

# Sheer Curation of Experiments: Data, Process, Provenance

## Abstract

This paper describes an environment for the “sheer curation” of the experimental data of a group of researchers in the fields of biophysics and structural biology. The approach involves embedding data capture and interpretation within researchers' working practices, so that it is automatic and invisible to the researcher. The environment does not capture just the individual datasets generated by an experiment, but the entire workflow that represent the “story” of the experiment, including intermediate files and provenance metadata, so as to support the verification and reproduction of published results. As the curation environment is decoupled from the researchers' processing environment, the provenance is inferred from a variety of domain-specific contextual information, using software that implements the knowledge and expertise of the researchers. We also present an approach to publishing the data files and their provenance according to linked data principles by using OAI-ORE (Open Archives Initiative Object Reuse and Exchange) and OPMV.

**Keywords:** sheer curation; provenance; data repositories; experimental data; OAI-ORE; linked data; Fedora; OPM; OPMV.

## Introduction

This paper presents the work of the BRIL (Biophysical Repositories in the Lab) project, which has been implementing a repository for the “sheer curation” of the experimental workflows of a group of researchers in the fields of biophysics and structural biology. *Sheer curation* can be defined as “an approach to digital curation where curation activities are quietly integrated into the normal work flow of those creating and managing data and other digital assets. The word sheer is used to emphasis the lightweight and virtually transparent nature of these curation activities”<sup>1</sup>. Our approach depends on the process of data capture and interpretation being embedded within researchers' working practices, so that data capture is automatic and invisible to the researcher. Sheer curation is based on the principle that effective data management at the point of creation and initial use lays a firm foundation for subsequent data publication, sharing, reuse, curation and preservation activities.

The project uses a variety of domain-specific contextual information, which is available when the data is created, to automatically capture metadata or other information about the data and workflow that would not be accessible once the material passes into in a generic preservation environment, thus adding immediate value to the raw datasets. A key point is that it is not just the individual datasets generated by these experiments, but the entire processes or workflows that represent the “story” of the experiment, to support the validation and reproduction of published results. A particular challenge is that the researchers' work takes place in local environments within the department, entirely decoupled from the repository. In meeting this challenge, the project is bridging the gap between the “wild”, ad hoc and independent environment of the researchers desktop, and the curated, sustainable, environment of, say, an institutional or subject repository.

## Sheer Curation

It is increasingly accepted that performing digital curation and preservation in the early stages of data creation is more cost-effective in comparison to the potential loss that may be incurred through the destruction of data, for example because of the need to recreate, the loss of reputation, etc. (Rumsey, 2010). On the one hand, decisions taken during the early stages of a digital object's lifecycle may have an effect upon the preservation strategies that can be applied at a later date; on

---

<sup>1</sup> [http://en.wikipedia.org/wiki/Digital\\_curation#Sheer\\_curation](http://en.wikipedia.org/wiki/Digital_curation#Sheer_curation), accessed 6<sup>th</sup> September 2011

the other hand, if the digital objects are being preserved so that they can be reused in an informed manner, account has to be taken of the different practices of researchers across disciplines and the different natures of the data they create or gather, (Borgman, 2007, RIN 2008). In recent years there have been renewed efforts among digital curators to establish new methods of performing digital curation from the outset. The limiting factor among researchers is that they have time to meet only their own immediate, short-term requirements, and – even when they want to – they often do not have the resources, in terms of time, expertise or infrastructure, to spend on making their datasets reusable by others (Shearer, 2009; Key Perspectives Ltd, 2010).

One approach to this has been termed *sheer curation*<sup>2</sup>, which describes approaches in which digital curation activities are integrated into the workflow of the researchers creating or capturing data. The word “sheer” here is used in the sense of “lightweight and virtually transparent” way in which these curation processes are integrated with minimal disruption to their normal working practices<sup>3</sup>. This approach depends on the data capture or ingest process being embedded within the researchers' working practices, so that data capture is automatic and invisible to the researcher. Sheer curation is based on the principle that effective data management at the point of creation and initial use lays a firm foundation for subsequent data publication, sharing, reuse, curation and preservation activities.

For example, the UK Digital Curation Centre's SCARP project<sup>4</sup> (during which the term *sheer curation* was coined) carried out a number of case studies in which digital curators engaged closely with researchers across a range of disciplines in order to improve data *curation* practice through a close understanding of *research* practice (Lyon et al., 2010; Whyte et al., 2008). Other examples, this time from the business world, is given by (Curry et al., 2010), who discuss the role of sheer curation in the form of distributed, community-based curation of various types of enterprise data. Sometimes the concept is contrasted with *post hoc* curation, where the curation activities start after the period during which the digital objects are created and primarily used.

For work described here the approach was slightly different, embedding expertise about research practice – gained from researchers through a series of interviews – within the software that captures the data and ingests it into the archive. As digital objects are created during the process of an experiment, this software uses a variety of domain-specific contextual information associated with these objects to interpret them and, importantly, their relationships to other objects, thus capturing the experimental process. This contextual information is available when the data is created, but would no longer be accessible once the material passes into a generic curation environment; nevertheless, this information may be very important for subsequent publication, sharing, reuse, and preservation of the digital objects, even if it may not directly benefit their creators and primary users.

To illustrate the problem, consider the following. The researchers with whom we were working typically carrying out their work on a single desktop or laptop computer, and at the end of an experiment the associated data files – including all intermediate files – reside within a directory structure in a directory dedicated to that experiment. If the contents of this directory were deposited in a generic preservation environment, it would be possible to carry out a certain amount of digital preservation actions – the repository might extract metadata about the files, and might even include record the relationships represented in the directory structure – however, it would not be at all clear what all this data meant. The “story” of the experiment is represented implicitly in a variety of information such as the location of files in the directory hierarchy, metadata embedded

---

<sup>2</sup> By Alistair Miles of the Science and Technology Facilities Council.

<sup>3</sup> <http://alimanfoo.wordpress.com/2007/06/27/zoological-case-studies-in-digital-curation-dcc-scarp-imagestore/>, Accessed 3<sup>rd</sup> September 2011

<sup>4</sup> <http://www.dcc.ac.uk/projects/scarp>

opaquely in binary files, filenames, the contents of log files, and so on. However, the semantics of the collection as a whole would be lost in this *post hoc* curation model, and it is unlikely that the researcher would have the time or the ability to explain it in great detail. In the model of sheer curation that we have been following, the researchers' knowledge is elicited beforehand and encoded in software that processes the information as it is captured.

## Use cases

The research use cases that the project has been addressing are in the fields of biophysics and structural biology, a multidisciplinary area that interacts and collaborates with several research groups, both within the institution (e.g. Asthma, Cardiovascular, Cancer) and with industrial partners such as pharmaceutical companies. The research practices vary in detail, but from our use case analyses common patterns can be seen: (i) an initial stage in which raw data is captured from experimental equipment in a laboratory; (ii) various stages in which the data is processed and analysed; (iii) publication of outputs, which may include data (e.g. protein structures) as well as journal articles.

The parts of the research processes that fall under category (ii) largely involve *interactive* activities, where the researcher uses and responds to desktop-based tools, rather than the more automated workflows that are implemented using workflow engines; this is a key challenge for capturing and curating this material, as we shall see. While, for a particular type of experiment, the form of these research processes follows certain general patterns, they can be very unpredictable at a detailed level, partly as a result of their interactive nature – the flow of control often depends on the researcher's personal judgement and decisions, which are inaccessible as we are only able to monitor what actually happens in the digital environment.

A key aspect of the project throughout was the close involvement of the researcher communities. Firstly, it was necessary for the repository development team to obtain a clear and detailed understanding of the researchers' data, processes and tools, so that could implement this expertise in software capable of capturing the domain-specific metadata and other contextual information, such as data provenance. Secondly, a key idea behind sheer curation is that it should be integrated with the researchers' processes in as transparent and non-invasive a way as possible. Thirdly, services for access to and reuse of these specialised datasets need to be aligned very closely with the ways in which the communities would want to use them – a standard repository search and browse interface would not be sufficient. As a consequence, instead of a rather loose partnership between the developers and the researchers, these different areas of knowledge and expertise were closely integrated. In fact, one of the major problems encountered, which slowed the collaboration down in the earlier stages of the project, was the difficulty that the staff with ICT expertise encountered in gaining sufficient understanding of the science.

While the research processes followed by the various research groups followed a common pattern when viewed at a very high level of abstraction, as described above, there were significant differences between them, and they used different tools and dealt with different file types. Consequently, while we interviewed researchers from five separate research groups, for the implementation we focused our efforts on the use cases of two of the groups only: macromolecular crystallography and biological nanoimaging.

Macromolecular crystallography addresses the determination of the structure of large molecules, such as proteins, using X-ray diffraction. In high-level terms, an X-ray beam is directed at a crystal of the substance under investigation from many angles, resulting in a set (typically 360, although sometimes more) of diffraction images. Each image contains several hundred spots, whose location and intensity are determined, using specialised software, and then combined to produce a model of the atomic co-ordinates of the protein. This process has many steps, as well as dead ends and repetitions when analysis or processing steps do not work and need to be modified and repeated, all of which generates large numbers of interim files. While a small number of the resulting files are

published – for example, PDB files in the Protein Data Bank- the vast majority are not currently kept, or at least are not curated.

Biological nanoimaging involves the use of microscopes to capture high-resolution images of biological samples, on the one hand to carry out research into cell and tissue structures, and on the other to develop new methods of and algorithms for digital imaging and processing. The data may be pseudo-3D representations, constructed from a large number of horizontal sections, or in vivo imaging where the sample is imaged at multiple time points, as well as 2D images. The same datasets may be processed many times using different image analysis techniques, and many raw images are processed when developing new analysis tools. Again, much of the information generated in this process is not currently curated or retained.

## Challenges

Several separate environments are involved in the broader process of experiment, processing and curation. The capture of the raw data typically occurs in a laboratory environment, for example using a microscope and camera configuration, or a synchrotron and charge-coupled device, and the data is transferred to the researcher's desktop computer where all subsequent processing and analysis takes place<sup>5</sup>.

As we have seen, however, the processing environment is not at all predictable or tightly controlled. In fact the project operated across two quite distinct and independent environments, which are completely decoupled from one another. On the one hand, there is the desktop environment where the data processing and analysis takes place, typically located in the researcher's department and under the control of the researcher, subject to whatever requirements the department places on that environment. On the other hand, there is the repository environment that is managed at an institutional level. A major challenge faced by the project was to bridge the gap between the "wild", *ad hoc* and independent environment of the researcher's desktop, and the managed and curated environment of the repository.

The tools used by the researchers are typically developed by people working within the discipline, either by the researcher communities themselves, or by suppliers of laboratory equipment, and they are designed to operate in the researchers' local environment using data that is accessible via the local file system (although there are some web-based services). The tools thus are entirely outside the control or influence of the repository staff, who are thus obliged to take them as they come.

## Objectives

The project had the following broad objectives:

1. To implement a "sheer curation" environment by embedding a repository within the experimental workflows of the targeted researchers, so that, as far as possible, capture of data and metadata occurs automatically, invisibly to the researcher, and with no (or very little) change to the researchers' normal practice.
2. To manage in the repository not just individual datasets but entire experimental workflows, modelled as compound objects incorporating data, metadata and provenance information. This will make it possible to verify published results or reproduce the processing and data on which the conclusions are based and which justify them.
3. To capture automatically domain-specific metadata and other contextual information that is available at the point of data creation, but that could not be extracted later in a generic preservation environment.

---

<sup>5</sup> For practical, organisational, reasons, the project did not address the research process from the initial data capture in the laboratory, but only from the moment that this raw data was transferred to the processing environment.

As observed above, a generic preservation environment would be unable to interpret fully the context or nature of many of the files generated during processing, which would appear simply as files in a directory structure, the semantics of the collection as a whole being lost. In our sheer curation environment, we incorporate domain-aware processing to extract this implicit semantics, and thus build up the provenance graphs required for Objective (2). Once this information has been extracted and stored, it can be transferred without loss to other, more generic, preservation environments.

## Implementation

In all the use cases that we examined, a researcher works through an experiment at the same desktop machine – indeed, as all processing is done locally, this was only to be expected. This simplified our implementation by allowing us to focus on capturing the researcher’s process, i.e. we could restrict ourselves to looking at information flow in one direction only – from the desktop to the repository – during the processing of the data. The approach we took was to use a lightweight client, running on the researcher’s computer, to “scavenge” information from the researcher’s work area and transfer it to the repository environment. A further simplification was made possible by the fact that a researcher works on an experiment in a dedicated directory, so all files derived during processing are saved in the file hierarchy within that directory, which means that we only needed to “watch” this directory.

Specifically, each time that a file is created, modified or deleted within the watched directory, the file is uploaded and a message is sent to the repository containing the nature of the action, the original pathname and the timestamp. On the repository side, this information is interpreted and used as the basis for creating digital objects, extracting domain-specific metadata from the objects, and inferring relationships between objects, which are then ingested into the repository, which was implemented using Fedora Commons<sup>6</sup>, all relationships being stored in the RELS-EXT datastream of the Fedora digital objects. Much of the processing that takes place is concerned with analysing the information that is available and exploiting it to infer the details of the researcher’s workflow. Although this workflow is outside our control, in any particular category of use case (in our case, either macromolecular crystallography or nanoimaging) its structure is broadly known at a high level. This means that certain files are expected at certain stages, and in addition the workflow generates as a by-product a lot of information that can be used to infer inter-object relationships, for example in file headers and log files<sup>7</sup>.

Our approach can also be made to work in situations where it is not possible to use the “watcher” client on the researcher’s desktop. To implement this, the directory is submitted for deposit in its entirety at the end of the experiment, and the files are transferred to the repository one by one in timestamp order, thus simulating the live creation of the files. This worked because most files are just created once and not updated, although modifications to the software were required to deal with those few files that were updated, mainly log files to which the tools appended status information each time they were executed.

## Provenance

Data provenance is a particular kind of metadata that describes the derivation history of digital objects. It is widely applied in the digital library community as a way of documenting the activities that occur during the lifecycle of a digital object (PREMIS, 2011), and in the e-science community as a way of recording the scientific process with a view to verifying or reproducing it (Simmhan et al., 2005).

---

<sup>6</sup> <http://www.fedora-commons.org/>

<sup>7</sup> Of course, the researcher could in principle “sabotage” this approach by, e.g., moving files to other directories or by renaming files at random.

The Open Provenance Model (OPM) is an emerging standard for modelling provenance that aims to enable the digital representation of the provenance of any object, whether itself digital or not, in a generic manner so as to support the exchange of provenance information between different systems, the building of common tools etc. (see Moreau et al., 2011 or Moreau et al., 2008 for more details). OPM represents provenance of data as a directed graph, where the node types correspond to the fundamental OPM entities artifact, process and agent, and the arc types indicate the nature of the causal dependency relationships between the nodes. The basic concepts of OPM are illustrated in Figure 2, which includes all the node types but an incomplete set of relationships. The three types of entity are: (i) artifact, an “immutable piece of state, which may have a physical embodiment in a physical object, or a digital representation in a computer system”, (ii) process, an “action or series of actions performed on or caused by artifacts, and resulting in new artifacts” and (iii) agent, a “contextual entity acting as a catalyst of a process, enabling, facilitating, controlling, affecting its execution”. Thus in Figure 2, Agent Ag controls Process P which takes as input Artifact A1 and generates Artifact A2.

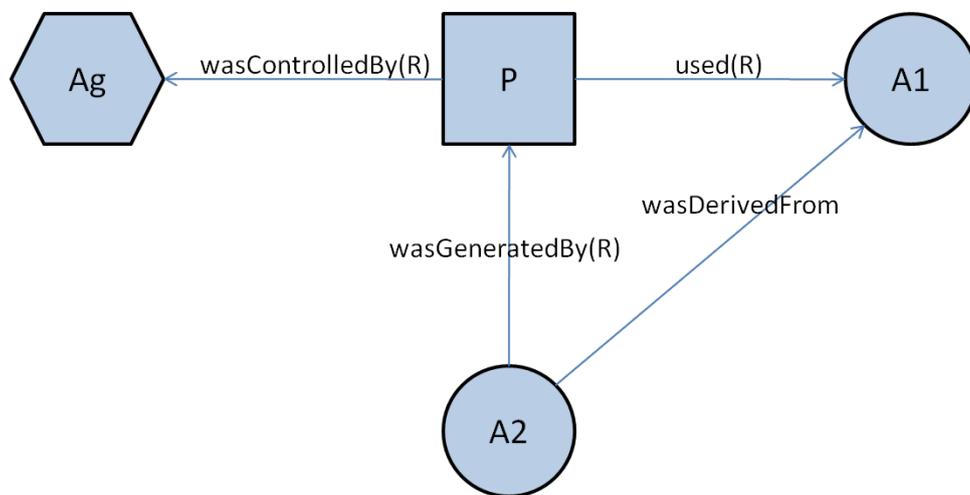


Figure 1: Fundamental concepts of OPM

An XML serialisation of this abstract, graph-theoretical model has been developed (Moreau & Groth, 2010). However, for this project we used OPMV (Open Provenance Model Vocabulary), a more lightweight profile of OPM that is implemented as an OWL-DL ontology (Zhao, 2010). We considered OPMV to have several advantages for our purposes: it was simpler to use in practice; it was more understandable both to users and developers as it closely matched the graph-based generic model; it was easier to combine with other RDF- or OWL-based vocabularies such as Dublin Core<sup>8</sup> or FOAF<sup>9</sup>; it provides better compatibility with semantic web technologies for the publication of experiments as linked data..

In our case, artifacts corresponded to the files generated during the processing (and in some cases compound objects comprising sets of related files), processes corresponded to the processing steps in the researcher’s workflow<sup>10</sup>, and agents corresponded to users who initiate processes and the software tools that are used in the processing. These entities were modelled as repository objects in our Fedora repository implementation, in the case of processes and agents as objects containing

<sup>8</sup> <http://purl.org/dc/terms/>

<sup>9</sup> <http://xmlns.com/foaf/0.1/>

<sup>10</sup> The OPM model allows processes to be represented at multiple levels of granularity. As most of the cases we addressed were very interactive, each logical step in the processing corresponded to an action on the part of the researcher (e.g. to request to execute a command or a script). This need not be the case in general however, as a single user action could correspond to the execution of multiple processes, and contrariwise multiple processes could be combined into a higher-level process.



standards for the description and exchange of aggregations of Web resources”<sup>11</sup> (Pepe et al., 2009). In brief, this allows *aggregations* of Web resources to be defined and assigned a persistent URI<sup>12</sup>; the Web resources that make up an aggregation are known as *aggregated resources*, and when an aggregation’s URI is dereferenced it resolves to a *resource map*, which is a description of the aggregation (for example, in the form of an RDF graph). One of the experimental workflows that we have been addressing may naturally be regarded as an aggregation in OAI-ORE terminology, and in our case the aggregated digital objects include not only the files generated – the “artifacts” in OPM terminology – but also the objects corresponding to OPM processes and agents, and the repository object that describes the experiment itself.

Note that, in addition, the aggregation may aggregate additional objects that are outside the repository, for example publications or presentations based on the research, and (for the crystallography users) entries in the Protein Data Bank<sup>13</sup>. Moreover, the resource map that describes the aggregation may include additional triples that make assertions about the aggregation, the aggregated resources, the resource map itself, or about any Resources that are related to these, with the restriction that the resource map must be a connected graph. As well as including information about semantic types and other metadata, the resource map could include links to other objects of interest that are outside the control of the researcher, for example additional research related to the issues addressed by the experiment, other experiments that reuse the data, or attempts to repeat the original work<sup>14</sup>.

The representation of experiments using OAI-ORE resource maps can be used for publishing the experimental data to the Web in a linked data-compatible fashion – the resource map is precisely a “map” of the aggregation in RDF graph-based form. Resource maps corresponding to experiments can be published to the Web and thus crawled by appropriate Web agents to aid subsequent discovery, analysis and re-use. In particular, by using RDFa one can embed a resource map within a human-readable webpage that describes the aggregation, corresponding to the “splash page” for the experiment within the repository<sup>15</sup>. Note that multiple experiments may use the same raw data, so the graphs may overlap and form a wider network of information.

In addition, the resource maps can be used as the basis for interactive user interfaces for browsing the experiments. An interface that just allowed a user to browse through lists of the constituent objects of an experiment, or search for files based on metadata fields, would not be very useful. While a user may wish to find experiments themselves in this way, they are likely to want to be able to drill down into experiments in a way that corresponds to their conception of how the experiment took place, in terms of processing steps and the files generated at each step. We have used the RDF within in resource maps to provide a natural basis for driving such a graphical user interface, illustrated in Figure 4.

---

<sup>11</sup> <http://www.openarchives.org/ore/>

<sup>12</sup> For example, a DOI, as described in Brase (2009). See also Burton & Treloar (2009) for a discussion of identifiers in the context of the Australian National Data Service.

<sup>13</sup> <http://www.rcsb.org/pdb/home/home.do>

<sup>14</sup> See <http://www.openarchives.org/ore/1.0/datamodel#GlobalRels>

<sup>15</sup> See [http://www.openarchives.org/ore/1.0/rdfa#URI\\_Choice](http://www.openarchives.org/ore/1.0/rdfa#URI_Choice). It is also possible to include the splash page itself as one of the aggregated resources belonging to the aggregation

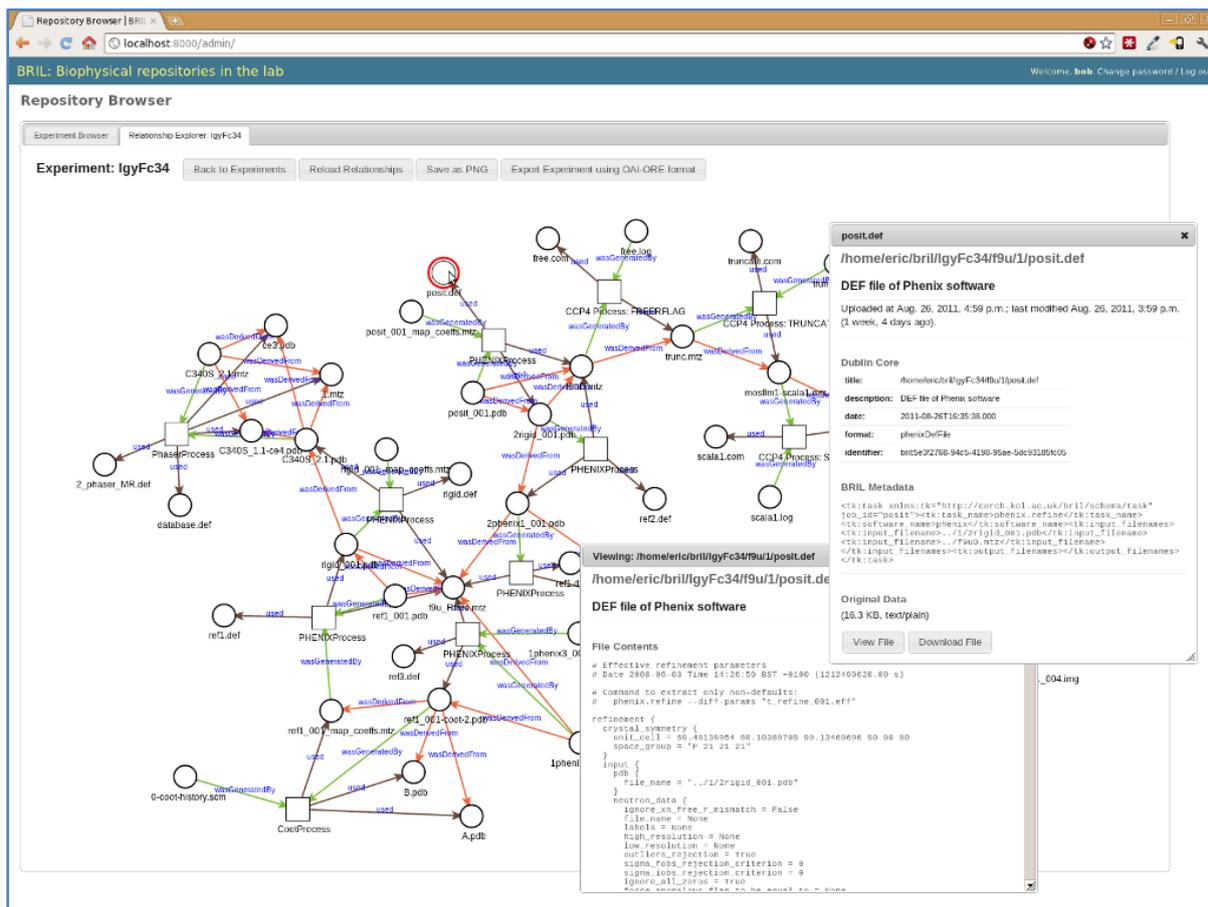


Figure 3: Experiment Browse Interface

Another way in which the OAI-ORE representation can be used is as an inter-repository exchange format. The full dataset corresponding to an experiment is a complex graph consisting of many nodes. If such an object is to be transferred between repositories – for example, in our case, from the sheer curation environment to a generic institutional preservation environment – then we require a common exchange format that describes adequately the structure of the object (as Dissemination Information Package) and provides enough information for the target repository to interpret the package correctly (as Submission Information Package). Several mechanisms have been proposed for this, including the Repository eXchange Package (RXP) proposed by Caplan et al. (2010). We have been investigating the use of resource maps for this purpose, where the destination repository accesses the content of the aggregated objects via their URIs.

## Conclusions

The environment in which the experimental processing takes place in our use cases is very unpredictable and “untidy” compared to the more controlled processing of that occurs in e-science workflow environments. However, it cannot in general be assumed that scientific data processing occurs in integrated environments that are subject to close monitoring and control. Indeed, this sort of model – where much of the processing is outside one’s control and only loosely coupled to the curation environment – occurs in other disciplines, so such automated approaches to capturing the progression from raw data to published results have the potential for broader application.

Mechanisms for publishing the data and provenance arising from an experiment in their entirety are likely to become increasingly important as community and political pressures drive a movement towards increased openness in science and the verifiability and reproducibility of published scientific results. Moreover, by publishing in a form compatible with linked data standards the results can be more easily reused by various clients and for various purposes (see Zhao et al (2008)).

However, the approach does require a quite detailed understanding of the researchers' work practices, which has to be elicited from them via user engagement activities that can be very time-consuming, particularly when the development team has no background knowledge of the discipline in question. The resources required for this were somewhat underestimated at the start of the project. The extent to which this effort is justified would have to be addressed on a case-by-case basis.

## Acknowledgements

We gratefully acknowledge funding for the BRIL project by the Information Environment programme of the Joint Information Systems Committee (JISC), UK.

## References

- Borgman, C. L. (2007), *Scholarship in the digital age: Information, infrastructure, and the Internet* (Cambridge, MA: MIT Press).
- Brase, J. (2009), "Datacite - a global registration agency for research data", COINFO '09: Fourth International Conference on Cooperation and Promotion of Information Resources in Science and Technology, 257-261, <http://dx.doi.org/10.1109/COINFO.2009.66>.
- Burton, A., Treloar, A. (2009), "Designing for Discovery and Re-Use: the 'ANDS Data Sharing Verbs' Approach to Service Decomposition", *International Journal of Digital Curation*, Vol. 4, No. 3, 44-56, <http://ijdc.net/index.php/ijdc/article/viewFile/133/172>.
- Caplan, P., Kehoe, W., Pawletko, J. (2010), "Towards Interoperable Preservation Repositories: TIPR", *International Journal of Digital Curation*, Vol. 5, No. 1, <http://www.ijdc.net/index.php/ijdc/article/view/145>.
- Curry, E., Freitas, A., O'Riain, S. (2010), "The Role of Community-Driven Data Curation for Enterprises". In *Linking Enterprise Data*, Part 1, edited by D. Wood (Boston, MA: Springer US), pp. 25-47. [http://dx.doi.org/10.1007/978-1-4419-7665-9\\_2](http://dx.doi.org/10.1007/978-1-4419-7665-9_2).
- Key Perspectives Ltd (2010), "Data dimensions: disciplinary differences in research data sharing, reuse and long term viability. SCARP Synthesis Study", Digital Curation Centre, <http://hdl.handle.net/1842/3364>.
- Lyon, E., Rusbridge, C., Neilson, C., Whyte, A. (2010), *Disciplinary Approaches to Sharing, Curation, Reuse and Preservation*, DCC SCARP Final Report, <http://www.dcc.ac.uk/sites/default/files/documents/scarp/SCARP-FinalReport-Final-SENT.pdf>.
- Moreau, L., Freire, J., Futrelle, J., McGrath, R. E., Myers, J., Paulson, P. (2008), "The open provenance model: An overview". In IPAW, J. Freire, D. Koop, and L. Moreau, Eds. *Lecture Notes in Computer Science*, vol. 5272. Springer, 323–326.
- Moreau, L., Groth, P. (2010), Open Provenance Model (OPM) XML Schema Specification. Latest version <http://openprovenance.org/model/opmx-20101012>.
- Moreau, L., Cliord, B., Freire, J., Futrelle, J., Gil, Y., Groth, P., Kwasnikowska, N., Miles, S., Missier, P., Myers, J., Plale, B., Simmhan, Y., Stephan, E., and Van den Bussche, J. (2011), "The Open Provenance Model core specification (v1.1)". *Future Generation Computer Systems*, Vol. 27, No. 6, 743-756. <http://dx.doi.org/10.1016/j.future.2010.07.005>.
- Pepe, A., Mayernik, M., Borgman, C.L., Van de Sompel, H. (2009), "From Artifacts to Aggregations: Modeling Scientific Life Cycles on the Semantic Web", *Journal of the American Society for Information Science and Technology*, Vol. 61, No. 3, 567-582, <http://arxiv.org/pdf/0906.2549>.
- PREMIS (2011), PREMIS Data Dictionary for Preservation Metadata version 2.1, <http://www.loc.gov/standards/premis/v2/premis-2-1.pdf>.

Rumsey, A. S. (ed.) (2010), *Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information*, Final Report of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access, [http://brtf.sdsc.edu/biblio/BRTF\\_Final\\_Report.pdf](http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf).

Shearer, K. (2009), *Survey of Digital Preservation Practices in Canada*, Library and Archives Canada, <http://www.collectionscanada.gc.ca/digital-initiatives/012018-3100-e.html>.

Simmhan, Y., Plale, B., Gannon, D. (2005), "A survey of data provenance in e-science", SIGMOD Record, Vol. 34, No. 3, 31-36, <http://portal.acm.org/citation.cfm?id=1084812>.

Whyte, A., Job, D., Giles, S., Lawrie, S. (2008), "Meeting Curation Challenges in a Neuroimaging Group". The International Journal of Digital Curation Issue 1, Vol. 3. <http://www.ijdc.net/index.php/ijdc/article/view/74/53>

Zhao, J., Goble, C., Stevens, R., Turi, D. (2008), "Mining Taverna's Semantic Web of Provenance", Concurrency and Computation: Practice and Experience, Vol. 20, No. 5, 463–472, <http://onlinelibrary.wiley.com/doi/10.1002/cpe.1231/full>

Zhao, J. (2010), Open Provenance Model Vocabulary Specification. Latest version: <http://purl.org/net/opmv/ns-20101006>.