and spreadsheets, which may not be ideal formats.

One of the greatest advantages of standardisation projects such as Crisp would be having a permanent record of experiments that could be searched in a similar way to an email in-box, Professor Eastoe said.

Dr Jones added that tying a researcher to his or her work and publications would also help to track how large infrastructure projects contribute to science. “The funding agency can get a better handle on the impact of the money they are spending.”

Broadening the system to universities and libraries is on the agenda. For the time being, however, the system is being used only in Crisp centres and others that belong to a similar initiative called Pandata.

As always, rolling the system out further will depend on funding. But Dr Jones is confident. “The European Commission has made it a key element of its plans for Horizon 2020 funding…within the next three to five years we expect to see this thing fully joined up,” he said.

holly.else@tsleducation.com