International Journal of Digital Curation

Editorial

Alexander Ball
Digital Curation Centre

It is with great pleasure that I introduce this first issue of IJDC Volume 9. The year 2014 promises to be an exciting one for the journal, and it starts with our largest issue to date. Contained within are 23 papers that originate from the Ninth International Digital Curation Conference, held in San Francisco in February 2014, plus a further three papers and articles received through general submission.

Of the conference papers, seven are peer-reviewed research papers while the remaining 16 report on areas of practice. As in previous years, time did not permit the authors of the latter to present their work in detail, so even if you saw the original presentations I would strongly encourage you to read the full story from the pages of this journal.

Let us begin, however, with the paper that won the conference award for best research paper. In repository circles, one of the distinctions often made is between *publishing* something and merely *making it public*. The idea is that publishing is a process for fixing a resource in its best possible state. 'Just sharing' makes no guarantees of quality or permanence, but opens the way for the community to make or suggest improvements. Taking inspiration from open source software, Kansa, Kansa and Arbuckle argue that in the data context we have an opportunity to achieve the best of both worlds: well curated datasets that improve in the light of reuse.

The examples given by Kansa, Kansa and Arbuckle of curatorial attention given to zooarchaeological data resonate with the passionate plea of Peer, Green and Stephenson, that all stakeholders should strive for data quality, though the latter authors are rather less keen on the idea of post-publication community curation. It is notable that Peer, Green and Stephenson direct their arguments towards repositories. They highlight how the situation is inverted for data compared to scholarly papers: it is the repositories, data centres in particular, that are seen to publish data in the sense explained above, where journals might merely make it public. Far from being a concern, it is an opportunity for symbiosis; indeed, several papers this issue explore how the publishers of research papers can coordinate with the data repositories that are publishing the underlying data.

Castro and Garnett present a plugin being developed for the Open Journal System publishing platform that will allow supplementary data to be deposited in a Dataverse repository, and display a prominent data citation on the corresponding article pages. Callaghan, Tedds, Lawrence et al. discuss such links and some other possibilities to which that plugin might aspire, such as pushing citation notifications to the repository, or displaying metadata, visualisations, or maps of geospatial coverage provided by the repository.

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/

 doi:10.2218/ijdc.v9i1.296

The latter paper is the result of research carried out by the PREPARDE project. Another of its outputs is a set of guidelines aimed at journals editors and publishers, helping them to choose a data repository with which to partner, or at least recommend to their authors. I am pleased that Callaghan, Tedds, Kunze et al. are formally publishing it within the IJDC's pages. The guidelines may also be of interest to researchers seeking an appropriate home for their data, and thus to institutional support staff who might advise them.

Akers and Green suggest that academic libraries go a step further, and become members of Dryad. This is not quite as partisan as it seems. It relates to Dryad's particular funding model: by becoming members, libraries can arrange discounted submission fees for their researchers and influence the governance of the repository. The attractiveness of this is of course proportional to the number of datasets likely to be deposited there by researchers from the institution in question.

One of the issues that Akers and Green touch on is that, due to their very nature, institutional data repositories offer datasets less visibility than disciplinary data centres. Ball, Ashley et al. describe a project that promises to address that, at least within the UK, by developing a national data discovery service. The work described owes much to Research Data Australia, with which it shares a vision of making relevant data easy to find, even if one only uses Google.

Making data easy to find is one step closer to getting it reused, and as Wallis points out, without reuse it can be hard to justify the effort to make data reusable, or even know where to begin. She gives two examples where data producers have identified groups that might want to reuse their data, and asked them in what types of data they are most interested and how they would like them documented. Acting on this information means the data producers can concentrate their efforts where they will have most benefit, data reusers can spend less time making sense of the data, and the dataset itself is more likely to have impact.

Wallis' examples come from the surface and climate modelling communities, in whose data there is widespread interest. Will other disciplines find it so easy to identify potential reusers? There are some others I can think of, such as medicine and social science, where there might be interest not only from other research groups but also groups with less worthy motives. Comerford considers the possibility of participants being identified from anonymised data. He presents a technique for balancing this disclosure risk against the loss of data utility that results from more aggressive anonymisation. This will be of interest to anyone deciding how to make sensitive data public safely, but also underlines the need for graded access to sensitive data, with greater restrictions put on access to less anonymous versions.

Ferreira et al. provide a more general treatment of risk management in the context of data management plans (DMPs). They recommend supplementing DMPs with a dedicated Risk Management Plan based on the principles of ISO 31000, and provide a worked example from the MetaGen-FRAME project. A quite different planning gap is identified by Miksa, Strodl and Rauber. They propose DMPs should be complemented by, or perhaps complementary to, a Process Management Plan. This would describe how the data was collected, processed and analysed, and detail how this process will be (and has been) shared and kept re-enactable in the face of changing formats, platforms and services. With both proposals, the question is whether the additional effort is outweighed by the benefits to both the original researcher and the wider community, and whether that

is generally true or specific to certain disciplines.

It is important to ask such questions, given that DMPs themselves can be far from trivial to write. That is why the new data support programme for engineering researchers at the University of Michigan, as reported by Nicholls et al., focuses on DMP writing, and why tools such as DMPonline and DMPTool are proving so popular. Both tools have recently undergone extensive revision, in response to feedback from users. It is instructive to compare how requirements were gathered in each case: the highly structured method described by Getler et al. in connection with DMPonline version 4, with the more ad hoc approach recounted by Strasser, Abrams and Cruse in connection with DMPTool 2.

Of course, it is easier to write a DMP if there is disciplinary or institutional infrastructure already in place to which to refer. If you are in the process of setting up data management services at your institution, you might find inspiration from those who have already done so. Abrams et al. describe the DataShare service piloted in 2013 and launched this year at the University of California. Minor et al. report on a series of data curation pilots run between 2011 and 2014 at the University of California, San Diego to inform that campus's Research Data Curation Program. Lastly, Norman and Stanton use a sequence of case studies to illustrate how research data management support at the University of Sydney evolved between 2012 and 2014.

Both the latter two papers emphasise the need for training in research data management skills to complement the technical side of the infrastructure. If you are designing a curriculum to deliver such training, you could do worse than review Molloy, Gow and Konstantelos' summary of the curriculum framework developed by the Digital Curator Vocational Education Europe (DigCurV) project. The framework has already been used by University College London, the University of Aberystwyth, and Purdue University Libraries. Then, for a novel approach to course delivery, consider how the pilot course outlined by Shadbolt et al. supplemented theoretical work with immersive placements: participants practised their new skills on real data alongside active researchers.

There are circumstances where formal training is not possible, and one has to fall back on less direct methods of influencing behaviour. The question considered by Darch is a classic example: how can a Citizen Science project, reliant on contributions from range of volunteers, ensure the data they generate is credible? Darch takes Galaxy Zoo as his case study, and finds the way in which volunteers are credited has a remarkably high impact on the quality of their contributions. Meanwhile, based on the experience of the eBird project, Lagoze recommends making unusual contributions harder to submit and subject to greater scrutiny than expected ones. Both papers warn that even Citizen Science has its limits of scale, and suggest how machine-learning approaches can be used to stretch those limits.

From classifying galaxies and birds to identifying web resources, now, as Van de Sompel et al. consider better ways of mapping between persistent identifiers and web addresses. The typical approach is to append an identifier to the URL of a resolver service, thus yielding an address that redirects to a metadata record or landing page. But under that model there is no standard way of deriving the web addresses of an object's online resources (if any) from its identifier, nor of deriving the identifier from the web addresses. I will not repeat the proposed solution here, but I can reveal that OAI-ORE Resource Maps and HTTP headers are involved.

The paper ends by noting the importance of common metadata standards in scholarly communication, a sentiment underlined in triplicate by Ball, Chen et al. in their paper on

 doi:10.2218/ijdc.v9i1.296

the work of the Research Data Alliance's Metadata Standards Directory Working Group. It relates how the Digital Curation Centre's Disciplinary Metadata Catalogue – targeting standards relating to research data – has been expanded through volunteer effort, and how the information collected there might be used in a more advanced platform.

Moving away from the conference papers to general submissions now, Emmelhainz takes up the metadata theme with a set of recommendations for controlled vocabularies suitable for use with qualitative cultural anthropology data. The question arises because the field is highly diverse and has proved rather hard to organise conceptually.

Robinson reports on an immersive project to create a digital archive of the work of fashion designer Zandra Rhodes. The paper covers a lot of ground, from appraisal and selection, through digitisation and cataloguing – yes, controlled vocabularies were used alongside a metadata schema derived from several already in use – to some highly engaging forms of display.

Finally, Browning describes the building of a quite different archive, this time of the broadcast content of two current affairs television channels. The paper provides a fascinating insight into how technologies have evolved over the archive's 25-year history, and how various innovations have been used to add value to the collection.

That is a lot of content to digest, so while you consider where to start there are a few things to say about the journal itself. I mentioned at the beginning that this would be an exciting year for the IJDC, and it began with the long-awaited introduction of a data policy requiring that data behind articles be shared and preserved in a custodial environment. You may also notice some small changes to the format of papers, including the restoration of the dates on which papers were submitted and accepted.

By far the biggest change is yet to come, though. As of next issue, we will move to a rolling publication model. This means we will be publishing papers online as soon as they are ready, rather than saving them up to be published all at once. Please, therefore, do not be alarmed when Volume 9, Issue 2 is first published that it seems rather short. It will lengthen as the year progresses! You will know when the issue is complete by the addition of an editorial, meaning in effect that you will have a whole volume of papers to read before you hear from the editorial team again. I had better let you get on with it.