The International Journal of Digital Curation
Issue 1, Volume 6 | 2011

Editorial

Kevin Ashley,
Digital Curation Centre

March 2011

It has been almost nine months since the last edition of this journal, but on seeing the amount of content we have for you in this issue you may appreciate why it's taken longer than we planned. There are 21 papers overall, 12 peer-reviewed papers from IDCC10 in Chicago and 4 from iPres 2009 in San Francisco. There's another 5 general articles from iPres 2009 as well. Before looking at these more closely, it's interesting to reflect on the change in the nature of much of the work reported here. In developing the award-winning digital preservation management tutorial at Cornell 10 years ago, Anne Kenney and Nancy McGovern defined a five-stage maturity model for digital preservation in institutions. Beginning with isolated, individual projects, it moved through stages which included embedding in institutional processes (at which point digital preservation as a separate action can become invisible, and certainly unremarkable) and finally includes cross-institution cooperation and embedding (Kenney and McGovern, 2003). The articles submitted to *IJDC* reflect this maturation over the past six years. Although we still see useful and interesting papers about isolated projects, much of the work now concerns scale, embedding and services which involve multiple institutions. Digital curation as a field is undergoing rapid maturation.

The first group of papers deal with issues that relate in some way to the problem of scale: dealing with quantities of material far too large to expect human effort to be expended on each one. Many discuss forms of automation, but others deal with techniques to make large quantities of information comprehensible, and in particular to help humans make decisions on where human intervention is required. Esteva et al. discuss the use of visualisation to examine the preservation condition of large record collections, where a single image can tell us a great deal about such things as the relative risk posed by formats used in collections of millions of documents. Visualisation of this sort can clearly be used to tell us other things about document collections, many of which will be interesting to users of the collection as well as to its custodians. Hsu & Brown describe a technique to determine the software dependencies in software embedded in CDROMs, the target being to be able to construct appropriate emulation environments automatically and on demand. The problem isn't as simple as it might appear, even in the subset that they examine (MS Windows 3.0 and following.) Their reference to "DLL Hell" is shown to be all too real, and they demonstrate that there's already a pressing need to translate folk knowledge about software requirements into a more systematised form, capable of machine processing. Gelernter & Lesk describe automation applied to the task of matching data from

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.

heterogeneous data sets – attempting to use ontologies and reasoning both to match variables which are essentially the same (even when coded differently), as well as the more subtle matching required to relate a field called "Age" with one called "Date of Birth." Suchodeletz et al. look at automation within emulated environments. Tools run inside emulators are proving to be of increasing use as an aid to migration, amongst other tasks, but they are often designed for human interaction via keyboard or mouse. To use them on an industrial scale requires a means of automating such interactions, controlled by the world outside the emulator. Some of their work, and the problems they encounter, are reminiscent of the challenges faced in developing and using the XTest and XTrap extensions of the X Window system (Annichiarico et al., 1991; Drake, 1991). Meanwhile, Tarrant, Hitchcock & Carr describe marrying the notions of Linked Data with information about format risk and migration to produce a service which gives us more than the data that went into it. With increasing scale we hope we gain increasing efficiency, but costs are still likely to rise. The paper from Kejser, Nielsen and Thirifays describes work undertaken in Denmark to produce a cost model for dealing with digital materials at cultural heritage institutions. Two aspects of this paper struck me. They have attempted to deal systematically with the problem of assessing format complexity, widely acknowledged to be related to cost in some way. Using the length of the format's documentation (as well as other factors) as a proxy is innovative and intuitively correct. Second, this paper contains what I believe to be *IJDC*'s first reference to material in a web archive – ironically from the CEDARS project, itself concerned with long-term preservation.

Data reuse can be examined from a number of perspectives, and the next group of papers does just that. Donnelly and North report on their work with the MESSAGE project, in which Donnelly was able to effectively embed himself as an observer in the project team. Useful insights into attitudes both to primary data use and data reuse from a mixed academic and commercial research team are offered and the work is set against the context of similar investigations. The paper benefits from articulating the views of the researchers directly as well as that of Donnelly as observer. Amongst their observations is that data changes in importance and in its meaning, even when the data itself does not change. The study also shows how data collection can change not just our opinion about the answers to research questions, but about the importance of the questions themselves. Finally, Faniel and Zimmerman consider the existing literature on data reuse and attitudes to it and propose a list of research questions to support further progress. They feel that the problems we face include not simply the quantities involved in the data deluge, but other changes in the number of actors involved and the increasing range of intermediaries. Data centres have been the key actors in enabling data reuse for over 40 years in some disciplines; Collins reports on a study carried out by Technopolis on behalf of the UK's Research Information Network on researchers' attitudes to such data centres in the UK. Most believe that the existence of such centres does not stimulate novel research questions, but does improve the efficiency of research in general. At what some might call the extreme end of data reuse is open science – not simply exposing your data after publication, but exposing your work and methods from the moment you start developing research questions. Whyte & Pryor report on an exploratory study on researcher motivations for open science and open data. It includes a comprehensive review of the literature, examining benefits not just to researchers but also to their institutions and funding bodies. Collating information on these benefits is work that I hope the DCC will be undertaking in the coming year. Wynholds takes a closer look at one problem that's related to those benefits, that of

data citation. Unlike other cited objects, the edges of data are fuzzy and the nature of what needs to be identified is unclear. Wynholds received the award for best student paper at IDCC10 in Chicago, December 2010, for this work.

Another thematic thread identified in this edition is that of institutional and professional change. Kim, Addom and Stanton look at education requirements for e-science professionals and the corresponding job skills. Their work is intended to help LIS schools adjust their curricula to meet the requirements of the market. I read their conclusions as being that the job is about curation, communication and infrastructure. This work won the award for best paper at IDCC10. In a general article based on a presentation at iPres 2009, Bermès and Fauduet look at the human challenges presented by the move to digital at the Bibliothèque nationale de France. Over 10 years, their organisation has also gone through many of the stages identified in Kenney & McGovern's maturity model. The change appears almost complete – as they observe, when the whole library is digital, you don't need a digital library department any more. Not every institution is as advanced as BnF as the paper from Sinclair et al. demonstrates. They report on a survey of over 200 institutions conducted in early 2009 by the PLANETS project. It examined awareness, preparedness and planning for digital preservation at libraries and archives across Europe. High levels of awareness exist, but only about half have translated this to policies and budgets. At the other end of the organisational scale, Prom comments on the problems faced by (often) lone archivists in making digital curation a "systemic institutional function." He describes the straightforward methods he has developed to help them. This work has already received much positive comment from archivists and records managers, but not all of Prom's observations chime with my own experience of training archivists over the past 15 years. With this, as with all our articles, we would be interested in your observations. Innocenti et al. describe work from DL.org on policies for interoperability between digital libraries (a term which is used here to encompass archives, repositories and other digital collections). Or is it about interoperable policies? As one of a number of co-authors on this paper, I have my opinion, but I'll let you decide.

SHAMAN is an EU-funded FP7 project looking at many aspects of digital curation in industry. Wilkes et al. describe one set of findings relating to preservation of engineering content, with particular reference to product life cycle management. They present a solution architecture – and another lifecycle model. Sperberg-McQueen describe an approach to test the validity of format conversions in which XML and its assertions are used to formally prove or disprove successful migration. Their assertions are bold and one wonders how widely they can be applied given the document-centric character of the technique. The paper critically examines the problems and limitations of the approach it describes. Dappert and Farquhar, in another paper derived from an iPres 2009 presentation, reflect on the interactions between digital preservation services and digital preservation metadata. This leads to a view that preservation metadata should not be thought of as static observations, but as information to support processes and services; the question is not what we want our metadata to say today, but what do we want it to do in the future? They develop a practical data dictionary to support this approach.

I end my descriptions with two unrelated but equally remarkable papers. Guttenbrunner et al. present a case study on techniques used to recover data from early

home computer systems which were recorded on audio cassettes. As well as giving detailed information on the discoveries made about one particular format (Philips Videopac+ G7400) they have useful general observations on this class of problem. Their tools are already able to read original media that the original systems cannot; I expect that better signal-processing techniques could help them achieve even more.

I thought I was well-informed about fonts, how systems and software process them, and the problems they can present in preservation. That is, until I read the paper from Brown and Woods in this issue. Based on analysis of two large document collections, they show that the problem of identifying fonts and appropriate substitutions is even worse than I thought. In this clear and detailed exposition of their work, they show that 8% of documents are likely to suffer significant preservation problems as a result, and that government document collections (where one might expect more conservative font usage) are no better than others.

So, a rich selection of material in this issue. There's more to come in Volume 6, Issue 2 – perhaps including your comments on the material here. Journals are not monographs and no one expects a journal to express a single point of view. Although I and my colleagues on the editorial board apply selection criteria to the material before you, we don't necessarily agree with the conclusions of all the authors – in some cases not even with the premises from which they begin. You may find yourself in a similar position after reading one or more of the articles here. If so, we encourage you to write to us at ijdc@ukoln.ac.uk. Only through debate can we refine and test the ideas we are collectively exploring in digital curation.

References

Kenney, A. & McGovern, N. (2003). The five organizational stages of digital preservation. *Digital Libraries: A Vision for the Twenty-first Century. A Festschrift to Honor Wendy Lougee on the Occasion of her Departure from the University of Michigan*. Retrieved March 9, 2011, from http://quod.lib.umich.edu/cgi/t/text/text-idx?c=spobooks;idno=bbv9812.0001.001;rgn=div1;view=text;cc=spobooks;node=bbv9812.0001.001%3A11.

Annicchiarico, D., et al. (1991). *XTrap: The XTrap architecture*. Maynard, MA: Digital Equipment Corporation.

Drake, K.J. (1991). *Some proposals for a minimum X11 testing extension.* UniSoft Ltd.