International Journal of Digital Curation

Editorial

Alexander Ball
Digital Curation Centre

This year, 2015, marks some special moments in the history of the Digital Curation Centre (DCC). In February, in London, we held the 10th International Digital Curation Conference. The theme, appropriately enough, was 'Ten years back, ten years forward': it was a celebration of all that had been achieved over the past decade, but also a chance to reflect on how that progress might continue into the future.

This is also the year in which we publish the 10th volume of the IJDC. I have the great privilege of presenting the contents of the first issue to you in these pages. Included are new versions of 22 of the papers from the conference, thus you may well detect that same decadal theme seeping through into this journal. Indeed, the theme has even spread to one of the two original peer-reviewed papers we have in this issue. But more on that later.

Let us begin with the paper that won the conference award for best peer-reviewed paper. Matthews et al. describe the progress made by the UK Science and Technology Facilities Council over the past ten years towards preserving the 'memory of science'. In other words, not only the bits and bytes of the data, but also the contextual information needed to understand them and their importance, and the links between the various parts of the scholarly record. This has included the policies and strategies needed to run a sustainable service, techniques for effective bit preservation, commitments to identifiers and other important metadata, infrastructure for capturing contextual links, and tools for working with Research Objects: defined aggregations of resources linked by their provenance.

One of the ways provenance can be recorded is in the form of a workflow: a sequence of inputs, processes and outputs chained together to turn raw data into a set of meaningful results. Systems like Kepler and Taverna have evolved to make workflows easy to monitor, and their outputs easy to track, but despite the benefits many researchers prefer to write workflows in familiar scripting languages. Catering for this demographic, McPhillips et al. present YesWorkflow, a toolkit that uses special comments in scripts to provide the functionality of a scientific workflow system.

Workflows typically engage external services by means of Application Programming Interfaces (APIs). While there is a temptation for service providers to take a 'build it and they will come' attitude to APIs, the most successful examples have taken some pains to ensure they correspond to the needs of users. Edmond and Garnett report on research conducted in the context of the Europeana Cloud project into whether researchers want or need API access to cultural heritage datasets. They discuss a range of different factors

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.v10i1.376
DOI: 10.2218/ijdc.v10i1.376

that are relevant: not only the tools available for managing workflows, but also the quality of the content on offer, the critical mass of services and data available through APIs, and the pervasiveness of the relevant skills.

In connection with the latter, Edmond and Garnett consider whether Software Carpentry might provide a model for propagating skills among researchers.[1] Software Carpentry is a volunteer organisation that teaches coding skills to researchers in short, intensive workshops using community-maintained learning materials. The question is addressed in more depth by Teal et al., who describe how a sister initiative, Data Carpentry, is teaching data analysis skills using the same approach. They also explain how pre- and post-workshop questionnaires are used to assess the quality and impact of the training. From the results so far, it seems the approach is both popular and effective, perhaps answering some of the questions posed by Tibbo as she considers how to design a workshop or continuing education programme.

Tibbo is careful to consider the historical perspective, and the role played by graduate-level programmes in iSchools. This resonates with the concerns of Lyon and Brenner, who argue that iSchools are ideally placed to provide a ‘Capability Ramp’: a means of building data skills, workforce capability, and practical experience incrementally in response to current and future trends in digital curation. The Capability Ramp model focuses on interactions between students, data scientists and domain experts. For example, immersive placements of students with domain experts helps those students extend their domain knowledge. Lyon and Brenner provide one example of this: the Immersive Informatics course developed by the University of Bath and UKOLN and taken forward by the University of Pittsburgh. Mayernik et al. provide a second: the Data Curation Education in Research Centers (DCERC) internship programme developed by the Graduate School of Library and Information Science at the University of Illinois at Urbana-Champaign, the School of Information Sciences at the University of Tennessee at Knoxville, and the NCAR Library in the National Center for Atmospheric Research (NCAR). The programme has been refined iteratively over its three years of operation to ensure both students and mentors are better prepared, and to give students greater understanding of the broader professional context.

Another way of increasing workforce capability is to provide training for support staff already in post. Grootveld and Verbakel describe the *Essentials 4 Data Support Course* run by Research Data Netherlands. This is the successor to the 3TU.Datacentrum course *Data Intelligence 4 Librarians*; the change in name recognises that research data support is provided not only by librarians but also IT and domain specialists. The paper describes various ways in which the course has been improved in response to feedback. I note, for example, that the course is now structured around the data lifecycle instead of skill sets, the better to reflect the concerns of the researchers being helped.

Funders increasingly expect researchers to have data management skills already, of course, but it is an open question whether they are acquiring them in the right way at the right time. For example, in her UK-based survey of doctoral students and supervisors, Abbott found that neither group were confident in their data management skills, suggesting that institutions should not expect such skills to be passed on within the supervision relationship. Rather, Abbott argues, more specialist training courses are needed.

Carlson and Stowell Bracke present a possible way forward. They describe a data

[1] Software Carpentry: https://software-carpentry.org/

literacy programme piloted by the College of Agriculture at Purdue University. Despite not earning credit for it, the students enrolled on the programme demonstrated full commitment and it is encouraging to see that the course has had a lasting impact on the way they conduct their research.

Those designing training and services for researchers in the realm of digital curation need to know the requirements and needs they are addressing. The Data Curation Profiles Toolkit, also developed by Purdue, is a popular resource for ascertaining those needs through interviews, but until very recently the usability of the tool had not been tested. Zhang et al. addressed this point by conducting a study of the perceived usefulness and perceived ease of use of the toolkit, the results of which are published in this issue. It is clear the study participants wanted the toolkit to be quicker and easier to use; some suggested creating online tools for preparing interviews, transcribing them, and developing the Data Curation Profiles themselves.

Data interviews were a key part of the methodology used to set up a catalogue of large, externally funded datasets for the benefit of researchers at New York University School of Medicine. Read et al. describe how the school went about selecting datasets to include and metadata to collect about them, recruiting local experts to provide guidance on using the data, and designing the catalogue interface around the needs of end users.

Profiling the needs of a discrete group of researchers is one thing, but how might one go about profiling the needs of a community that is not so clearly identified? Kim addresses this question in the context of preserving blog content from different communities. The paper presents a promising solution that uses content and link analysis to identify and compare different communities, perhaps paving the way for Web archives to derive quantified models of the designated communities they serve.

Identification on the Web is also the concern of Bolikowski, Nowiński and Sylwestrzak, though the issue they address is quite different. It is an oft-repeated truism that the ‘persistent’ part of ‘persistent identifier’ comes from an organisational commitment rather than any technological ingenuity, but that does not mean the technology is irrelevant. Inspired by cryptocurrencies such as BitCoin and initiatives such as LOCKSS, Bolikowski, Nowiński and Sylwestrzak suggest a fully decentralised system based on public/private keys, digital signatures and proof of cryptographic work to validate the minting and updating of identifiers. The idea is to eliminate dependency on any one organisation in the system, though I wonder if the extra robustness this gives above simple agreements between organisations (of the form: if I go under, you take over) is worth the environmental impact and cost of the extra computation. To be fair, they do offer some mitigations for this issue: I encourage you to read the article and decide for yourself.

Identifiers are vital for efficient known-item searches, but speculative searches depend on information on what a resource is about. For reasons of scalability, authors are often asked to provide this information themselves, so what questions should be put to them in order to get the best results? Willoughby, Bird and Frey consider three approaches: asking for descriptive words and phrases, asking specific questions about the content, and asking for search terms that might be used to find it online. They all have their uses, but what struck me was the confirmation that author-supplied metadata is richer in contextual information but noisier than metadata provided by a professional cataloguer.

Chao investigates the feasibility of a further alternative: text-mining journal papers for information to populate a dataset’s metadata record. To do this, she maps between the methods sections of soil ecology articles and the metadata schemes commonly used for soil ecology data. These mappings indicate a promising level of correspondence, but the gaps are more interesting in a way: they have implications both for the further development of the metadata standards and for the information that journals should encourage authors to provide.

While the task of turning the manual mappings into automated extraction routines might seem daunting, Chao might take comfort from Vellino, who seeks to do precisely this in the context of climate data. The Berkeley Earth project integrated climate data in some 20 different file formats, using a set of manual mappings to a common format.[2] Vellino asks whether machine-learning techniques might enable further formats to be understood without human intervention, using the existing mappings as a training set.

For a lesson in how to make metadata more amenable to transformation to a common format, we need look no further than the article by Parton et al., which describes the development of the MOLES 3 metadata standard by the Centre for Environmental Data Archival (CEDA) in the UK. MOLES 3 is a specialisation of ISO 19156 (Observations and Measurements), and may be straightforwardly transformed into the European INSPIRE profile of ISO 19115 (Geographic information – Metadata). Within those constraints, it was designed to preserve as much as possible of the expressive power of its predecessor, MOLES 2, so that the existing records could be migrated across automatically with a minimum of loss. This was particularly important as CEDA is the designated place of deposit for certain families of NERC-funded data, and thus had nearly 6000 records to migrate.

In the first of our non-conference papers, Swauger and Vision ask what factors influence where researchers deposit their data. Of the eight factors examined, the policies of funders, journals and institutions certainly come out highly, as do matters of subject specialisation, ease of submission, and the perceived trustworthiness of the repository. One of the ways repositories can demonstrate their trustworthiness is through certification; Swauger and Vision mention ISO 16363, but at the time of writing this is only three years old and certification is some way off for most repositories.

Of more immediate relevance is the Data Seal of Approval, which was developed between 2005 and 2008. Dillo and de Leeuw provide an excellent overview of the history of the scheme. They also discuss how it fits in with other certification schemes and outline what the coming five years might hold for it. Over 40 repositories have acquired the seal so far, and the aim is to raise this number to 60 by 2020. The list is dominated by specialist social science and humanities archives; I wonder if there are any plans to attract institutional data archives?

We have a trio of articles from the perspective of those providing institution-level support for research data management (RDM). Macdonald and Macneil provide an update on recent developments at the University of Edinburgh, including plans for a Data Asset Registry (in fact delivered by the institutional Current Research Information Management System), a Data Vault, and an Electronic Lab Notebook system that integrates with the rest of the data management infrastructure. Fitt, Rouse and Taylor describe how Oxford Brookes drew up its RDM policy and roadmap, and used tools such as the Data Asset Framework (DAF) and CARDIO to plan the implementation of those documents. Lastly, Cox and Williamson publish in some detail the results of a the DAF survey conducted by the University of Sheffield in 2014. The article lays down a challenge for other institutions

[2] Berkeley Earth: http://berkeleyearth.org/

to do likewise in order to facilitate benchmarking. There is certainly an appetite for this in principle, but as the DAF methodology can be finely tuned to particular local needs I fear the surveys will never be conducted with the level of consistency needed to make fair comparisons between institutions.

Moving from institutional perspectives to a national one, we come finally to our second non-conference paper. Hutař and Melichar reflect on the digitisation projects that took place in the cultural heritage sector in the Czech Republic between 2002 and 2014. In this 'long decade of digital preservation', a digitisation programme that was ramped up initially in reaction to widespread flooding developed into a sustained and large scale enterprise backed up by investment in archival infrastructure. The paper ends with a refreshingly frank assessment of the current state of affairs: while understanding of digital curation issues is not as pervasive as it perhaps should be, the authors are hopeful that by applying for ISO 16363 certification, the National Library's digital repository can lead the way in promoting good practice.

That brings to an end our look at the contents of this, the first issue of Volume 10. Happily, there is not long to wait before the first articles and papers of the second issue are published. These will include further decadal reflections originally presented at the International Digital Curation Conference, with a wide range of original peer-reviewed papers to which to look forward.