The International Journal of Digital Curation
Issue 1, Volume 3 | 2008

Editorial

Chris Rusbridge,

Digital Curation Centre

July 2008

The 3rd International Conference on Digital Curation was held late last year in Washington DC. I was not able to attend, but I gather it was a very successful event. Most of the papers in this issue of the International Journal of Digital Curation stem from that conference, or its predecessors. The nine papers in this issue are all very different, although there is a sinuous thread linking them; some do share a sub-theme, and come to conclusions that are the same in some ways and very different in others.

Taking a broad view across several related projects with chemistry, e-Science and data curation perspectives, Frey shows how today's technology radically changes the basic record-keeping element of laboratory practice, the laboratory notebook. In the process, some aspects of record keeping necessary for proper curation become more difficult: writing the huge data volumes that can result from modern laboratory instrumentation into a lab notebook is clearly impractical. However, with care, better curated data can result. Curating the data at source, he argues is essential; the context cannot be re-constructed at a later date. Curation is the job of the scientist, and not to be left to the archivist or librarian at a later stage.

Wallis et al write from the archivists' point of view; in this case the field is environmental sensing. Their related concerns that data and the technological infrastructure be better documented at the earlier stages of experiments are coloured by their perspective, and suggest earlier involvement by the archivist. In practice, the message is the same.

The main context that Wallis et al were investigating related to short-term deployments of instrumentation arrays. They do note that other projects, involving collection of data with the intent of creating longer time series, had necessarily developed better data curation practices. Dürr et al describe work on collection of hydrology data that does indeed take account of curation needs.

Although Dürr et al involved the scientists in their collection strategy, the datasets seem to be collected as end products rather than an integral part of the experimental process. Among their conclusions was the requirement for very strict quality control on ingest. This is particularly understandable where data are treated in a "fire-and-forget" way by the researchers, rather than handled as inputs to subsequent experimental phases, where the data quality is of more pressing concern to researchers.

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. ISSN: 1746-8256 The IJDC is published by UKOLN at the University of Bath and is a publication of the Digital Curation Centre.

One of the roles envisaged for data curation archives is to act in some "community proxy" roles. Vardigan et al describe just such an activity in the Social Science community, where the data archive community is well established (with over a 40-year history), and has developed standards to document social science datasets. With the development of DDI 3.0, they are providing additional features for metadata re-use and for use throughout the data life cycle, rather than just the end-product dataset.

Metadata are also the focus for Patel and Ball, who examine how Representation Information may be used as a basis for curation and preservation strategies. They look in particular at applying these strategies in the fields of crystallography and engineering, and, among the challenges and issues associated with this, they note the need for robust IT infrastructure and global collaboration.

Taking a broader, perhaps more "helicopter view" and without specific disciplinary focus, Day reviews collaborations in the sciences and in digital preservation. He then looks at concepts of trust and control from management sciences, before laying out the series of attempts to standardise trust in the context of "trusted digital repositories" (TDRs). The wonderful notion that trust relates to a willingness to be vulnerable to the actions of others, regardless of any control mechanisms, in exchange for some perceived reward or advantage, makes clearer the distinction between genuine trust and any formal stamps that checklist or other certification activities could provide.

A network of collaborations on a grand scale is the subject of Anderson's paper on the US NDIIPP Network. In particular this paper reveals the considerable degree to which simple differences in organisational background, expectations and practice can influence collaborative effort. A critical observation is that "interoperability for long-term preservation is data-centric and not system-centric".

Looking at the data containers rather than the actual data, Pearson and Webb discuss the vexed issue of file format obsolescence. However, they are less concerned with when and how this might occur than with how and when a preservation repository might discover and decide to act on that obsolescence. In particular, they report on an Australian development, the Automatic Obsolescence Notification System, or AONS II. This work does to a certain extent encompass the technology watch element of preservation services, and they suggest that it needs development to become more of a shared activity. Individual repository managers could then look at, and perhaps contribute to, the risk factors for particular formats, while taking their own decisions on actions, which will vary greatly depending on environment and requirements. Obsolescence is a slow process, and the timing of actions may not be critical, although too late may not be realised until… too late!

Finally Moore presents work towards a theory of digital preservation, viewed as validation of communication from the past. Preservation is then the minimum set of preservation processes to achieve defined policies, and the minimum set of metadata to validate the assessment. Moore places this in the context of his rule-oriented data grid developments. There are some wonderfully terse observations here, for example a "preservation environment is the software middleware that shields records from the rapid evolution of technology". A key feature of this work is its emphasis on automation and rule-based decisions. Much early work on digital preservation, and many of its casual assumptions, were based on significant human involvement. The social science repositories mentioned in Vardigan et al, for instance, tended to validate individual deposits. But as we move into the territory both of large-scale national library legal deposit environments on the one hand, and of the science data deluge on the other, scalability factors dictate that automation must take over. The idea that our scientific and cultural heritage might depend in the future on completely automated preservation actions is deeply troubling, and I am not clear that we have even started to have this particular conversation. However, I have great faith in our technologists, and we can be sure that nothing will go wro