THE IMPACT OF E-SCIENCE

K.J. Peach, e-Science Centre, CCLRC, Chilton near Didcot, Oxon OX11 0QX, UK

Abstract

Just as the development of the World Wide Web has had its greatest impact outside particle physics, so it will be with the development of the Grid. E-science, of which the Grid is just a part, is already making a big impact upon many scientific disciplines, and facilitating new scientific discoveries that would be difficult to achieve in any other way. Key to this is the definition and use of metadata.

INTRODUCTION

E-Science has become a popular research activity, particularly in some countries, covering a very wide range of projects and scientific domains. It is related to, but not synonymous with, the Grid – the Grid is an essential enabling technology for e-science, and requires many of the constructs needed by e-science (metadata, portals etc) for its successful implementation, but should not be confused with it.

While particle physics certainly has enormous computational needs that represent a significant computing challenge, it is not unique. There are equivalent challenges across the whole spectrum of research, from medicine and the life sciences, through the environment, chemistry, physics and engineering. The *scientific* problem may be different, but the underlying *methodologies* have much in common.

In this short paper, I will review some of these common issues, and indicate some of the major unresolved problems. Note that many of the themes in this paper are brilliantly illustrated in the presentation [1] by Mark Ellisman on the BIRN project

WHAT *IS* E-SCIENCE?

A simple Google™ search on the word *e-Science* yields more than 170,000 entries [2]. Even allowing for some noise, this represents a significant volume of activity. There are websites on e-science from government agencies and scientific policy makers, academics and universities, learned societies, healthcare and industry. While there is a great deal of activity, there is no agreed definition of e-science. *My* definition of e-science is "*the science that can be achieved through the use of computers to connect different sources of data about a subject, usually collected independently, to extract new information beyond that which is in each data set taken separately, to generate new knowledge and understanding*". The crucial concepts, which drive much of the developing technology, are associated with the issue of *different sources of data*, and the use of *computers* to assist the individual scientist to process the data and extract the *information* that leads to the *new* knowledge and understanding – that is, the *science*.

Achieving the e-science goals requires the development of several underpinning technologies.

- Reliable and secure networks; many applications in all domains (medicine, life sciences, engineering, environmental sciences, and the other physical sciences, as well as particle physics) have, or will have, very large data volumes that *might* need to be transported over the network, or have relatively modest volumes (images of tens to hundreds of gigabytes) that need to be transported from the repository to the application with low latency, requiring substantial bandwidth.)
- A consistent metadata description of the data – see below.
- Comprehensive portals, providing both the functionality and assurance that the individual researcher needs.

Many applications will also require Grid-scale computing, that is massive storage and large scale computation.

THE IMPORTANCE OF BEING: METADATA AND META-ANALYSIS

Meta-analysis – defined roughly as an "analysis of other people's analyses" – is becoming increasingly popular, and important. Of course, this is not new to particle physics – the Particle Data Group produces the annual compendium [3] of particle physics results – a typical result is shown in **Figure 1**

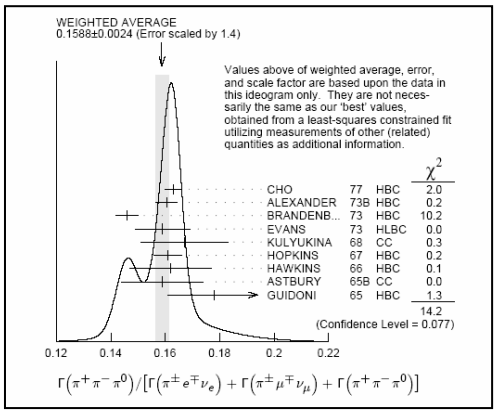

Figure 1: A meta-analysis from the Particle Data Tables.

Another example of contemporary significance in the UK is the recently published search for a correlation between the measles, mumps and rubella (MMR) vaccination and the increase in autism. The authors not

only did some new primary research, but also carried out a systematic review of other studies (a *meta-analysis*), the results of which considerably strengthened their original findings, namely that there is no demonstrable link. I suspect that the process of carrying out this meta-analysis was not particularly automatic.

There are two points to make about meta-analysis. The first is fairly obvious – while it may at first sight seem like an easy way of doing research (other people do the hard work) there is no doubt that the *confidence* in the conclusions of a serious meta-analysis over those from a single experiment is considerably enhanced; there are some very clear and straightforward conclusions to be drawn from **Figure 1**. The second point is that, without a well-defined metadata language, it is difficult to see how the process of meta-analysis can be automated, or made robust and easily auditable.

To be of general value, metadata needs to be comprehensive, standardised, verified and certified. It should also be no more complicated than necessary. It is useful to divide the metadata into three categories.

1. "geographical" metadata

This includes all of the information about the location of the data, and the characteristics of the "container". In a distributed environment, this is presumably generated, verified and certified by the system, probably the Grid.

2. "environmental" metadata

This includes all of the information about the conditions under which the data were obtained. Some of this is common to all subjects (who, when, where, how), but there is clearly some subject-dependence in the description of *what* the data describe.

3. "structural" metadata

This is subject specific, describing what the data *means*.

It is customary when discussing metadata to refer to the *ontology* – roughly defined as the set of *relevant* entities. The important feature of the ontology is that it provides a specification of all of the attributes (in this context, within a restricted domain) that are knowable. This is not quite as trivial as it looks, since the elements of the ontology may not be evident or agreed, and may (for example) depend upon scale or granularity. (If the domain of study is the migration of antelope or caribou, it may not be useful to start with atoms and molecules.) It is important that the metadata language – which can be viewed as the implementation of the ontology – be common within a domain, and where necessary consistent between domains. (In principle, the metadata vocabulary could be different, but unambiguous translation is only possible if the underlying ontology is the same.)

Many areas of science have made significant progress in the development of appropriate metadata descriptions. This is essential if data from different sources about the same, or similar, systems are to be combined automatically and transparently.

In particle physics, the definition and creation of the geographical metadata is reasonably well advanced – the successful LCG Data Challenges provide testimony. There is good progress in the definition and automatic collection of the environmental metadata (the "electronic logbook"). However, there is not yet general acceptance of the need for structural metadata, although there are discussions in the community; there is no accepted ontology. Whilst this lack of an agreed structural ontology will not prevent the higgs from being discovered, it may mean that the discovery is later than it could have been, and that the available data is not used optimally to explore its nature. This point is well illustrated by the difficulties and delays in combining the data from the four LEP experiments to search for a possible higgs signal.

PORTALS

Part of the Chambers Dictionary definition of a portal is *"a website, often incorporating a search engine, that provides access to a wide range of other sites"*. In the context of e-science, a more restrictive definition is needed. The portal provides structured access to data, applies the appropriate access and security policies, and guarantees the provenance of the data. A well-designed portal helps the researcher by providing a comprehensive suite of operations, managing the workflow and providing the researcher with the information that is needed to answer the questions posed. Before discussing the features of a portal, it is perhaps instructive to look at some websites that are not e-science portals.

The first example is Google™. There is no doubt that Google™ is an extraordinarily valuable tool – this paper could not have been written without it. However, it is not an e-science portal. To illustrate, I typed "carrot juice cures piles" into Google™, and expected to find a few articles where the words "carrot juice", "cures" and "piles" appeared in different contexts – and indeed there is at least one such site [6]. However, what was surprising was that a large number of sites did provide cures for piles involving carrot juice, for example, one [7] which advises sufferers to "drink a juice of turnip leaves, spinach, water cress and carrots (equal quantity)".

A more relevant example is the Particle Data Group website. This provides access to certified information in a structured way, but does not allow the user to manipulate the data. For example, it might be interesting to see the effect of omitting the data point with the large χ^2 from not only **Figure 1**, but also from all related plots.

On the other hand, the Durham Data base of reactions [8] is *almost* a portal. As well as having a well-defined metadata, it also gives access to the actual data. Although it does not (yet) incorporate tools to allow a meta-analysis to be done directly form the web page, it would not take much time to edit the data files provided to define the input to a further stage of analysis.

There are now many examples of data portals. The GEODISE project (Grid Enabled Optimisation and Design Search) aims [9] to "bring together and further the technologies of Design Optimisation, Computations Fluid Dynamics, Grid computation, Knowledge Management and Ontology in a demonstration of solutions to a challenging industrial problem". The MyGrid project

aims [10] to develop "open source high-level middleware to support personalised *in silico* experiments in biology on a Grid". Discovery Net (High Throughput Informatics) aims [11] to "design, develop and implement an advanced infrastructure to support real-time processing, interpretation integration, visualisation and mining of vast amounts of time critical data generated by high throughput devices". DAME [12] is an advanced Aircraft healthcare diagnosis system

CCLRC is developing a Data Portal [13] with the aim of offering a single method of browsing and searching the contents of all of the CCLRC data resources through the use of a central catalogue holding metadata about all of these resources. The structure of the metadata follows a formal scientific metadata model that is also being developed. The relationship between the portal and the metadata model is shown in Figure 2.

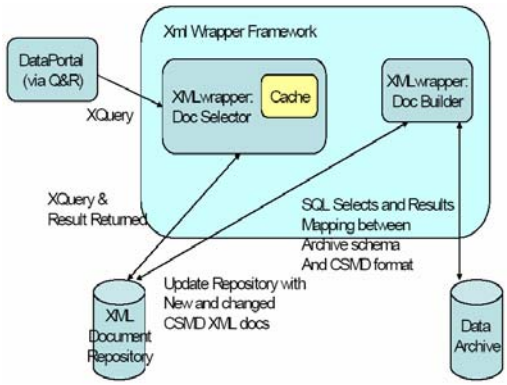

Figure 2: Portals and Metadata, from [13].

Much of the utility of e-science will depends upon the functionality of the portals that are developed. What still has to be developed is the certification of the portals – what gives the individual researcher confidence that the portal is comprehensive and accurate? The effectiveness of a gate depends upon the integrity of the gatekeeper.

E-SCIENCE AND THE GRID

E-science applications, including particle physics, require access to data, processing, analysis, simulation and visualisation. Many of these features need computing power at the leading edge of what current technology can deliver. Now, while it is true that the *cheapest* way of providing a given amount of raw compute power, and any associated data storage, is to locate both in the same place, there are some sound reasons for developing the distributed paradigm. Some of the data sources are naturally distributed, and so connecting them will require either transporting the data or using the network. The distributed computing model is inherently more resilient, and provides the potential for ensuring that critical data is always available, even if a whole region is rendered inoperable (through natural catastrophe or failure of, say, the electricity supply). The development of Grid technology will, in principle, allow massive computing resources to be harnessed in emergency. Finally, there is the psychological effect that the Grid has in engaging distributed communities and the pragmatic observation that resources might be available for local computing that would not necessarily be available otherwise.

The development of the LHC Computing Grid is a major step toward producing a practical, heterogeneous, large-scale distributed computing model able to deliver the huge amounts of computing (both CPU cycles and data storage) that the LHC needs. It is important that, in developing the LCG, general solutions are adopted so that the experience gained is available to other applications requiring high performance computing.

There are some Grid issues that need to be addressed. The comparison with the electricity grid is instructive – the "deliverable" is power (volts times amps) – and there are not many other things that need to be specified (frequency, phases, voltage). Moreover, once the Grid has delivered the power, the involvement of the generating and distributing companies is limited to sending the bill – they do not share in the "intellectual property" that their supply of power enables. However, for the computing Grid, there is no simple analogue to "power" – a computing problem is a complex mixture of CPU power, local memory, I/O capacity (to memory, cache, disk and, via the network, to other processors), and persistent storage. The issue is even more complicated because of the "negative inflation" that arises through the inexorable application of "Moore's Law" to all of these components. These issues are already affecting the distributed computing models of the present generation of particle physics experiments (BaBar, CDF and D0), and are being discussed in the context of the Memoranda of Understanding needed for the LHC. Finally, as well as the physical resource represented by the computing hardware and fabric, there is the delicate issue of the value of the real Grid resource – the data, the information derived from these data, and the knowledge abstracted form the information. (Who receives the Nobel Prize – the person who provided the data, the person who processed the data and extracted the information, or the person who realised its significance?) It is essential that data policies are defined that cover these issues.

DATA CURATION

The rapid expansion of all things "e" has created an urgent needed to address the long-term preservation of digital information. We do not know in advance whether any particular piece of information has any long-term value, but we do know that a great deal of digital information is already effectively lost – stored on media that can no longer be read.

To illustrate the problem, the first website returned by the Google™ search for "*e-science*" is shown in **Figure 3**. Now, I do not know if the information contained in this website was really important, but (given the Google™ search algorithm) at some time many other authors of websites thought that it was. The issue is that, now, it is very difficult to judge – the information that it contained

has vanished from the web. (At least those responsible for this site had the courtesy to say "goodbye" – many just stop supporting the site.)

Figure 3: The first website returned by the Google™ search for "*e-science*"

The first question to be addressed is "what to preserve?" – raw data, reconstructed data and associated montecarlo, selected data (skims or DSTs), n-tuples, distributions, collaboration notes, or final publications. The second question is "for how long?" – 5 years, 50 years, 500 years or 5000 years.

The first observation is that the issue (unlike the "paper" library) is *not* the cost of physical storage – at least until Moore's Law for storage technology fails. However, the "cost of access" to the data is an issue. Preserving the data (the "bits and bytes") is a chore, but can be automated as new technology is introduced. More difficult is preserving the information contained in the data, and the knowledge derived from it. The essential requirement for both of these is that the metadata (to access the information) and the ontology (to translate this into knowledge) is stored and managed *with the data*. There is also the related issue of precisely *who* has access to the data/information/knowledge, and when.

This is not an issue while the experiments are running. However, it is an issue for scholarship – the history of the development of science – and *may* be an issue if there is a need for a re-analysis in the light of subsequent discoveries. Even if it is possible to read the data from older experiments, it is very often difficult to interpret the results because much of the "environmental metadata" has been lost.

All of this emphasises the need for the development of *data curation policies*, and the commitment to fund the consequences. The issue, not just for particle physics or even science but for society, is to identify the digital "Tablets of Stone", that is, something that is readable for hundreds or thousands of years.

E-SCIENCE IN ACTION

There are many current examples of e-science (conforming to the definition given above). While these share many of the same challenges as those facing particle physics, they also have some rather different constraints. Several of the projects in the medical domain (for example [1]) have very serious concerns for privacy and confidentiality – for example, the need to ensure that the subject cannot be identified from the data while guaranteeing its completeness and integrity. In the industrial arena, there are serious issues of security of access from a distributed enterprise. All applications require protection from malicious intrusion.

E-science is a new methodology. There is an enormous investment being made world wide, with lots of enthusiasm – which is just as well, because there are many serious issues to be addressed. Particle physics can, and must, play its part in addressing these issues, so that e-science can achieve its full potential.

ACKNOWLEDGEMENTS

I have chosen to illustrate a number of the issues by quoting real examples, chosen almost randomly. It is a pleasure to acknowledge the hard work of the many people who have developed, and are developing these applications. I would particularly like to thank Carole Goble and John Gordon for providing me with copies of presentations from which some of the examples have been adapted.

I would also like to thank the organisers of CHEP04 for the invitation to give this presentation, which has stimulated me to think about these issues, in some cases for the first time.

REFERENCES

[1] M. Ellisman, "The BIRN Project: Distributed Information Infrastructure and Multi-scale Imaging of the Nervous System", these proceedings.

[2] In a different session, A. Sutherland (Oracle, these proceedings) noted that a Google™ search on *Grid Computing* produced more than 1,300,000 entries.

[3] S. Eidelman *et al*, Phys.Lett. B592 (2004) 1; see also http://pdg.lbl.gov.

[4] L. Smeeth *et al*, The Lancet, 364 (2004), 963.

[5] A definition of Ontology plucked from the web is "the specification of a conceptualization", Tom Gruber, Stanford. Collins Dictionary definition includes "the set of entities presupposed by a theory".

[6] See http://www.earthfirstjournal.org/efj/feature.cfm?ID=204&issue=v23n5.

[7] See http://www.fatfreekitchen.com/home-remedy/hemorrhoids-piles.html.

[8] M.R. Whalley, see http://durpdg.dur.ac.uk/HEPDATA/.

[9] See http://www.geodise.org/.

[10] See http://www.mygrid.org.uk.

[11] See http://www.discovery-on-the.net/.

[12] See http://www.cs.york.ac.uk/dame/.

[13] G. Drinkwater *et al*, *"The CCLRC Data Portal"*, UK e-Science All Hands Meeting 2003.

[14] B. Matthews, S. Sufi, and K. Kleese van Dam, "*The CCRLC Scientific Metadata Model*", DL-TR-02001