Moving from a scientific data collection system to an open data repository.

Tom Griffin, Brian Matthews, Alistair Mills, Sri Nagella, Arif Shaon, Michael Wilson, Erica Yang
STFC Rutherford Appleton Laboratory, UK

When many organisations operating open repositories for publications are considering moving to support data management in line with recent changes to the policies of funding bodies, this paper describes the reverse change, where an organisation operating an open data collection system has extended it to support data preservation with links to a publication repository. The lessons learned when moving from a data collection system to an open repository should be useful to those considering the move the other way.

The Science and Technology Facilities Council (STFC) operates large national scientific facilities (e.g. ISIS[1] neutron source, Central Laser Facility and Diamond synchrotron) for UK scientists and provides access for UK scientists to large international facilities (e.g. CERN, ESO) and satellites. These facilities contain instruments which produce large amounts of data which have been made available for many years to the scientists who have performed experiments for analysis in their own institutions.

One of these facilities is the ISIS neutron source which is a large £500 million microscope used to find the structure and properties of chemicals to make new materials, medicines etc.. which has been operating since 1984. The particle accelerator collides its proton pulses at one of the two tungsten targets which generate the neutrons that are then directed to a suite of instruments, each optimised to explore different properties of materials. Inside the instrument experimental scientists position a material for investigation. Neutrons travel into the material and are detected when they come out. The directions in which the neutrons emerge tell us about the arrangement of the atoms inside. The amount of energy lost by the neutrons as they travel through the material tells us about the atomic dynamics. An understanding of why individual substances behave as they do is fundamental to the development of new materials with properties tailor-made to their application.

ISIS instruments produce 2-120 files of raw and calibration data per experiment in NeXus[2] or RAW format, which is analysed in Mantid[3] and other software to produce molecular structures in 5 different structure formats (e.g. Cambridge Structural Database[4]). The 834 experiments undertaken in 2009 produced 0.5 million files, containing 0.5 TB of data. Over 25 years ISIS has produced 8 million files in 250,000 experimental datasets archived as 8TB of data, of which less than 100 datasets are commercial and therefore have constraints on making them available for re-use. Data up to five years old is checked occasionally but older data is effectively dormant – we want to change this in order for the UK to get more value from the data, and its investment in the facility.

The main reasons for preserving the data are:

1. For scientists to access their data from facilities;
2. Validation of scientific results by other scientists;
3. Re-use of data in meta-studies to find hidden effects/trends;
4. To test new theories against past data;
5. To do new science not considered when data was collected without repeating experiment.

[1] ISIS is a the name of the neutron facility, it is not an acronym.
[2] http://www.nexusformat.org/
[3] http://www.mantidproject.org/Main_Page
[4] http://www.ccdc.cam.ac.uk/products/csd/

Raw data is produced by the instruments in conjunction with calibration information (e.g. instrument temperatures and settings) which are combined to form calibrated data. This is analysed by the experimental scientists in various ways, to derive various measures of the material, which are usually put together to form a model of the structure and properties of the material, which are written up in an academic publication, and which are deposited in one of the chemical structure databases (e.g. CSD). Figure 1 illustrates this process, showing the different forms of data involved, and particularly the many forms of derived data which may be produced.

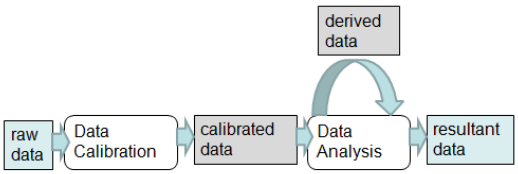

Figure 1: The flow of data from a scientific experiment

For many years STFC operated a data collection and archive service to achieve only the first of these reasons. The data was archived in a distributed file system at four separate locations to manage failure risks, and experimental scientists could gain access to their data by entering the "run number" into the web page shown in Figure 2, left where they could download the data. There were no links between the raw data, the experimental proposal, the resultant data submitted to chemical structure databases or the academic publication. There were also no preservation actions taken on the archive such as fixity checks or watching for external events, such as format obsolescence.

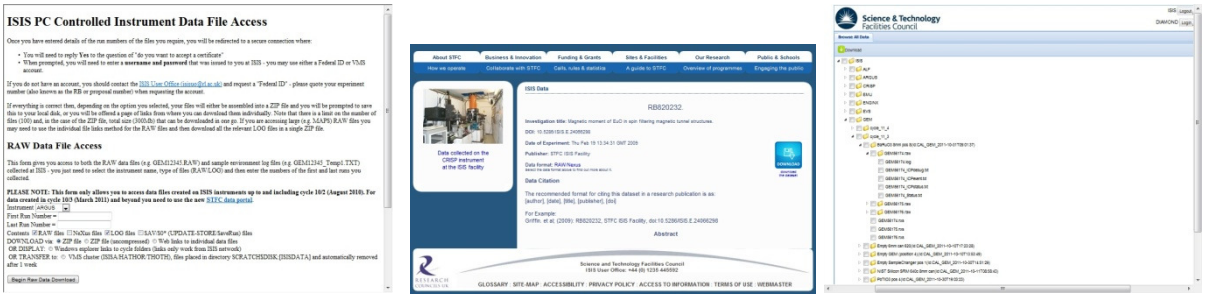

Figure 2: The user interface of the old scientific data collection system (left), landing page for a data DOI (centre) and the new iCAT data preservation infrastructure (right).

In recent years three main advances have been made to the e-infrastructure. Firstly, to also store data in a preservation system (the Safety Deposit Box from Tessella[5]) to perform fixity checks. Secondly, to publish DataCite Digital Object Identifiers (DOI) which supports both the citation of data in academic publications that link to a landing page such as that shown in Figure 2, centre; and the searching of data through the DataCite metadata (Starr & Gast, 2011) search utility[6]. Thirdly, cataloguing of the data by the Core Scientific Metadata Model (Sufi *et al*, 2003) for its inclusion in the open iCAT[7] service (Flannery *et al*, 2009) so that users can search and browse for data using the TopCat user interface shown in Figure 2, right. These changes have resulted in a data collection and preservation architecture shown in Figure 3, above the dotted line. They support the basic functionality of collecting and archiving data for the experimental scientists, linking to published citations and data search engine for data discovery, data access and download for analysis and

[5] http://www.digital-preservation.com/wp-content/uploads/SDB4.pdf
[6] http://search.datacite.org
[7] http://code.google.com/p/icatproject/

verification. These functionalities reliably support the first reasons for preserving the data, and the accessibility required for the second, but they do not provide the flexibility and support required for users unfamiliar with the data to use it for novel purposes.

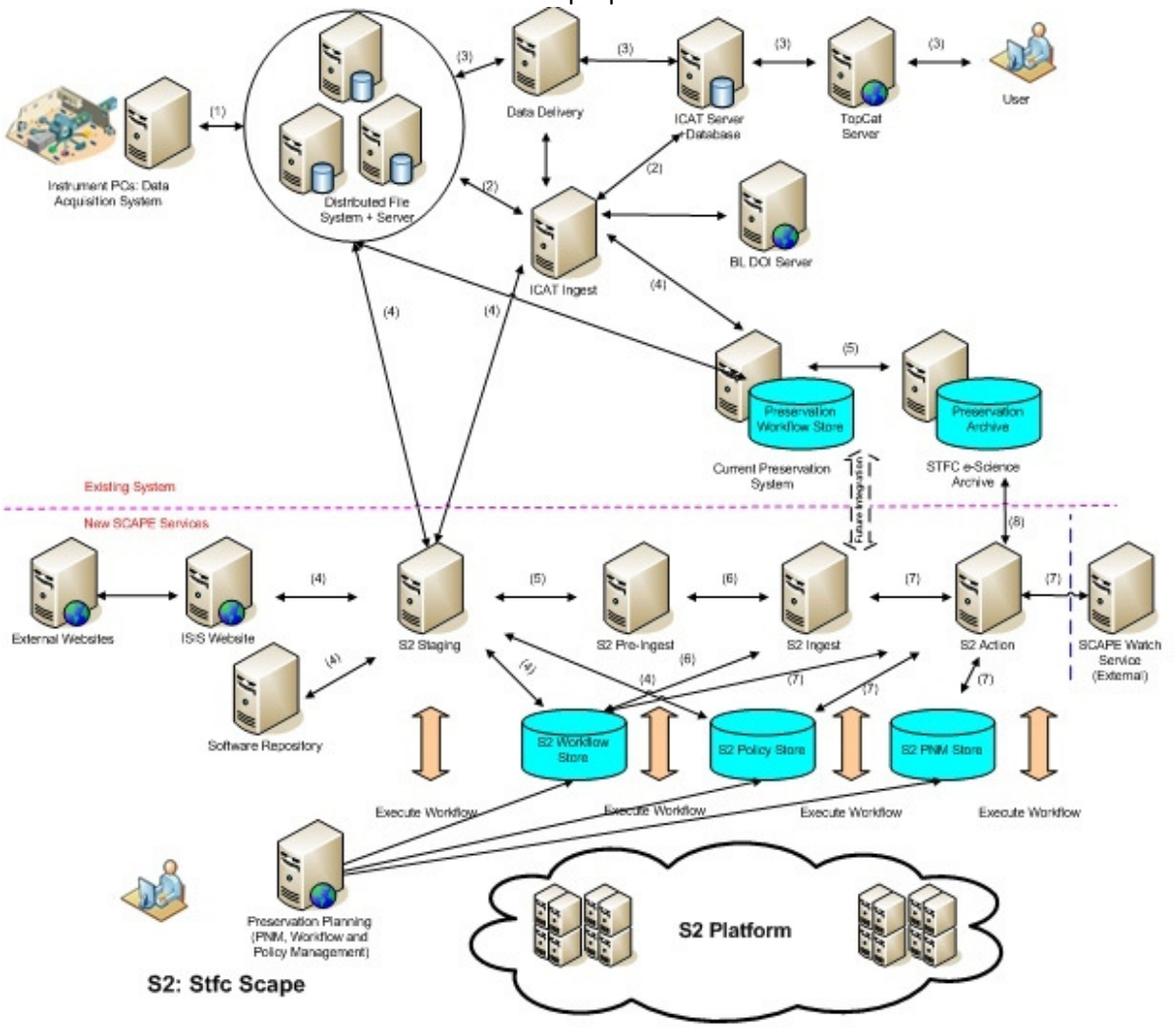

Figure 3: The architecture of the scientific data collection system (above dotted line), with the extensions for the preservation infrastructure (below dotted line).

For different user groups to use the data, the system must support multiple security and data policies (Wilson *et al*, 2011) to protect the experimental scientists, provide provenance records of the scientific analysis to support validation of published results (Yang *et al*, 2010), provide descriptions of the data and preserve the analysis software needed to re-use the data (Conway *et al*, 2011), and it must foster trust in potential users who are unfamiliar with it, so that they will use it. The proposed extensions to the preservation system shown in the lower part of figure 3 address these issues by providing watch services for external events such as format obsolescence, a policy store for data and security policies to be applied to access, by storing provenance records of data to clarify analysis validation; by providing mechanisms to support the re-use of data rather than just its archiving, and by providing ways to foster trust.

Support for the re-use of data rather than just its archiving, comes from providing support for preservation planning supported by preservation network models (Conway *et al*, 2011) which provide a basis for archiving not only the data itself, but also the software used to analyse it in a software repository, and support for storing data descriptions including links to external websites to provide context information so that scientists coming to the data whether the software used

originally still works, or not, have enough information to interpret the meaning of the data. In order to perform cross-disciplinary studies, or to test new theories with data sets, it is necessary to manipulate them into common forms equally compatible with new software. For scientists to do this requires both detailed descriptions of the semantics of the data themselves, and broad descriptions of their context to provide hooks to new theories and techniques that may not have existed when the data were collected. It is unclear exactly what level of description will be needed for any potential future use the data may be put to, but the proposed approach at least supports an extensible preservation network which can link any available information which can be archived as potential future support.

Trust is fostered in two ways. Firstly, by increasing transparency through an open software, where the code can be inspected by the preservation archive, rather than by a closed proprietary system where users must take the risk on the manufacturer's warranty alone. Secondly, by supporting the complete ISO OAIS standard, so that the whole preservation system can be externally certified to comply with best practice, as an external warranty.

Conclusion

Experience of moving from a data collection service to a data preservation service has shown the need for several mechanisms to support the future discovery, access, interpretation and re-use of the data, rather than just for archiving the data for future viewing.

Acknowledgements

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 270000 to the ENSURE project, and n° 270137 to the SCAPE project.

References

E Conway; B Matthews; S Lambert; D Giaretta ; M Wilson; N Draper (2011) Managing risks in the preservation of research data with preservation networks. In Proc. 7th International Digital Curation Confernce (IDCC2011), Bristol, UK, 05-07 Dec, URL: http://epubs.stfc.ac.uk/bitstream/7171/Final_ManagingRisks_Conway_IDCC11.pdf

Flannery, D.; Matthews, B.; Griffin, T.; Bicarregui, J.; Gleaves, M.; Lerusse, L.; Downing, R.; Ashton, A.; Sufi, S.; Drinkwater, G.; Kleese, K. (2009) , "ICAT: Integrating Data Infrastructure for Facilities Based Science," *e-Science, 2009. e-Science '09. Fifth IEEE International Conference on* , vol., no., pp.201-207, 9-11 Dec. 2009, doi: 10.1109/e-Science.2009.36

Starr, J.; Gast, A. (2011) isCitedBy: A Metadata Scheme for DataCite, D-Lib Magazine, Vol. 17, Nº. 1-2, 2011.

Sufi, S.; Matthews, B; Kleese van Dam, K. (2003) An Interdisciplinary model for the representation of Scientific Studies and associated data holdings, Proc. UK e-Science All Hands Meeting '03, Sept, ISBN 1-904425-11-9. URL: http://www.eminerals.org/papers/AH_2003_metadata.pdf

Wilson, M.; Crompton, S.; Matthews, B.; Orlov, A. (2011) , "Enforcing Scientific Data Sharing Agreements," *E-Science (e-Science), 2011 IEEE 7th International Conference on* , vol., no., pp.271-278, 5-8 Dec. 2011, doi: 10.1109/eScience.2011.45

Yang, E.; Matthews, B.; Wilson, M. (2010) , "Enhancing the Core Scientific Metadata Model to Incorporate Derived Data," *e-Science (e-Science), 2010 IEEE Sixth International Conference on* , vol., no., pp.145-152, 7-10 Dec. 2010, doi: 10.1109/eScience.2010.48