IJDC | *Peer-Reviewed Paper*

Towards the Preservation of the Scientific Memory

Brian Matthews
STFC Scientific Computing Department
Rutherford Appleton Laboratory

Shirley Crompton
STFC Scientific Computing Department
Daresbury Laboratory

Catherine Jones
STFC Scientific Computing Department
Rutherford Appleton Laboratory

Simon Lambert
STFC Scientific Computing Department
Rutherford Appleton Laboratory

Abstract

In this paper we consider the requirements for preserving the memory of science. This is becoming more challenging as data volumes and rates continue to increase. Further, to capture a full picture of the scientific memory we need to move beyond the bit preservation challenge to consider how to capture research in context, represent the meaning of the data, and how to interpret data in relation to other scientific artefacts distributed in multiple information spaces. We review the progress of scientific research into the digital preservation of science over the last decade, emphasising in particular the research and development programme of STFC. We conclude with a number of observations into the future directions of research and also the practical deployment of policy and infrastructure to effectively preserve the scientific memory.

Received 16 January 2015 | *Accepted* 10 February 2015

Correspondence should be addressed to Brian Matthews, Rutherford Appleton Laboratory, Didcot, Oxfordshire, OX11 0QX, UK. Email: brian.matthews@stfc.ac.uk

An earlier version of this paper was presented at the 10th International Digital Curation Conference.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

http://dx.doi.org/10.2218/ijdc.v10i1.361
DOI: 10.2218/ijdc.v10i1.361

Introduction

Considerable progress has made in digital preservation over the past few years, but the task of preserving digital information over the long-term remains challenging. Data volumes are ever increasing, so there is a need for preservation solutions to scale. Further, the complexity of the structural and semantic dependencies of digital data also needs to be preserved to enable reusability. These challenges have particular relevance when it comes to scientific data.

Firstly, modern instruments and experiments generate data in very large volumes and at very high rates, which makes storing, managing and accessing data difficult. Thus the cost implications of maintaining data archives for the long term can be a substantial barrier to preserving all data; for example, due to data volumes synchrotron x-ray sources do not generally guarantee to keep data beyond a limited period of a few months. Secondly, scientific data is highly specialised to its scientific domain and the techniques used to collect data. This means that data formats, vocabulary, software and methods are often particular to that domain, thus specialised knowledge is needed to handle and interpret scientific data, access to common services, such as format characterisation services, are of limited value and the reuse potential of the data may be limited to a small and specialised community. These two factors mean that the value of preserving data needs to be considered carefully.

Further, science data is rarely self-contained, but subject to interpretation in the context of its collection. A complete understanding of the data is only possible if information on its purpose, coverage, collection methodology, environment, errors and tolerances, calibrations and other information describing how and why the data was collected is also available. To maintain an understanding of the data, this contextual information needs to be recorded and made available to the reuser, and thus also be subject to preservation requirements.

Raw science data is rarely an end in itself (unlike a document, or a film, which is in a final form for presentation to the user), but rather an item which is then subject to further processing, generating “derived” or “analysed” data via the use of specialised software packages; subject to aggregation or filtering across data; used to generate visualisations; and described, discussed and conclusions drawn in both formally published (e.g. journal articles) and unpublished (e.g. reports, but also on web pages) materials. To get a complete picture of the science undertaken and to understand how conclusions were arrived at, we need to capture all these digital artefacts and the relationships between them, to form a provenance trail of the scientific outputs.

Science data collection, analysis and reporting are frequently highly collaborative activities, with distributed teams, components and information. Digital artefacts may be generated by different people and in different places, and stored and copied in different locations. There may be different attributions and rights to different parts of the record that need to be respected. Thus to maintain the context of the science, it is necessary to manage distributed digital artefacts, with varying rights, access controls and data management policies.

Managing the preservation of physical files and their bits is essential. Bit and format preservation apply to all digital objects regardless of the use of the file. This includes: maintaining persistent identity; ensuring the integrity of the object; knowing what format the object is in through characterisation; and ensuring that the format is readable.

However, to be able to use data for the intended original purpose or reuse data for a new purpose needs more than physical integrity, it requires knowledge about the data from a scientific point of view. Research data may also be in binary format and so can't be visually inspected. Consequently, there will be a need for supplementary information that is not contained within the data files themselves. Some of this context, such as experiment set-up, needs to be captured at the point the experiment was undertaken; other context, such as analysed data or a journal article, may appear months or years after the experiment was undertaken. In general, the creation of the additional context or semantic information does not happen once and then is preserved, but is added to over the lifetime of the digital object. Consequently, we need to take a whole lifecycle view on the preservation of science.

Thus we consider the preservation problem as not one of how to preserve scientific data, but rather of how to preserve the *scientific memory* in the digital era. It is thus a problem of knowledge management, as much as the technical challenge of maintaining bit identity. How this knowledge can be identified and captured is a complicated process, as it resides in a range of places for a variety of purposes.

Over the past decade, the Science and Technology Facilities Council (STFC) has developed a core infrastructure for storing, managing and archiving data for its scientific communities. However, the STFC has recognised the complexity of the problem of preserving the scientific memory and as a consequence there is an active research and development programme to investigate some of these issues in support of its data infrastructure, funded by a number of projects including CASPAR[1], SoftPres[2], ACRID[3], ODE[4], SCAPE[5], SCIDIP-ES[6] and APARSEN[7]. Whilst we fully recognise that there is much vital research undertaken elsewhere, the STFC programme forms a good summary of the requirements of, and approaches to, preserving the scientific memory.

In the rest of this paper, we consider a number of challenges required to preserve the scientific memory, describe some work undertaken at the STFC to develop these themes, and conclude with a number of outstanding areas for further research.

The Challenges of Preserving the Scientific Memory

This characterisation of the scientific memory allows us to identify the requirements of a systematic approach to its preservation and derive a number of technical challenges that need to be addressed via research. We also discuss some recent research activities at the STFC which have contributed to the development of a general approach to preserving the scientific memory.

[1] Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval (CASPAR): An EU project within the 6thFramework Programme 2006-2009. See http://www.casparpreserves.eu/

[2] SoftPres: Tools and Guidelines for Preserving and Accessing Software Research Outputs. A JISC funded project, 2008-09.

[3] Advanced Climate Research Infrastructure for Data (ACRID): A JISC funded project, 2010-11.

[4] Opportunities for Data Exchange (ODE): An EU project 2010-12. See: http://www.alliancepermanentaccess.org/index.php/community/current-projects/ode/

[5] Scaleable Preservation Environment (SCAPE): An EU project within the 7th Framework Programme, 2011-14. See: http://www.scape-project.eu

[6] Science Data Infrastructure for Preservation – Earth Science (SCIDIP-ES): An EU project within the 7th Framework Programme, 2011-14. See: http://www.scidip-es.eu/

[7] APARSEN: An EU project within the 7th Framework Programme, 2011-14. See: http://www.alliancepermanentaccess.org/index.php/aparsen/

Preservation Analysis

The case for preserving science is not entirely obvious; not all science data is equally valuable, and it may be the case that the cost of maintaining large volumes of data may outweigh the benefit derived. Thus each collection of science data needs a separate analysis of the preservation case, detailed further in Conway et al. (2011). This includes:

- **Developing a business case:** What are the costs and benefits associated with the preservation of this data? What future technological and social risks can be anticipated in preserving the science with associated costs?
- **Developing a preservation policy:** An analysis of the collection to be preserved to determine criteria for retaining artefacts, including who the target audience is and their expected level of competency ("designated community"), and how long the data should be retained.
- **Developing a preservation strategy:** A detailed description of the approach taken for preservation, including hardware and support, a replication strategy, what related information is to be collected and managed, the tools and services used, and the processes and procedures to maintain the archive.
- **Preservation watch:** The process and procedures for maintaining the accessibility and usability of the archive in the face of changes in technology and in the designated community.

The barriers and drivers to data preservation and exchange were considered in ODE in order to facilitate enhanced data sharing in the future (Darby et al., 2012; Dallmeier-Tiessen et al., 2014). In this study, a conceptual model was developed to characterize the process of data sharing and the factors that give rise to variations in data reuse. This included technical, psychological, social, organizational, legal and political components. The model was developed by a wide ranging consultation, and identified sub-models of process, context, and drivers, barriers and enablers. These provided a comprehensive description of the factors that enable or inhibit the sharing of research data. It was intended that by implementing the enablers research communities could overcome the barriers to data reuse to facilitate future research.

More specifically, consideration has also been given to scientific data management and preservation in the arena of "big science", that is in large-scale, typically multinational and long-term collaborative research programmes, such as those found in space and particle physics, and also in the use of large-scale facilities, such as neutron and synchrotron sources (Gray, Carozzi and Woan, 2012; Bicarregui et al., 2013). In these programmes, the need to care for data has been recognised; without good data management, the core science may not be done and the potential to extract the most science from the data will be missed. Such programmes do invest resources into data infrastructure to manage and distribute the data. Nevertheless, the case still needs to be made for best practice, especially for the long-term retention and reuse of data.

When it comes to a particular science data scenario, the business case still needs to be made, including a cost-benefit analysis. Cost analysis for digital preservation is reasonably well understood (e.g. Shehab et al., 2013). The benefits arising from data preservation in the scientific domain are harder to determine, as many of the benefits of freely available data are hard to measure.

The Keeping Research Data Safe (KRDS) model of benefits (Beagrie, 2011) divides the outcomes of a data curation activity between direct benefits (the positive impacts gained) and indirect benefits (the negative impacts avoided by investing in data curation). The guide then discusses how this framework might give particular outcomes, but in a fairly unsystematic manner. In SCIDIP-ES a more systematic characterisation of the outcomes has been proposed (Caruso et al., 2013). This can be combined with the KRDS approach to provide a more detailed analysis of the potential benefits accruing from the preservation of data.

The benefits can be divided two main categories: *utility* and *substitutability.* Utility factors consider the value of data for re-examination and reuse. Thus if the utility of the data is high, then the benefit of the data is high. Further, the data is more valuable if the data is desirable – that is it sought out for re-examination – especially in new contexts and situations. Data will also have more impact if it is reusable – that is presented in a manner which encourages reuse; if it is easier to comprehend and to integrate with other data, it is likely to be reused, and thus have higher utility.

Substitutability factors assess whether an alternative data set of acceptable quality can be substituted if the primary data is not available. The user may be able to replace the current data by accessing a reasonable substitute from elsewhere, which may not be the same data, but another data set from which the same information content can be produced. Alternatively, the data may be reproduced – that is the user may able to generate new data afresh at a reasonable cost. Substitutability factors are more frequent in science data than may be appreciated; if the cost of generating or finding an acceptable substitute for the data is lower than the preservation costs, or provides higher quality, then the case for preserving data is weaker.

Successful preservation is enabled only when preservation planning, monitoring and operations are put in context with institutional preservation policies. Often in digital preservation, those policies are expressed as mission statements in high-level strategic documents, which make it a challenging task to align preservation planning, monitoring and operations with them. In SCAPE three levels of policies were defined (Sierman et al., 2013) to reflect different levels of control in an organization, from strategic levels to operation levels. In order to make those policies understandable to the planning and monitoring component, an ontology was created which enables the definition of machine readable policy models.

To support detailed preservation strategy analysis, the concept of the Preservation Network Model (PNM) (Conway et al., 2011; Conway et al., 2012) has been developed by CASPAR and subsequent projects. This method is based on the OAIS model (CCSDS, 2012) and considers the dependencies between digital objects and the representation information components that give them context and how these dependencies impact the cost of preservation, and their maintenance over the long terms. PNMs have been supported by the federated preservation tools and services developed in SCIDIP-ES. However, the practical application of this technique in a variety of scenarios that are tailored to the particular needs of the domain community needs to be explored further to make it a practical approach, and also to manage different preservation strategies (e.g. emulation and migration).

Bit Preservation

Science data needs to be kept safe and accessible for the long term and at scale. This requires the management of data at the “bit level”, that is maintaining its physical integrity. Much of this is intrinsic to the good management of a data centre, with resources in place to maintain availability as part of the active use of data. This is usually known as bit preservation, and involves the following aspects.

- **Replication:** Ensuring that copies of data are maintained, including at different locations.
- **Integrity checking:** Checking the data against corruption, typically via checksums which test whether the physical bits stored have been changed.
- **Media refresh:** Moving the data periodically onto new (tape or disc) media to mitigate against the effects of physical decay of the media material; also transferring to new storage technology as physical media become obsolete.
- **Scaling:** All issues of bit preservation are subject to scaling issues; these tasks become harder was data volumes increase, both in terms of total volume of data, and also in number of data units (e.g. files) stored.

Bit preservation is not usually seen as a research challenge within the big science data domain. In a sense it is "business as usual"; Bicarregui et al. (2013) discusses how "big science" projects in particular can factor in digital preservation as a product of good data management; issues are long term resourcing and good planning, rather than specific bit preservation challenges. For "bench" science, there are subject repositories that collect data, and again it is resourcing and managing these collections which are challenges rather than bit preservation per se. Nevertheless, there are some outstanding research challenges. Scaling means that bit preservation tasks, such as file integrity checking, file format checking and verification, and media refresh may take a long time. Generating a check sum for a very large file (of 100s of Gb or larger) may take many hours and may be impractical. Experiments on scientific archives with Hadoop in SCAPE have shown that while there is utility in using such approaches to parallelise specific preservation actions, there is also a need to tailor the approach to the specific needs of the archive. The overheads of adapting a working repository within a Hadoop architecture and using legacy systems and software may overwhelm the advantages.

Stepping beyond bit preservation to the preservation of syntax, there is a need to characterise the format of data and validate whether data conforms to declared format standards, and to migrate data from obsolete to new formats, while preserving data semantics. Similarly, this is seen as within the scope of data centres, which would take advantage of general purpose characterisation tools, such as DROID[8], although many scientific formats are highly specialist and would not typically be covered by such tools.

Cataloguing, Access and Publication

In order to be discoverable, sharable and reusable, data needs to be catalogued, published and made accessible for searching and browsing. Again, this is an aspect of good practise in data centre management, and involves:

- **Persistent identifiers:** Maintaining the identifiers of artefacts over time, so that the references to those artefacts are stable over time. This ensures that the identity of artefacts can be trusted to remain constant. Note that persistent identifiers need to refer to a variety of digital objects, including software, workflows, and aggregations as well as documents and data.
- **Metadata:** Well defined metadata formats and clear and consistent descriptions of data and other artefacts are essential for its discovery and use.

[8] Digital Record and Object Identification (DROID): http://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/droid/

- **Domain-specific ontologies:** Further formal vocabularies and relationships to describe data in terms of their domain semantics.

These processes are becoming part of the normal expectation of digital repositories. Within the STFC, there has been an ongoing effort to support digital repositories and the cataloguing and publishing of information in support of the large-scale facilities operated by the STFC and others. This includes using an enterprise scale data catalogue (ICAT) as a middleware component, instantiating a well-established metadata model, with supporting tools and services for assigning DOIs, providing access, and managing data upload and download (Flannery et al., 2009; Matthews, Sufi, et al., 2009).

Preserving the Science Context

Preserving science data in context requires a broader point of view on the preservation challenge, including the collection and maintenance of information that provide insight into how the data should be interpreted, and thus preserve the scientific activity. This entails the selection, elicitation, capturing and linking the appropriate information, which could include the following:

- Information about instruments, sensors, samples, data sampling conditions, parameters measured, coverage, units and data rates.
- Information on the intention of the observation, its methodology, and the actors involved in the data collection.
- Information on the environment in which the data has been collected which has an influence on its interpretation, and calibration information on the instruments so that data can be normalised against reference measurements.
- Information on errors, tolerances and biases known to affect the data.
- Tacit knowledge concerning the science, which may be captured in laboratory notebooks, websites, blogs, social media, annotations etc.

The concept of representation information in OAIS is intended to capture the context in which data should be interpreted, and the notion of PNMs discussed above was developed to support the specification of this wider contextual information. Realising this however, has proven complex. Recently, the SCIDIP-ES project has developed an infrastructure that contains considerable support for capturing, packaging and sharing the representation information as a dependency graph (Shaon et al., 2012; Crompton et al., 2014).

Others have considered a linked data approach to support the links and dependencies between items needed to support capturing contextual information. One approach to this was taken in the ACRID project, where information about climate data was packaged into a linked data structure, using OAI-ORE[9], containing information about the observations used to collect the data as well as links to the data itself. Thus the data package can carry information about its collection context, increasing its trustworthiness (Shaon et al., 2011). This is similar to the Research Object approach discussed below.

[9] Open Archives Initiative Object Reuse and Exchange (OAI-ORE): http://www.openarchives.org/ore/

Preserving Provenance

Preserving science provenance extends the notion of the science context to cover the wider scientific lifecycle, so how the science progresses from experiment to generate intellectual outputs is recorded. Thus to record the full picture of how research results are derived, we need to preserve different types of research artefacts, for example, raw and derived data, software, workflows, visualisations, publications and also the relationships between them. To preserve the full provenance of science, we consider:

- Capturing the dependencies and relationships between artefacts generated and used in the scientific process;
- The specific preservation needs of different types of digital artefacts, including data, software, visualisation, documents, and workflows;
- Navigating through provenance structures to address particular digital artefacts in context;
- Aggregating and packaging aggregations of artefacts as digital objects in their own right.

Provenance extends the requirement to capture context to the whole lifecycle. Again, there is a need to capture networks of relationships between artefacts. This has been explored for modelling relationships between object (e.g. Groth and Moreau, 2013), but is not well supported in current preservation architectures.

Thus there is a need for networks of relationships to be captured and stored to record science research; the Research Object[10] approach, which builds on Linked Data concepts, (Bechhofer et al., 2013) is well-suited for this and has been further explored in SCAPE (Matthews et al., 2013), where a specific approach using Investigation Research Objects, tailored to the specific needs of facilities science, has been developed and used to construct Archival Information Packages.

Research Objects link data together in a provenance graph, and provide a boundary to its scope. A research artefact can be linked to a number of other research artefacts. An investigator, workflow or instrument can participate in a number of investigations; a publication may use the output of several investigations to support its results. If this is represented as a simple web of linked data, then it would be difficult to distinguish which artefacts and relationships are relevant to which research object. OAI-ORE provides a boundary to determine membership of the Research Object, which can then be assigned an identifier its own right.

Collecting information together to preserve context and provenance brings with it the need to preserve additional classes of digital artefacts, particularly workflows and software. The Workflow4Ever European project considered preserving workflows and has developed the Research Object concept to capture workflows (Belhajjame et al., 2012).

The preservation of software is also needed to capture how data is used. However, software has characteristics that make its preservation more challenging than other digital objects. Software is inherently complex, normally composed of a large number of highly interdependent components. Software is also highly sensitive to its operating environment, dependent on items including compilers, runtime environments, operating systems, documentation and the hardware platform with its built-in software stack. Preserving a piece of software thus involves preserving much of its own context.

[10] Research Object approach: http://www.researchobject.org

Handling this complexity is a major barrier to the preservation of software, so much so that the preservation of software is often seen as a secondary activity, less critical than the preservation of the data it manipulates. However, in many cases, data becomes unusable without the software to handle it and recreating software from partial information can be a near-impossible task.

Models have been developed for the systematic preservation of software. Matthews, Shaon, et al. (2009), Matthews et al. (2010) and Matthews et al. (2012) discuss the issues that arise when considering the preservation of software, including the motivations for its preservation; the complexity of software, influencing what items should actually be preserved; the different strategies that are undertaken in the preservation of software (e.g. emulation and migration); and the criteria for judging whether software has been preserved to an adequate level of quality.

Preserving the Science Memory in a Distributed Environment

Science is a collaborative endeavour, with teams of people engaged in projects, each contributing their own artefacts to the common collection, together with their views and comments. As a consequence, the artefacts may be distributed in different locations with different ownerships. Thus we need to consider:

- The location of artefacts in different locations, potentially with copies and versions of artefacts in different places;
- Maintaining a link structure across repositories in different places, which are under different jurisdictions and may change at different rates;
- Managing the trust relationships between people and organisations to provide the appropriate guarantees that there can be stability of preservation;
- Attribution and rights management so that credit can be properly assigned to contributions to the scientific activity.

The linked data approach proposed by Research Objects also works well within a distributed environment. There is no necessity for artefacts to reside in the same archive, and links can be external as well as internal.

Outstanding Challenges

Before we can provide a complete infrastructure for preserving the scientific memory, there still remain a number of areas which require further investigation. The APARSEN project in its common vision document presents an overview of broad areas of development (APARSEN, 2014); here we concentrate on some themes arising from our perspective as presented above.

Preservation Analysis

In organisations whose focus is on the creation and management of data, the business case for preservation as an ongoing activity is not yet fully accepted. As discussed earlier, there are costs and benefits associated with preserving science; the benefits in particular are not well explored. Further, while the importance of bit preservation is well-understood, the notion of "functional preservation", that is maintaining understandability of data, is still under development.

Creating human readable preservation policy is a complex and time consuming business. To be able to write effective policy the key characteristics, or significant properties, of the object(s) need to be identified and the environment required to ensure these are maintained needs to be described. This is complex for all objects; but the data within the digital file for scientific data is an area which is still new and the potential compromises not yet identified. So, for example, one may decide that having a photograph/image is acceptable in black and white (for some designated communities) as it is known that what is being lost is the definition provided by colour and this may not be vital to the information content of the image. However, does anyone know what is the equivalent situation for a specialised data file from a neutron spallation source?

Building a Preservation Infrastructure

Both SCAPE and SCIDIP-ES have built components of a preservation infrastructure. SCAPE has a collection of tools which while powerful, are not specially tailored to preserving science (Kraxner et al., 2013). SCIDIP-ES has taken an OAIS based approach and the emphasis on preserving representation information is a step in the direction of preserving more of the science context; especially when used to describe domain parameters and necessary software. However, defining and describing representation information is not straightforward even with these tools. A detailed analysis of the preservation scenario is needed, which is difficult for domain specialists, rather than information specialists, to carry out. There is a need for guidelines and processes for specific domains; European Space Agency's Long Term Data Preservation guidelines present an approach for this (LTDP, 2012), and it needs proving in other domains.

The SCIDIP-ES approach uses representation information and preservation description information to represent context. This is a powerful approach, but proves complex to manage in practice. The Research Object approach provides an intuitive model and builds on the Linked Data infrastructure. Thus an approach that combines the SCIDIP-ES approach to OAIS with Linked Data would be a strong candidate to build a preservation infrastructure that can preserve scientific memory. This approach would bring the SCIDIP-ES information model into the Research Object world, using its ontology for OAIS as a basis, and combining this with other relevant linked-data vocabularies. This Research Object view allows us to add rich science context, so that archival information packages can be generated which capture the relationships between entities rather than treat them in isolation. This Linked Data approach would also allow the tools to be more loosely coupled in a Linked Data framework, thus exposing representation information via Linked Data endpoints.

Research Objects for Provenance

There are outstanding challenges posed by the initial developments in preserving data in Research Objects. Research Objects try to encapsulate a scientific objective, bringing all the items of interest together and grouping them. This raises the issue of what constitutes a "complete" Research Object. In a particular domain, we could reasonably expect that research objects of a particular type would have particular artefacts and relationships present. This would be the output of a preservation analysis in the particular context of the domain of under study. This would allow the definition of a domain-specific Research Object template, and an assessment of the completeness of research objects to be established.

The immutability of the Research Object is not clear: there are items which are immutable, such as the experiment and associated raw data, and there are others which are extensible, such as supplementary data and publications. Further, this issue of change means that the Research Object, with its unique identifier may become so different that it needs to be considered to be a new entity with a new identifier. An example would be when the underlying experimental data is migrated from one format to another, is this same object? Should there be links to both versions even though the fact that a migration has occurred means that there was some preservation risk to the original data? Research Objects, with their notion of boundaries, are well suited to notions of versioning, where we can relate objects together as they change, thus keeping the old boundary stable.

Preserving the Science Memory in a Distributed Environment

There remain issues of trust and sustainability in a distributed architecture. If archive managers are going to link to external sources, they require some guarantees. They require that artefacts kept in other archives are stable, do not change and maintain their identity (especially in dereferencing of persistent identifiers); accurate, the information that they offer is truthful and accurate to some specified means; accessible, the rights to accessing the artefact do not change; and meaningful, are provided with sufficient context in their own right to be understood as objects of interest to science. Trust relationships need to be established between repositories to ensure these properties. There is also a need for sustainability in the long term, with due consideration for managing archive change and archive migration.

Preserving Tacit Knowledge

Most preservation approaches concentrate on capturing the explicit knowledge of the science, encapsulated in databases, file-stores, documentation, registries, ontologies etc. However, for a true understanding of why the science was undertaken, we also need implicit or tacit knowledge, which is kept informally in peoples' minds or within the dialogue which goes on between people. It uses the prior knowledge and experience of scientists, their developed intuitions, and their observations on the conduct of the experiment. This knowledge is notoriously hard to capture. It may be written in tools such as blogs, social media, Electronic Laboratory Notebooks etc.; further work is required to manage the preservation of these types of record and link them appropriately to the explicit scientific knowledge. Research in business knowledge management could be of particular use here, with its emphasis on the elicitation of tacit knowledge, using techniques such as interviews (which may include media such as video), storytelling, after action review, or communities of practice, which can then be captured and preserved with the data.

A Final Word

We wish to move from a point of view of preserving artefacts, such as documents or data, to preserving research itself. It is the knowledge of the science which makes the artefacts useful in the future, both to understand and validate the work undertaken in the past, and to give sufficient understanding of these artefacts so that they can be reused in the future. Thus we see that preservation should be seen as knowledge management. A vision of a preservation system should try capture and preserve both the explicit knowledge of the science, embodied in data, documents and their relationships, but also the implicit knowledge, trying to capture the experience and intuitions behind the decisions made in the scientific process.

References

APARSEN Project. (2014). *Report on a common vision of digital preservation: Progress to Year 3* (Deliverable D11.3). Retrieved from http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2014/06/APARSEN-REP-D11_3-01-1_1_inclURN.pdf

Belhajjame, K., et al. (2012). *Workflow-centric research objects: A first class citizen in the scholarly discourse.* Paper presented at ESWC2012 Workshop on the Future of Scholarly Communication in the Semantic Web (SePublica2012), Heraklion, Greece.

Charles Beagrie Ltd. (2011). *Guide to the KRDS benefits framework* (Version 3). Retrieved from http://www.beagrie.com/KRDS_BenefitsFramework_Guidev3_July%202011.pdf

Bechhofer, S., et al. (2013). Why linked data is not enough for scientists. *Future Generation Computer Systems, 29*(2), pp 599–611. doi:201110.1016/j.future.2011.08.004

Bicarregui, J., Gray, N., Henderson, R., Jones, R., Lambert, S., & Matthews, B. (2013). Data management and preservation planning for big science. *International Journal of Digital Curation, 8(1),* 29–41. doi:10.2218/ijdc.v8i1.247

Caruso, G., Briguglio, L., Matthews, B., Tona, C., & Albani, M. (2013). *Modelling data value in digital preservation.* Paper presented at the 10th International Conference on Preservation of Digital Objects, Lisbon, Portugal.

Consultative Committee for Space Data Systems. (2012). *Reference model for an Open Archival Information System (OAIS)* (Magenta Book CCSDS 650.0-M-2). Retrieved from http://public.ccsds.org/publications/archive/650x0m2.pdf

Conway, E., Giaretta, D., Lambert. S., & Matthews, B. (2011). Curating scientific research data for the long term: A preservation analysis method in context. *International Journal of Digital Curation, 6*(2), 38–52. doi:10.2218/ijdc.v6i2.204

Conway, E., Matthews, B., Giaretta, D., Lambert, S., Wilson, M., & Draper, N. (2012). Managing risks in the preservation of research data with preservation networks. *International Journal of Digital Curation, 7*(1), 3–15. doi:10.2218/ijdc.v7i1.210

Crompton, S., Giaretta, D., Matthews, B., Brocks, H., Engel, F., Shaon, A., & Marelli, F. (2014). SCIDIP-ES: A sustainable data preservation infrastructure to support OAIS conformant archives. In D. Katre & D. Giaretta (Eds.), *APA/C-DAC International Conference on Digital Preservation and Development of Trusted Digital Repositories* (pp. 78–86). New Delhi: Excel India. Retrieved from http://www.ndpp.in/APA-DPDTR-2014/

Darby, R., Lambert, S., Matthews, B., Wilson, M., Gitmans, K., Dallmeier-Tiessen, S., … Suhonen, J. (2012). Enabling scientific data sharing and re-use. In *Proceedings of the IEEE 8th International Conference on e-Science.* IEEE. doi:10.1109/eScience.2012.6404476

Dallmeier-Tiessen, S., Darby, R., Gitmans, K., Lambert, S., Matthews, B., Suhonen, J., & Wilson, M. (2014). Enabling sharing and reuse of scientific data. *New Review of Information Networking, 19*(1), 16-43. doi:10.1080/13614576.2014.883936

Flannery, D., Matthews, B., Griffin, T., Bicarregui, J., Gleaves, M., Lerusse, L., … Kleese, K. (2009). ICAT: Integrating data infrastructure for facilities based science. In *Proceedings of the 5th IEEE International Conference on e-Science.* IEEE. doi:10.1109/e-Science.2009.36

Gray, N., Carozzi, T.D., & Woan, G. (2012). *Managing research data in big science* (LIGO Project report P1000188). University of Glasgow. Retrieved from http://arxiv.org/abs/1207.3923

Groth, P., & Moreau, L. (2013). *PROV overview. An overview of the PROV family of documents* (W3C Working Group Note). Retrieved from http://www.w3.org/TR/prov-overview/

Kraxner, M., Plangg, M., Duretec, K., Becker, C., & Faria, L. (2013). *The SCAPE planning and watch suite: Supporting the preservation lifecycle in repositories.* Paper presented at IPRES 2013: the 10th International Conference on Preservation of Digital Objects. Lisbon, Portugal.

LTDP Working Group. (2012). *Long term preservation of Earth observation space data: European LTDP common guidelines* (ESA report GSCB-LTDP-EOPG-GD-09-0002). Retrieved from http://earth.esa.int/gscb/ltdp/EuropeanLTDPCommonGuidelines_Issue2.0.pdf

Matthews, B., Shaon, A., Bicarregui, J., Jones, C., Woodcock, J., & Conway E. (2009). *Towards a methodology for software preservation.* Paper presented at the 6th International Conference on Preservation of Digital Objects, San Francisco, USA.

Matthews, B., Sufi, S., Flannery, D., Lerusse, L., Griffin, T., Gleaves, M., & Kleese, K. (2009). *Using a core scientific metadata model in large-scale facilities.* Paper presented at the 5th International Digital Curation Conference, London, UK.

Matthews, B., Shaon, A., Bicarregui, J., & Jones C. (2010). A framework for software preservation. *International Journal of Digital Curation, 5*(1), 106–118. doi:10.2218/ijdc.v5i1.146

Matthews, B., Shaon, A., & Conway, E. (2012). How do I know that I have preserved software? In J. Delve, D. Anderson, M. Dobreva, D. Baker, C. Billenness, & L. Konstantelos (Eds.), *The Preservation of Complex Objects: Volume 1. Visualisations and Simulations*. University of Portsmouth.

Matthews, B., Bunakov, V., Jones, C., & Crompton, S. (2013). *Investigations as research objects within facilities science.* Paper presented at the 1st Workshop on Linking and Contextualizing Publications and Datasets, Valletta, Malta.

Shehab, E., Lefort, A., Badawy, M., Baguley, P., Turner, C., Wilson, M., & Conway, E. (2013). *Modelling long term digital preservation costs: A scientific data case study.* Paper presented at the 11th International Conference on Manufacturing Research, Cranfield University, UK.

Shaon, A., Callaghan, S., Lawrence, B., Matthews, B., Osborn, T., & Harpham, C. (2011). *Opening up climate research: A linked data approach to publishing data provenance.* Paper presented at the 7th International Digital Curation Conference, Bristol, England.

Shaon, A., et al. (2012). *Towards a long-term preservation infrastructure for Earth science data.* Paper presented at the International Preservation Conference 2012, Toronto, Canada.

Sierman, B., Jones, C., Bechhofer, S., & Elstrøm, G. (2013). *Preservation policy levels in SCAPE.* Paper presented at IPRES 2013: the 10th International Conference on Preservation of Digital Objects, Lisbon, Portugal.