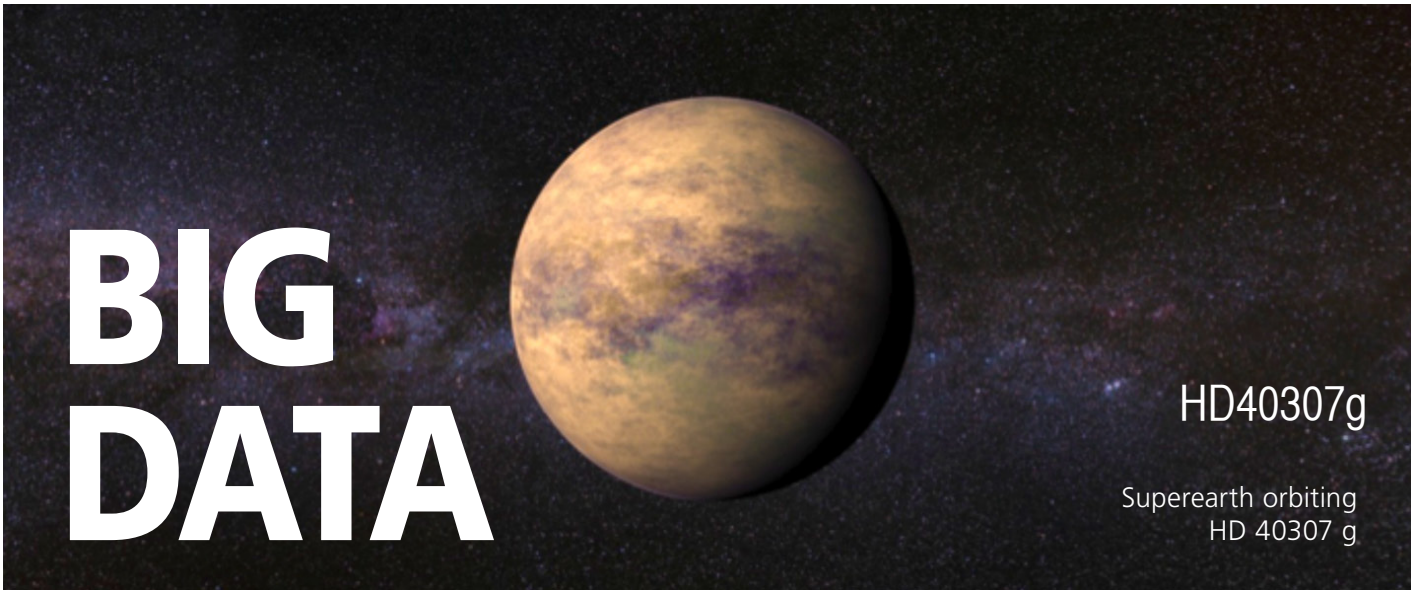
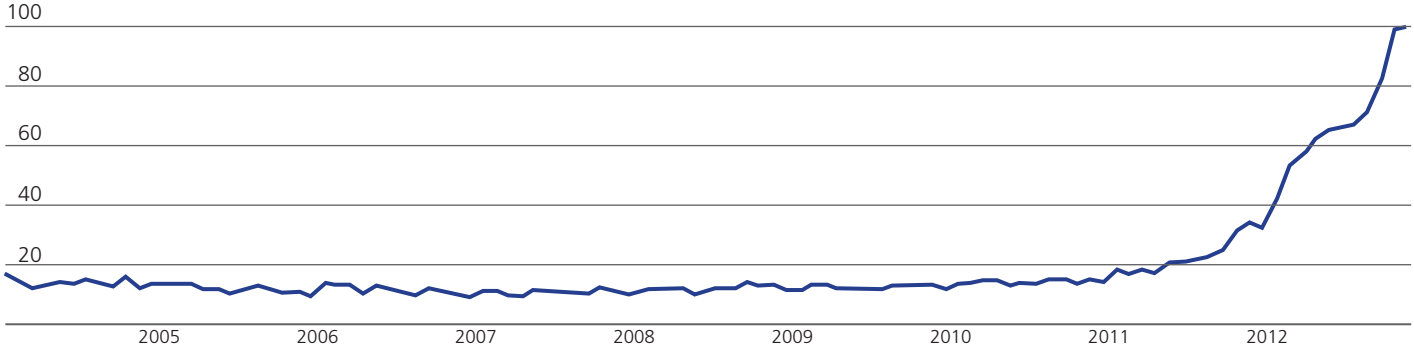
e science TALK
e-ScienceBriefings
Talking about e-science



In November 2012, Mikko Tuomi of the University of Hertfordshire and Guillem Anglada-Escude of the University of Göttingen announced the discovery of a new 'Super-Earth' – a rocky planet five times larger than ours – orbiting around the habitable zone of its parent star, where surface water would be liquid. They did this by analysing old data sets using new methods. This discovery demonstrates the importance of keeping and curating data so it can be reused later. But as science continues to produce a deluge of data, is keeping it all even viable – and will a future researcher from a different or even completely new field be able to understand it? This challenge has led to the concept of 'Big Data'.

Big Data is about the petabytes of results from particle physics, systems biology and Earth simulation science – how we deal with that volume of data and how we use it. But it's also about the variety of data being produced. Life sciences, social sciences and cognitive sciences produce data of many different types, including images, for example, as well as text-based data, so categorising and storing it all becomes a challenge. And in medicine, as data becomes obtainable at an ever-faster rate, there is an opportunity to mesh data from different source – from physiological feedback and genetic screening – to determine the course of intervention particular to individual patients.

Big Data is not just confined to science: it pervades other areas, including commerce and government. Many online retailers and search engines know so much about our interests and buying habits that they feel confident enough to tailor advertising and suggest products that we might like to buy. Some of the time, at least, they get it right. Science works differently to commerce, however. Science still needs theories to operate and to make sense of that data. One area where that distinction may be blurred is smart cities, where science and technology are employed to regulate our living processes. Already, masses of data once kept under lock-and-key is being shared by governments openly, allowing app developers, for instance, to tap into data on public transport, or refuse collection, and then present it in a useful way to the consumer.



Big data is a big deal, and more people are searching for it (Google Trends)

Data as Infrastructure

The term 'data infrastructure' was used by US President Bill Clinton in 1994, but it has actually taken some time for the idea that data is infrastructure to catch on. In the global knowledge economy, data is a basic component that leads to commerce and economic growth. Big Data owes its existence to e-infrastructures that governments have invested substantially in, but it was to facilitate the handling of Big Data (principally from particle physics at first) that e-infrastructures were built in the first place. Since that initial conception, Big Data has widened its scope to include diverse data including all manner of text- and number-based data, graphics and audio files, and increasingly metadata.

Data is infrastructure: astronomical data are encapsulated in the gears of an orrery, while data is the basis of the stock market, and of whole economies

Laurence Field, CRISP Data Management – *"Data management is a key aspect of what we're trying to do and covers many topics: data archiving, data preservation, persistent identifiers of data, data access and identity management. Physics research infrastructures have their own bespoke solutions, so we're trying to provide common solutions where possible. There are also new research infrastructures coming online that have additional requirements and we are trying to include these in those common solutions. Even though the requirements of research infrastructures may be different, we're able to identify common challenges, for example identity management, so there are always areas where we can work together."*

Metadata: Making sense of data

Data provides information and information leads to knowledge. To make sense of the data in the first place, you need to describe and manage the data. Metadata is data about data. 'Tagging' photographs, weblinks, even music in the apps we use is something that we're all used to; it helps us organise our lives. For the website delicious.com, tagging Web pages with useful category-words has helped users build their own signposted guide to the web and share those guides with others – here, metadata is a way to signpost points of interest[1]. This concept was later adopted by other social media platforms such as photosharing and blogging sites, and a particular implementation of metadata – the hashtag, used by Twitter and other microblogging sites – has entered the popular consciousness[2]. Even the file-and-folder analogy that was useful for personal computers and small networks has given way to the new paradigm of search, whether on our personal machines or 'in the cloud', helping users find files even if they can't remember what they called them.

In science, metadata is an important way of organising data to deal with it now, but it also makes it more searchable, and therefore useful (and, crucially, reusable) in the future. This is especially important when researchers, funders and governments are under pressure to demonstrate 'added value' to publicly funded research. Without metadata, the rush to make large datasets open – which is one way they are being asked to add value – would leave future researchers and those from other disciplines with a meaningless glut of information. Furthermore, solutions to the Grand Challenges of climate change, pandemic, biodiversity and energy security are expected to require innovation at the interface between traditional scientific and engineering disciplines, necessitating cross-disciplinary collaboration.

It is extremely important, therefore, that metadata is universally understood by scientists, no matter what field they work in. This means that standards are important and should be flexible enough to make sense to researchers from fields not originally anticipated to have an interest in the data, and even from fields that didn't exist when the data was collected. Collecting data costs money curated and preserved so that it can be reused to extract the maximum possible benefit from it.

"Without context, data rots," said Ross Wilkinson of the Australian National Data Service. "It needs to be integrated into other datasets and publicised." Creating a scientific 'data commons' not only aids discoverability, connectivity and value, explained Wilkinson, but is also incumbent on scientists funded by the public: "Data coming off an instrument that is publicly funded is intended for the community…it's not for the sole use of the grant holder, but is intended to be shared."

Metadata is data about data, helping scientists to find it in the future

1 /Scrupulous employment of metadata in what is termed 'white hat search engine optimisation' helps Web users find the content they're looking for, and is enshrined in principles of good Website design

2/ having initially been widely adopted by IRC earlier in internet history

EUDAT

Enshrined within European Union Law is the free movement of goods, capital, people and services between its member states. To take on the grand challenges of climate, energy, food and global health requires a fifth freedom: data. The European data infrastructure project EUDAT, which has been running since 2011, was set up in response to the recommendations of the High Level Expert Group on Scientific Data. It aims to meet the challenges of the rising tide of scientific data in Europe. The project has benefitted from being able to take lessons learned from other initiatives around the world, such as dataone in the US (since 2009) and the Australian National Data Service (ANDS, since 2007).

Peter Wittenburg, Max Planck Institute for Psycolinguistics and EUDAT Scientific Coordinator – "During the early internet it became obvious that there are cross-disciplinary common services that could be used by all research disciplines. The same email system could be used by physics and humanities researchers, for example. In the same way, we're realising now that there are common components – building blocks – for global data infrastructures. One crucial building block we're working on is a world-wide registration and resolution system for persistent identifiers. Every single piece of data would have a PID, just as computers connected to the internet have a unique IP address."

EUDAT offers common data services relevant to a wide spectrum of communities, which it has achieved in part by having worked from inception with five exemplar projects from different research areas: LifeWatch, covering biodiversity; ENES, covering climate modelling; EPOS, seismology and vulcanology; CLARIN, for linguistics, and Virtual Physiological Human (VPH) for medicine. Working with a wide range of disciplines has helped EUDAT to specify the requirements those initial projects have from a pan-European data infrastructure. This in turn has helped codesign a range of services useful to researchers across all major fields. EUDAT have also extended their reach to communities from across the biomedical, environmental, physical and social sciences and humanities.

Matthew Dovey, Programme Director, JISC – *"In some branches of applied science, being able to make accurate predictions can be of more practical importance than understanding the underlying models. For example: determining future weather patterns, or choosing between different but established medical treatments based on a patient's lifestyle. Here, Big Data can be used to identify trends and patterns with improved reliability. Ever increasing sophistication of analytical tools may even one day replace the role of the theoretical scientist in hypothesising new models. Scientists then have the task of devising experiments to challenge and test these computer-generated models."*

3/ Joint Infrastructure Services Committee
4/ Support Infrastructure Models for Research Data Management

Coordinating effort

What specific skills are required to make use of these emerging data infrastructures? The social infrastructure is something the UK body JISC[3], which looks after digital strategies and standards for post-secondary education in the UK, is looking into. The SIM4DRM project is gathering evidence at an international scale in order to develop a model of best data management practices. The project will provide a 'cookbook for all stakeholders, organisations, policy makers and funders'. Another initiative, the project Knowledge Exchange, has a vision of 'making scholarly and scientific content openly available on the internet." They are particularly interested in the areas of Open Access, Licensing, Repositories, Research Data and Virtual Research Environments. Many changes are needed to establish the open-access publishing environment including quality training/incentives (e.g. more journals for data publications) and understanding of the benefits and costs of re-using publishing and archiving data sets.

Life Sciences

One of the biggest challenges for Big Data in life sciences is the variation in terminology - even groups working on the same disease but in different model organisms will often use different terms for the same thing. BioMedBridges aims to link together datasets so that researchers can find the information they're looking for even if they don't know the terminology specific to the model organism it has come from.

Stephanie Suhr, BioMedBridges – *"BioMedBridges is about making the most efficient use of life sciences data, linking up existing resources and creating bridges between different research communities to get them to agree on common data standards and formats. This involves cultural as well as technological challenges – different communities can use different terminologies. In one project use-case we are linking diabetes and obesity-related data from human patients and from mouse models, which requires translation between the terms used by both research communities. Systematic use of extensive mouse data resources by clinical researchers will be an extremely powerful tool for new scientific discoveries and therapies."*

Standards

The first steps towards metadata standards are already being taken. One fundamental decision, announced at the EUDAT 1st conference in Barcelona by CSC Director Kimmo Koski, was that metadata in Europe would be in English. Top-down decision-making is crucial. But also recognised by EUDAT and others is that the scientists themselves must be involved in the development of standards, otherwise they could run the risk of being irrelevant to researchers. The eventual goal is to have metadata auto-generated and data tagged with standard, searchable terms as it is collected. A basic example already in place is the way some search engines allow searching for images with a particular colour, which is achieved by simple image analysis as the Web is crawled.

CLARIN – Speaking the right language?

A big step towards the 'semantic web', where information is linked together in a smart or intelligent way, is turning natural language queries, such as "list all the instances of e-science in European policy reports from 1998", into filtered searches. Commodity services that aim to provide natural language search have improved vastly over the last few years. CLARIN's aim is leverage the nuanced precision of natural language queries into accurate searches of data in the social sciences, a field dubbed eHumanities. In addition to being one of the exemplar projects for EUDAT, CLARIN has linked up with other projects working in linguistics to form the DASISH consortium, which takes its name from the projects it comprises: DARIAH, CESSDA, ESS and SHARE, as well as CLARIN. All work in complementary areas of linguistics.

One of CLARIN's tools that is already in place is the Virtual Language Observatory – a search tool linking to a vast Europe-wide corpus of linguistic datasets covering everything from psycholinguistics – how the brain learns and interprets language – to endangered languages, especially from rapidly changing regions such as the Amazon basin, including those the indigenous Trumai, Aweti, and Kamayurá peoples.

Tape library, CERN, Geneva 2, holding masses of data. By Cory Doctorow / CC BY-SA

Virtual Physiological Human

Peter Coveney, Director, UCL Centre for Computational Science – *"Big data is precisely 'where its at' for the medical domain. It's feasible to generate vast quantities of data, especially from gene sequencing – which can now be done very quickly – potentially a few minutes for an entire human genome. We're faced with the challenge and opportunity of marshalling it. Virtual Physiological Human is concerned, like many projects, with accessing patient data and using data mining and analytics techniques, but also in the business of modelling and simulation, which is used quite extensively in the physical sciences and engineering, but far less common in medicine and biology. Merging these together – e.g. a CT scan and simulation – will allow truly personalised medicine where a surgeon can be armed with the best information to make the right decision."*

5/ "Systems Biology wasn't possible until we had the data" – Ross Wilkinson, Australian National Data Service
6/ "Riding the wave: How Europe can gain from the rising tide of scientific data" EC, 2010

How Big Data is changing science itself

To Galileo, *esperienza* – what his senses could tell him about the Universe – was key to unlocking its secrets. Trusting his own experience (a word that shares the same root as experiment) over the doctrines of classical scholars, whose ideas were 'common-sense' but often scientifically inaccurate, caused a knowledge revolution and lay the foundations of the modern scientific method. Experience, refined by controlled experiment, has allowed scientists to determine the nature of reality and describe it with ever-greater clarity. Data drives decisions in science and other areas of human endeavour, challenging scientists and policymakers to refine their understanding of How Things Really Work.

Big Data in science is a challenge requiring input across and between disciplines, and even outside the realms of academic science towards the citizen scientist. But there are tremendous benefits to having so much data available to science: for one, it allows us to test and modify theories as never before, with greater accuracy and agility. Big Data, like Galileo's *esperienza*, could be more revolutionary than evolutionary, because the availability of large data sets could spark off new areas of enquiry. It is already doing so in the field of systems biology, which couldn't exist without Big Data[5].

The solutions that e-science comes up with, in terms of e-infrastructures, will help lay the foundations for environmentally responsive smart cities that rely on the Internet of Things – where every electronic device is networked to make our lives easier and the impact we have on the environment smaller. Two years ago scientists, perhaps worried about the data tsunami, were invited to ride the wave[6]. Big Data is about meeting the challenge of riding that wave and sharing how to do so with others.

For more information:

- **www.sim4rdm.eu**
- **www.icordi.eu**
- **rd-alliance.org**
- **www.clarin.eu**
- **verc.enes.org**
- **www.epos-eu.org**
- **www.eudat.eu**
- **www.dataone.org**
- **www.ands.org.au**

EGI : www.egi.eu
Real Time Monitor: rtm.hep.ph.ic.ac.uk
iSGTW: www.isgtw.org
e-ScienceTalk: www.e-sciencetalk.org
email: info@e-sciencetalk.org

Scan this QR code into your smart phone for more on this e-ScienceBriefing

e-ScienceTalk is co-funded by the EC under FP7 INFSO-RI-260733