International Journal of Digital Curation

Editorial

Alexander Ball
Digital Curation Centre

It is with great pleasure I come to introduce the second issue of Volume 9 of the IJDC, though introduce is perhaps the wrong word. At the time of writing, many of the papers within have been published for some time as a consequence of the rolling publication model to which we moved with this issue. Perhaps you have already read some of the works I am about to describe? If so, I hope you found them informative, and that I can persuade you to sample some more of the issue's contents. These include eight further articles based on practice papers presented at this year's International Digital Curation Conference, a ninth derived from a paper presented at IASSIST 2014, plus three peer-reviewed papers received through general submission. The remaining article is Donnelly's review of the timely collection *Research Data Management: Practical Strategies for Information Professionals*, edited by Joyce Ray.

As Donnelly remarks, Research Data Management is an activity that involves multiple stakeholders, ideally working in concert. Dillo and Doorn present a prime example of this, namely the federated data infrastructure emerging in the Netherlands. They describe a four-tier model, in which end-user researchers work with a 'front office' provided by their institution or centre and focusing on active data management. Archival functions and expertise are provided by 'back office' shared services such as DANS, which themselves rely on infrastructure providers such as SURFsara for core services such as storage.

The model has both its appealing features and its issues, so it will be an interesting one to watch. As the need for data management support permeates through to more (and smaller) institutions, we can expect the shared services approach to become more attractive in more contexts. But it seems we are not quite there yet. Indeed, of the eight US universities studied by Akers et al., only one is experimenting with shared storage infrastructure, specifically a Dataverse Network. A more common scenario arising from these case studies is that of existing library services expanding to accommodate end-to-end data management support functions, most likely in response to the National Science Foundation's data management planning requirements.

The picture in the UK is similar. In particular, the set of expectations published by the Engineering and Physical Sciences Research Council (EPRSC, 2014) has caused many universities to review urgently how they support data management. My colleagues Jones et al. describe how the Digital Curation Centre has assisted UK higher education institutions to set up their own research data management policies and services; moreover, they explain how the first phase of this work was evaluated, and how the results have

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.v9i2.341
DOI: 10.2218/ijdc.v9i2.341

been used to reshape the current, second phase.

Such developments are remarkable since, in the UK at least, data management has traditionally been seen as a subject-based activity rather than an institution-based one. The shift towards a hybrid landscape – with some data managed by specialist centres and some by generalist facilities in universities – raises all sorts of issues, including how staff with relevant digital curation skills are distributed between them. Sands et al. provide a fresh insight on this, as they analyse the expertise and composition of the team at the University of California, Los Angeles that tackled the archiving of the data from the Sloan Digital Sky Survey. They highlight how economies of scale play a crucial role: certain data management knowledge can only be gained through experience, and the curatorial process can only ever be incomplete in the absence of domain expertise.

Pursuing this point, Pejša, Dyke and Hacker reflect on the how the Network for Earthquake Engineering Simulation (NEES) Consortium has established and influenced data curation practices over the past decade. The article describes how domain specialists from multiple laboratories collaborated to build a data commons for their field: they began by establishing data and metadata standards, and progressed to setting up a shared virtual research environment supported by a curation and preservation team. The usage statistics indicate increasing uptake by the community, and since 2011 the National Science Foundation has explicitly funded research that reuses the data curated there.

By way of contrast, Bachell and Barr examine an industry where a lack of preservation activity is putting digital cultural heritage at risk. Video game preservation is an area which is challenging not only on a technical level but also because of the intellectual property rights issues and commercial sensitivities. The issues are compounded for independent games developers, who do not have the turnover and therefore budget to invest in dedicated data management systems and expertise. It is good to see that, nevertheless, Bachell and Barr uncovered an awareness among them of the business case for preservation, and a willingness to consider lightweight solutions. This underlines how important our work in the digital curation community, researching and developing such solutions, is to small enterprises and lone researchers as well as professional digital curators.

The solutions presented in the remainder of this issue, though, will probably be of greatest interest to those responsible for data management services on a larger scale. Beginning at the data creation stage, we have two articles concerned with executable workflows. Song et al. describe a prototype system that takes a set of user requirements and, drawing from a catalogue of processing tools, suggests viable workflows that satisfy them. It can also suggest optimisations for a given workflow, such as eliminating steps with nil effect or converting a serial processing step into a parallel one. Meanwhile, Cuevas-Vicenttín et al. present a graph-based database for storing provenance information about scientific workflows. The system can be used to answer questions like how many times a process was invoked, which processes used a particular data object, or which data objects were raw inputs to a particular workflow. The current prototype only accepts trace files exported from VisTrails, but support for other workflow systems is planned along with integration with the DataONE infrastructure.

Moving forward to the point of ingest, it has long been recognised by archives that keeping everything that donors deposit is not only impractical but counter-productive. Niu offers some assistance to those tasked with establishing selection policies and procedures for digital repositories. She sets out a framework consisting of a set of selection methods from which to choose and a sequence of implementation steps to follow. While the framework appears simple, it is based soundly on archival literature, and the paper explores the nuances of its various elements.

Appraisal criteria can include many factors including the uniqueness of a file, the trustworthiness of its provenance, and the sensitivity of its contents. Meister and Chassanof demonstrate how to collect these sorts of information using the digital forensics tools supplied by the BitCurator system, as well as perform other ingest-related tasks like metadata extraction.

The time it takes to perform such tasks scales in line with the number and size of files to be processed, and there are times when even careful selection cannot keep timescales down to acceptable levels. Arora, Esteva and Trelogan describe how a team at the University of Texas was faced with precisely this issue, and sought to speed up the task of extracting metadata from a complex, evolving, multi-terabyte digital collection. They turned to parallel processing on a High Performance Computing platform, and while there were several hurdles to overcome, they nevertheless managed to achieve the desired result.

One of the final tasks of the ingest process is of course to prepare files for archival storage, but which formats should be used? For certain use cases, PDF – or more specifically PDF/A – has its attractions, but how suitable an archival format is it? Evans and Moore reflect on this question, drawing on their experiences working with the PDF/A-1 format at the UK's Archaeology Data Service. One of their arguments is that the value of PDF/A lies not so much in the format itself but in the way archivists and software developers collaborated to develop it, and that perhaps this conversation needs to happen whenever new software and formats are developed. The study by Bachell and Barr and the article by Pejša, Dyke and Hacker each confirm in their own way the positive effect this would have; it is really a corollary of that long-standing contention that in order to do digital curation properly, you have to plan for it right at the start.

References

EPRSC. (2014, October 9). Clarifications of EPSRC expectations on research data management. Retrieved from http://www.epsrc.ac.uk/files/aboutus/standards/clarificationsofexpectationsresearchdatamanagement/