E-INFRASTRUCTURE, SCIENCE DATA AND CRIS

***S C Lambert*[1]**

[1] *e-Science Centre, Science and Technology Facilities Council, Rutherford Appleton Laboratory, Harwell Science and Innovation Campus, Didcot OX11 0QX, UK*
Email: Simon.Lambert@stfc.ac.uk

ABSTRACT

Scientific research is supported by infrastructure, and e-infrastructure is one part of this. Repositories of data are a part of the e-infrastructure and have their own particular needs arising from the requirement for permanence of their data holdings. There are many threats to permanence, and there is a growing awareness of these threats and how they may be countered. Current Research Information Systems and other support to the research lifecycle, while focused on facilitating research activities in the present, will have a role in the preservation of the outputs of research into the future.

Keywords: e-infrastructure, Science data, Digital repositories, Digital preservation, Current Research Information Systems, Scientific metadata

1 SCIENTIFIC FACILITIES AND DATA IN THE CONTEXT OF INFRASTRUCTURE

In general, infrastructure supports and enables activity. In the world of modern science, large facilities are a major part of the infrastructure. Facilities such as STFC's ISIS neutron and muon source, the Central Laser Facility, and the Diamond synchrotron light source serve a wide range of user communities, both geographically and across disciplines. For example, many of the users of the Diamond facility work in the area of structural biology, but numerous other disciplines are present, from oceanography to archaeology. It is not only individual scientists who span this breadth; there is also the need to support collaboration, often on very large scale or widely geographically distributed. Here e-infrastructure, also known as cyberinfrastructure, comes into its own. The European e-Infrastructure Reflection group (e-IRG) defines the term thus: 'The term e-Infrastructure refers to this new research environment in which all researchers—whether working in the context of their home institutions or in national or multinational scientific initiatives—have shared access to unique or distributed scientific facilities (including data, instruments, computing and communications), regardless of their type and location in the world.' Another view emphasizes the unifying role of e-infrastructure, 'bridging the gaps between islands of functionality, developed for particular purposes' (PARSE.Insight Draft Roadmap, 2009).

The large facilities that comprise the most conspicuous components of the science infrastructure generate data, often on a very large scale, and increasingly there is pressure for this data to be archived and curated for the future or for other purposes. The statement of one of the UK's research funding bodies, the Natural Environment Research Council, provides an example: 'With regard to the acquisition of data, the NERC: regards datasets as a valuable resource in their own right; […] requires that recipients of NERC grants offer to deposit with NERC a copy of datasets resulting from the research supported, for use by other bona fide researchers, but without prejudice to the intellectual property rights' (NERC Data Policy Handbook, 2002).

Of course, not all data that is systematically stored originates from large facilities within the scientific infrastructure. The British Atmospheric Data Centre, based at STFC's Rutherford Appleton Laboratory, is NERC's Designated Data Centre for the Atmospheric Sciences and curates datasets from a great diversity of observations, some of it on quite a small scale. As an example, it includes data from the European Arctic Stratospheric Ozone Experiment of 1991–92: 'two CD-ROM set contains measurements made from 16 ground stations throughout Europe, flights made by the three aircraft involved in the campaign, numerous stratospheric balloons launched from Kiruna in northern Sweden and from ozonesondes from 28 European stations. In addition data from the total ozone monitoring network are included.' This is a long distance from the Large Hadron Collider at CERN but nonetheless requiring considerable effort in its collection and not reproducible.

Thus the scientific infrastructure requires the existence of a part of the e-infrastructure, namely digital repositories. The European Strategy Forum on Research Infrastructures (ESFRI) identifies digital repositories as one of its four e-infrastructure components and highlights a number of high-level requirements:

- availability;
- permanence;
- quality;
- right of use; and
- interoperability.

(ESFRI Roadmap, 2008)

Such digital repositories should not be thought of as only for raw data but potentially the outputs of every step of processing up to and including the peer-reviewed papers representing the final outcome of the work; that is to say, the 'records of science.' Whether repositories dedicated only to archiving preprints or postprints of journal papers are part of an e-infrastructure is debateable; they are often attached to particular institutes (universities etc.) and serve as a showcase for the intellectual output of the institute; or they may be associated with particular fields such as the well-known arXiv repository for physics and some other disciplines. Nonetheless, inasmuch as they all aspire to the requirements listed above, they can be regarded as a part of the wider e-infrastructure.

2 THE TEMPORAL DIMENSION: PRESERVATION OF SCIENCE DATA

The list of requirements from ESFRI shows that the geographical and cross-disciplinary spans of e-infrastructure are not the only dimensions of interest. With reference to science data, the temporal dimension also comes into play. The need for permanence is one of the key requirements that sets digital repositories apart from other components of the e-infrastructure. Whereas networks and high-performance computing are of value by serving a function at the time of use, the repositories must make a commitment into the future, to provide access, understanding, and long-term integrity of the data they hold.

What are the threats to permanence? Obviously the deterioration of physical media and the obsolescence of file formats come to mind; but for science data there is also a need to preserve semantics. Even if a data file can be interpreted as a sequence of numbers, it is still necessary to know what the numbers represent and the units of measurement, as well as perhaps additional information on how they were obtained, their accuracy, etc. Not only hardware and software change, but the environment and even what is 'common knowledge' change also; how can we ensure that the information trapped in the 'bits' remains understandable despite all these changes?

Many initiatives have been directed at this problem, including a number of projects funded by the European Commission's 7th Framework Programme. The PARSE.Insight project is one such, developing a roadmap for digital preservation in Europe, with a view to setting the agenda for the e-infrastructure in this area. It has conducted a large-scale survey among several communities of awareness, perception, and practices with respect to preservation of science data. The communities were researchers, data managers, publishers of academic journals, and funding bodies. A total of over 1700 responses were received, making it a valuable base of evidence. In addition, case studies are being conducted in a number of fields.

The results of the survey are presented in a publicly available report (PARSE.Insight Interim Insight Report, 2009). One of the key findings concerns the perception of threats to the preservation of data. Figure 1 shows the responses from researchers (from a wide range of fields) with respect to the threats. It can be seen that the majority perceive many of these threats as already 'important' or 'very important,' and that loss of understanding is already seen to be one of the leading threats.

An analysis of instances of data loss that had occurred in reality, outlined by the respondents to the survey, confirms the basis for this perception. The great majority of such instances were due to lack of sustainable hardware, software, or support of computer environments. These included obsolete media, hardware, software, and file formats, plus disk crashes and other damaged media, and were very common. However most of the other threats had already been experienced in reality by some of the respondents.

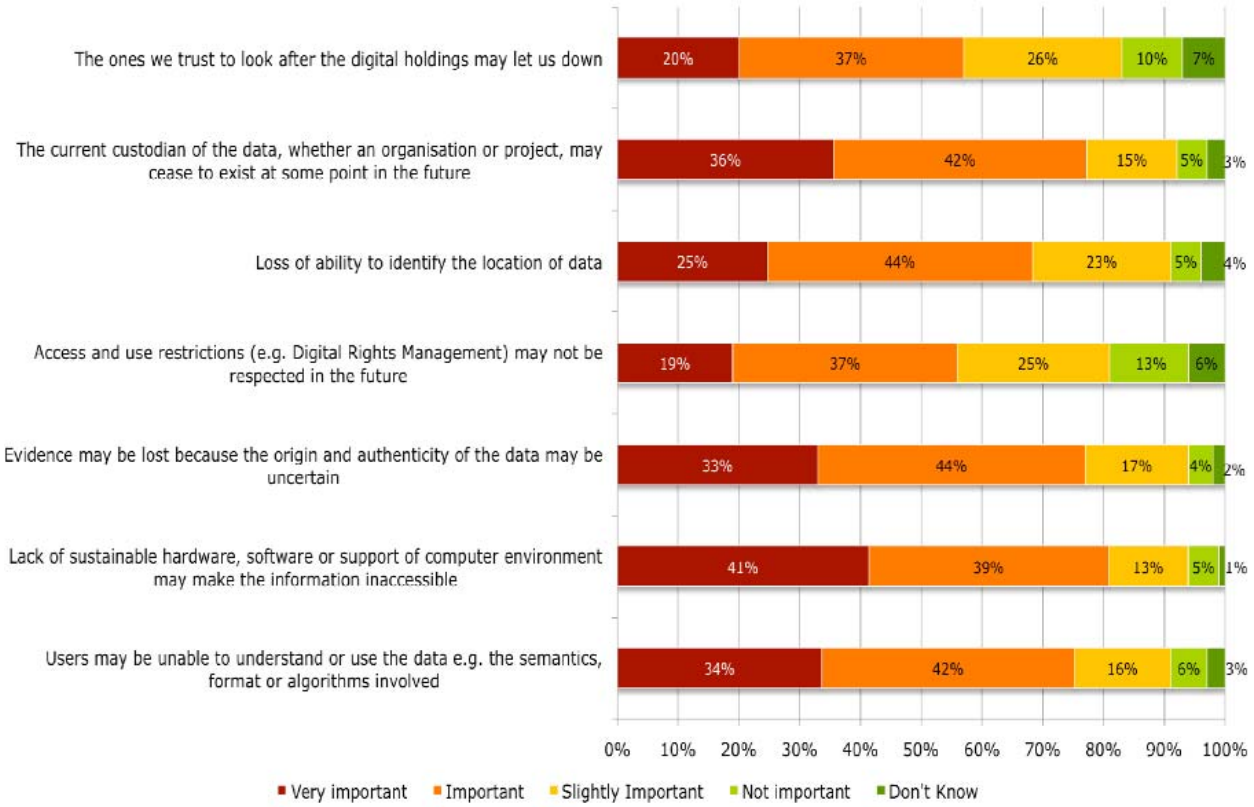

Figure 1. Responses to a question in the PARSE.Insight survey about threats to preservation of digital information. The respondents (in this case, researchers from across scientific disciplines) were asked to indicate how important they judged each threat.

It is clear that some of the threats to preservation are at the institutional level, whereas others may be addressed by providing some kind of supplementary information concerning semantics, authenticity, etc,. information that itself requires preservation, of course. The key standard in this area is the Open Archival Information System (OAIS) Reference Model, which defines the basic functional components of a system dedicated to the long-term preservation of digital information, details the key internal and external system interfaces, and characterizes the information objects managed by the system (CCSDS, 2002). This model introduces two key concepts:

- Representation Information: 'information necessary to render and understand the bit sequences constituting the Content Data Object' (Lavoie, 2004) and
- Preservation Description Information: supporting and documenting the repository's preservation processes, and including reference, provenance, context, and fixity.

With the Content Data Object itself, these make up the Archival Information Package, essentially the unit of data that is preserved.

It can be seen that there is a connection with Current Research Information Systems, as these also capture, represent, and store contextual information about the research process and its outputs. It should be understood that the word 'current' in 'Current Research Information System' means 'of current interest,' not necessarily research that is being carried out at this very moment. Thus there is a historical dimension to CRIS. The difference from the OAIS view is that the latter takes the data object as central.

3 SUPPORTING THE RESEARCH LIFECYCLE

Figure 2 depicts activities in the research lifecycle; it is these activities that the e-infrastructure should support, in a complex, distributed working environment. One of the expected benefits of e-infrastructure is to accelerate the lifecycle, thus increasing the volume of good science. This is not the only motivation, however: the curation of the outputs (data and publications) is also an essential support for the lifecycle. Indeed, background research depends on it, as does the development of proposals and experimental design. The ability to discover what relevant studies have been performed in the past is obviously crucial.

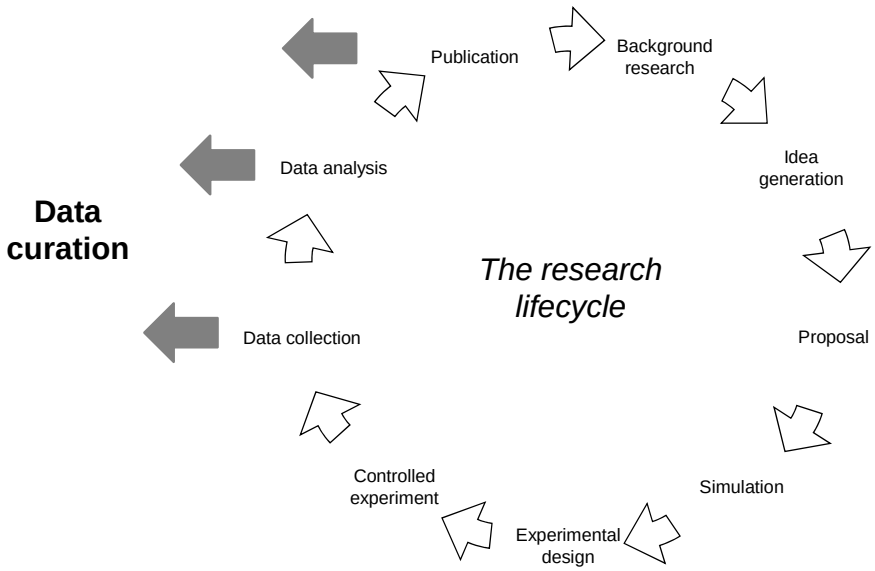

Figure 2: A depiction of the lifecycle of scientific research, adapted from Matthews (2008)

The lifecycle can be accelerated if the activities in it are integrated, saving time and effort in moving from one to the next since information has been collected at the appropriate stage in the process (a key motivation for CRIS, of course). This integration can be achieved through unifying metadata; for this purpose the Core Scientific Metadata Model has been developed at STFC. It is intended as a basis for support of the research lifecycle, a generic metadata model for all scientific applications with specialisation for each domain. It offers a common general format for scientific studies and data holdings. The basis is the concept of a *study*, which is how scientists tend to think of their work. Each study may comprise multiple investigations, which are individual experiments, simulations, or measurements. The attributes of a study may be summarised as:

- Topic (the subject of the study),
- Study description (including what the study is, who did it and when),
- Access conditions,
- Data description (organisation of the data into datasets and files),
- Data location,
- Related material (including references into the literature), and
- Legal note.

More information is given in Sufi and Matthews (2005). The metadata model is being used as the basis for a set of developments for unified access to STFC data resources with a common metadata catalogue database.

It is apparent that there is a certain degree of overlap and commonality among the three models: the OAIS Reference Model, particularly its Information Model; STFC's Core Scientific Metadata Model; and CERIF (Jeffery & Asserson, 2009) although they have different natures and motivations. The difference is one of perspective, and it is a difference that becomes significant with the passage of time. Table 1 illustrates this at a high level of abstraction.

Table 1. High-level comparison of three models supporting the scientific e-infrastructure

Model	Perspective	Class of systems
CERIF	People Projects Organisations Publications, patents, products	CRIS
Core Scientific Metadata Model	Scientific studies Investigators Datasets	E-infrastructure components for the research lifecycle
OAIS Reference Model (Information Model)	Producers and consumers (of information) Functions and components of archival information systems Packaging of archival information	Trusted digital repositories

The shift in perspective that occurs as time passes is from a focus on *process* to a focus on *product*. As an illustration of what this means, consider a well-known recent paper that looks back in time to examine research being done up to forty years ago: the study of past scientific views of global warming by Peterson, Connolley, and Fleck (2008). What approach did the authors use in their task of surveying and collating research from that period? By and large, the process aspects, programmes, projects, collaborations, funding, have withered away. That is an exaggeration; in fact one project is mentioned by name, and a number of important institutes (organisation units) are referred to. But the essential basis of this paper's argumentation is the published, peer-reviewed paper, the end-product of the research machinery that supported and facilitated the work. Numerous papers are cited, compared, and linked to understand the trends in the community's thinking at that time. It is interesting to note in passing that the published papers stand for their data: there are no references to the original datasets though some mentions are made of data collection exercises of particular significance.

One can therefore envisage a situation in which the role of CRIS, and indeed e-infrastructure components founded on metadata such as the Core Scientific Metadata Model, shifts over time, as the referents of their content recede into the past. They may make a transition from being systems that provide information of current interest about the research process, to forming the basis of a record of the context of datasets, part of digital repositories that are an aspect of the long-term e-infrastructure of science. They will provide important elements of that supplementary information specified by the OAIS Reference Model, the Representation Information and Preservation Description Information, by initially creating and then maintaining rich networks of contextual information about the datasets (and publications and all the other 'records of science') that become increasingly central in the quest for permanence in digital repositories.

4 ACKNOWLEDGEMENTS

The author would like to thank Keith Jeffery, Brian Matthews, and David Giaretta of STFC and colleagues on the PARSE.Insight project.

5 REFERENCES

Consultative Committee for Space Data Systems (2002) Reference Model for an Open Archival Information System (OAIS). Retrieved from the WWW, May 5, 2010: http://public.ccsds.org/publications/archive/650x0b1.pdf.

European Strategy Forum on Research Infrastructures (2008) European Roadmap for Research Infrastructures: Roadmap 2008. Retrieved from the WWW, May 5, 2010: ftp://ftp.cordis.europa.eu/pub/esfri/docs/esfri_roadmap_update_2008.pdf.

Jeffery, K.G. & Asserson, A. (2009) CERIF-CRIS FOR the European e-infrastructure. In this issue.

Lavoie, B.F. (2004) The Open Archival Information System Reference Model: Introductory Guide. *DPC Technology Watch Series Report 04-01*.

Matthews, B.M. (2008) Metadata for Information Management in Large-Scale Science. Presentation at *MPG EScience Seminar on Metadata Infrastructures*. Berlin, Germany. Retrieved from the WWW, April 20, 2010: http://epubs.stfc.ac.uk/work-details?w=50499.

Natural Environment Research Council (2002) *NERC Data Policy Handbook Version 2.2*. Retrieved from the WWW, May 5, 2010: http://badc.nerc.ac.uk/data/NERC_Handbookv2.2.pdf.

PARSE.Insight project (2009) *Deliverable D2.1: Draft Roadmap.* Retrieved from the WWW, April 20, 2010: http://www.parse-insight.eu.

PARSE.Insight project (2009) First insights into digital preservation of research output in Europe. Retrieved from the WWW, April 20, 2010: http://www.parse-insight.eu.

Peterson, T.C., Connolley, W.M., & Fleck, J. (2008) The Myth of the 1970s Global Cooling Scientific Consensus. *Bulletin of the American Meteorological Society* 89(9).

Sufi, S. & Matthews, B.M. (2005) The CCLRC Scientific Metadata Model: a metadata model for the exploitation of Scientific studies and associated data. In Talia, D., Bilas, A., & Dikaiakos, M. (Eds.) *Knowledge and Data Management in Grids*, CoreGRID 3, Springer.