The International Journal of Digital Curation
Issue 1, Volume 4 | 2009

Guest Editorial: Research Data: It's What You Do With Them

Malcolm Atkinson,
e-Science Institute,
University of Edinburgh

March 2009

These days it may be stating the obvious that the number of data resources, their complexity and diversity is growing rapidly due to the compound effects of increasing speed and resolution of digital instruments, due to pervasive data-collection automation and due to the growing power of computers. Just because we are becoming used to the accelerating growth of data resources, it does not mean we can be complacent; they represent an enormous wealth of opportunity to extract information, to make discoveries and to inform policy. But all too often it still takes a heroic effort to discover and exploit those opportunities, hence the research and progress, charted by the Fourth International Digital Curation Conference[1] and recorded in this issue of the *International Journal of Digital Curation*, are an invaluable step on a long and demanding journey.

The requirements are widely recognized, the Digital Curation Centre's Curation Lifecycle Model (Constantopoulos et al., 2009; Digital Curation Centre [DCC], 2008) clarifies the generic issues faced by a professional digital curator. Facilitating the complementary cycle from the data users' viewpoint of discovery, access, extraction, transformation, synthesis, analysis and publishing (more data) is the ultimate goal. Another large population of users contributes data, they carefully count plant species, record bird song, count insects or type in text to blogs. Others plan and build high-precision instruments and laboratories and drive them to generate more data that are contributed to existing or new repositories.

The requirements and the complexity of this diverse data ecosystem are widely recognized. The recent US report by the Interagency Working Group on Digital Data (2009) presents "a strategy to ensure that digital scientific data can be reliably preserved for maximum use in catalyzing progress in science and society". The European Union's INSPIRE directive (2007) seeks to bring consistency in geospatially referenced data, so that they can be shared across the Union; it is now referencing nearly 50 standards. It could be seen as an early of example of "farming" in this ecosystem – arrange the crop in a regular pattern and make it easier to harvest.

[1] Past International Digital Curation Conferences http://www.dcc.ac.uk/events/events_archive?et=3

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.

It has never been more evident that research into digital curation is necessary and valuable. The opportunities to deploy the research and gain significant benefits are growing rapidly. There are already outstanding examples of large and well-curated bodies of scientific data, such as those curated by the European Bioinformatics Institute, the National Center for Atmospheric Research (NCAR) (Jacobs & Worley, in press) and the British Atmospheric Data Centre (Lawrence, 2008). Their collections, policies and reputations are built on decades of subject-specific work. These towering examples of data organisation and data publishing continue to face immense challenges to keep up with the increased flow of data, the demands for computational access and analysis, and the new forms of data. These challenges are magnified as these separate resources operate in a global system where the research and policy questions require the integration and analysis of data from hundreds or even thousands of such resources. In both disciplines, much has been done to develop frameworks and standards that support semantic data integration *within their discipline*. But a research question today, such as understanding how a coastal ecosystem and the human community it supports are being affected by climate change, can draw on bioinformatics resources, social and geographic data, atmospheric data and oceanographic data. The challenge of facilitating such questions across traditional silos will change the data ecosystem itself.

But the focus must not rest only on the large data resources and the globally endorsed research questions; there are immense information riches in the much larger population of small data collections, the individual researcher's, the local research team's work, the archives of a small club of specialists observing populations of lichens and mosses over years. Many of today's key reference datasets will have had such modest beginnings. The Protein Data Bank (PDB)[2] started with just 7 biological macromolecular structures in 1971; today it holds more than 56,000 and serves more than 140,000 distinct users each month.

This means that the individual researcher and small group need unobtrusive help. Their data will start at a time when their primary imperative is their research goal, not data sharing. But practitioners who have spotted an ecological niche, a new key data requirement, will find their data in demand. Providing a priori requirements and insisting on data publishing plans before they start will either put them off so much that they fail to get off the ground, or slow them down so much that someone else grabs the niche. So just how do you provide the right help, the right tools and the right services at the right time?

The same democratization and easy access ramp are required for researchers who wish to use new combinations of data. They may have a new insight or intuition that will lead to important questions and answers for their research field, to local or global decisions and ultimately shape policy. There are normally hurdles to overcome, finding, understanding, accessing and transforming the data, and so on. If this occurs across the walls of traditional discipline boundaries, the investments that have been made to facilitate data exchange within the silo (approaching 50 standards for INSPIRE) introduce an insurmountable barrier of complexity. Yet those standards are still insufficient to facilitate the integrations needed by climate modellers (Millard et al., 2007). How can we simultaneously support new uses, often quite unexpected, and support the growing intensity of research questions inside a silo? Dramatic changes of

[2] Protein Data Bank (PDB) http://www.rcsb.org/pdb/

use are well illustrated by ships' logs over the centuries: initially data are recorded to improve safety during voyages; after collection in museums, they are poured over by historians, economists and sociologists to understand trade and life of that time. Today, their oceanographic and weather observations are digitized to calibrate long-term atmospheric models.

Partly because of the new digital data resources, every subject is making innovations in the way it collects and uses data. New instruments and observing procedures are introduced. New aspects of phenomena are recorded. New phenomena are recognized. New catagories are introduced and taxonomies are revised. New models and data mining strategies are used to calibrate, derive and extract data. These changes are key to the progress of research and must not be inhibited by the commitments made to meet the existing requirements. Yet such changes introduce extra challenges for the researcher who needs to integrate information from time series that traverse their introduction. These changes may disrupt well-established usage patterns and large investments in scientific workflows that used the data in their previous form. For integrating workflows that draw data from many different sources, their *integrated* changes may overwhelm the resources available for adapting to change. How can we facilitate and take advantage of the advances created by innovation and still use the existing investments in data-intensive research when changes are not relevant?

The resource about which we are most concerned, as we consider all of these data curation issues, is the time and intellectual effort of the combined community of data creators, data curators and data re-users. Individuals have to be enabled to pursue their role with minimum unnecessary labour and distracting complexity. But today we are aware as never before that energy is also a crucial resource, primarily because most means of generating it exacerbate global warming. Yet every spinning disc, every computer cycle and every data movement in the data handling systems built to support research and concommitant data uses, is costing energy. The data curation community therefore has to consider how to provide the world's research data services with minimum environmental damage. Perhaps this will be done by showing how data curation can be performed in remote centres using geothermal, wind and tidal current energy, in pooled data centres with all-optical technology for all data transmission. A transition to such forms of remote curation and data sociability is probably more a sociological than technical challenge, but there are certainly still technical problems to resolve.

The growing wealth and diversity of data, composed from many sources, collected through many instruments, measuring many different variables, using a wide variety of techniques, all varying over time, provide an immense potential for discovery. However, they have reached a degree of complexity that is already beyond the comprehension of individuals in many disciplines. The composition of more and more sub-models into multi-scale and multi-phenomena system models also challenges our ability to understand. As models are used to calibrate instruments, extract signals from high-amplitude backgrounds, normalize data and derive data products, while data are used to calibrate and validate models, there are potentially tautological arguments and error magnifiers.

Yet the goals of modern research drive us to look at systems of even greater complexity while the advances in technology tempt us to increase every scale factor. There are deep questions here about whether research is exhibiting the hubris of overreaching itself by building complex and interconnected conceptual and digital systems that are so coupled that they are liable to the forms of catastrophic collapse that the global financial systems recently exhibited. It is therefore time to redouble the effort to understand, explain and formalize these e-Science methodologies. It is time to understand how they should be applied in disciplines so that confidence in results is well founded and accurately stated. It is time to understand how to compose safely the different components researchers use. It is time to make the adoption of reliable methods easier, so that more researchers adopt them to improve their research.

The progress manifest in the differences between the first and fourth International Digital Curation Conferences represents a significant advance. It is a first step on a long and demanding journey.

References

Constantopoulos. P., Dallas, C., Androutsopoulos, I., Angelis, S., Deligiannakis, A., Gavrilis, D., et al. (2009). DCC & U: An extended digital curation lifecycle model. *The International Journal of Digital Curation, 4*(1).

Digital Curation Centre. (2008). *The DCC curation lifecycle model*. Retrieved March 4, 2009, from Digital Curation Centre Web site http://www.dcc.ac.uk/docs/publications/DCCLifecycle.pdf

European Parliament and Council. (2007). *Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an infrastructure for spatial information in the European Community (INSPIRE)*, published April 2007, Retrieved February 28, 2009, from http://eur-lex.europa.eu/LexUriServ/site/en/oj/2007/l_108/l_10820070425en00010014.pdf

Interagency Working Group on Digital Data. (2009). *Harnessing the power of digital data for science and society*. Report to the Committee on Science of the National Science and Technology Council. Retrieved February 28, 2009, from http://www.nitrd.gov/about/Harnessing_Power_Web.pdf

Jacobs, C.A., & Worley, S.J. (in press). Data curation in climate and weather: Transforming our ability to improve predictions through global knowledge sharing. *The International Journal of Digital Curation, forthcoming 4*(2). Retrieved February 28, 2009, from http://www.dcc.ac.uk/events/dcc-2008/programme/accepted-papers

Lawrence, B. (2008). *Costing metadata for curation*. Presentation to the Fourth International Digital Curation Conference, Edinburgh, UK. Retrieved February 28, 2009, from http://www.dcc.ac.uk/events/dcc-2008/programme/

Millard, K., Atkinson, R., Woolf, A., Stock, K., Longhorn, R., Higgins, C., et al. (2007). Developing feature types and related catalogues for the marine community - Lessons from the MOTIIVE project, *International Journal of Spatial Data Infrastructures Research*, Vol. 2, 132-162. Retrieved March 4, 2009, from http://epubs.cclrc.ac.uk/bitstream/2138/millard_etal.pdf