

# The International Journal of Digital Curation

Issue 1, Volume 4 | 2009

## The Third DCC-RIN Research Data Management Forum

Martin Donnelly,  
Digital Curation Centre,  
University of Edinburgh

May 2009

### Summary

A report on the third meeting of the Research Data Management Forum which was held in Manchester, UK on April 30 and May 1, 2009, with an overarching theme entitled “Value and Benefits”. The event was co-sponsored by the Digital Curation Centre (DCC) and the Research Information Network (RIN).

## Introduction

The third meeting of the Research Data Management Forum was held in Manchester on April 30 and May 1, 2009, co-sponsored by the Digital Curation Centre (DCC) and the Research Information Network (RIN). The event took “Value and Benefits” as its overarching theme.

### April 30, 2009

Welcoming the 42 delegates – comprising senior decision-makers, repository and data centre managers, digital librarians and academic/clinical researchers – to the event, DCC Associate Director **Liz Lyon** and RIN Head of Programmes **Stéphane Goldstein** each noted the timeliness of the event’s topic, and in particular the political sensitivities which currently surround issues of impact, value and cost. However, each took care to stress that that the event was not just to be concerned with financial values and benefits, and not just about costs.

The discursive element began with a keynote presentation<sup>1</sup> from **Astrid Wissenburg** of the Economic and Social Research Council, on the topic of “Value and benefits of data sharing and management: why, what, for whom and how?”

Astrid began by asking “What do we mean by value and benefits in the data sharing and management context?”, underlining the timeliness of the event by brandishing the current edition of the *Times Higher Education*, which providentially carried a cover story on measuring impact (Corbyn, 2009) and a related opinion piece by Philip Esler, Chief Executive of the Arts and Humanities Research Council (AHRC) (2009)<sup>2</sup>.

So why are the Research Councils obsessed with measuring impact and benefits? There are many factors, not least of which are political. For starters, it is likely that the new Research Excellence Framework (REF) will address impact more heavily than its predecessor, the Research Assessment Exercise (RAE); indeed, the Research Councils have already begun to make this more explicit in the funding application paperwork.

What kinds of impact are we talking about here? Essentially, the Research Councils are concerned primarily with economic impact/payback, while the research community is more focused on impact in terms of the advancement of knowledge. From a Higher Education perspective, academic impact is foremost; this in turn leads to economic/political/societal impact. But return on investment may take several years to become manifest, and the return may not be transparently economic. Different benefits of data may become apparent pre- and post-project, as well as emerging during the project lifecycle. Astrid gave the example of the 1958 Birth Cohort Survey, which had a significantly widespread impact on attitudes to smoking during pregnancy, demonstrated by falling rates of smoking among expectant mothers. However, drawing the causal connections, identifying the specific benefits (and, indeed, stratifying them) is difficult over the short, medium and long-term.

Who benefits from research data? The answer to this, at least, was simple:

<sup>1</sup> All presentation slides are available via the RDMF3 event page on the DCC website <http://www.dcc.ac.uk/events/data-forum-2009-april/>

<sup>2</sup> The discontinuation of the AHRC’s Arts and Humanities Data Service, of course, remains a sore subject for many in the data management community.

everyone, in all communities, from data creators, to other researchers (be they academic or political), to the end-users of research. The “how” is trickier. Astrid cited a growing portfolio of evaluation methodologies which are used to “measure” economic and societal impact, and a number of studies on the topic, including the Department for Innovation, Universities & Skills (DIUS) study on large-scale facilities in the UK, which demonstrated local economic benefit derived from local employees spending their salaries in local businesses, and the Research Councils UK (RCUK) ‘Excellence with Impact’ study (2007) which includes 18 case studies drawn from across the research councils. Frameworks are required, such as the UK Economic Impact Reporting Framework (EIRF)<sup>3</sup>, but there is a perception that this can be difficult to get to grips with.

The ESRC has two primary methods for policy and practice impact evaluations: the “payback” method, and “tracking forward.” The use of qualitative studies with mixed methods captures complexity and enables triangulation, and by drawing links between academic research and policy development (attribution, traceability, measurability and time lags) we begin to see direct and indirect pathways emerge; from outputs to impacts, to value and benefits.

Astrid ended her talk with a caveat which would be echoed by another speaker on the following day: that it is all too easy to spend more money on studying impact than the original research funding is worth, or for that matter the value of its benefit. In short, impact assessment cannot afford to become a runaway train.

The open discussion which followed Astrid’s presentation covered a variety of themes, including the maturity of appraisal methodologies, and the increase in value which occurs when datasets are well structured and able to be reused and combined with other data, thereby contributing to bodies of knowledge. Lastly it was noted that the ongoing lack of a uniform means of citing and crediting data (and, indeed, graphical representations of data in the shape of the “killer graph”) contributes to the difficulty of tracing final benefits and outcomes back to their original source.

## May 1, 2009

The first session of the day two addressed “the policy perspective,” with **Neil Jacobs** and **Simon Hodson** from JISC and **Stéphane Goldstein** from RIN each taking the opportunity to outline their respective organisational and operational views on the topic.

Neil began by citing a large number of reports: the OSI Infrastructure working group, Liz Lyon’s “Dealing with Data”, RIN’s “To Share or not to Share”, SPECTRA, Key Perspectives ‘skills Role and Career Structure’, his own “Keeping Research Data Safe”, JISC’s “IPR and Licensing Issues”, and the UK Research Data Service (UKRDS) feasibility study. This collection of reports served to demonstrate the multiplicity of current claims which relate to impact, but they provide little hard and fast evidence. In order to progress this, two classes of mechanism are required: firstly, a cost-benefit study and business case for an infrastructure which supports better monitoring of outputs and outcomes; and secondly, an enriched technical infrastructure which covers generic tools and domain-specific tools, both for researchers and for

<sup>3</sup>UK Economic Impact Reporting Framework (EIRF) <http://www.rcuk.ac.uk/aboutrcuk/eirf>

programme managers. JISC is supporting improvements in sector capacity and skills via initiatives such as the DC101 training programme and UK participation in the IDEA working group.

**Simon Hodson** gave an introduction to some specific aspects of JISC's new Research Data programme, noting that the DCC plays a key role within JISC activities in raising skills and capacity, and in linking this with subject-specific curation expertise. Simon gave a list of future work related to the new Research Data programme: firstly, in early June there will be a workshop at NeSC on the Data Audit Framework and related work; the programme's centrepiece will be a £1.5 million June call for proposals related to institutional data infrastructure; there will be a support project to pull together a picture of suggested business models for data management; there will be more value- and benefits-focused case studies, building on the findings of the SCARP<sup>4</sup> series; and there will be continued policy-related discussions with the research councils, with a view to continuing to embed a data management factor within council-funded research. The business case for research data is of key importance, and an advisory group will be established to produce a roadmap for it. Simon cited the UKRDS feasibility study, quoting to the effect that "not all datasets have potential value." Triage is therefore required, with greater study of a wider sample. Essentially we need to know who benefits, how they benefit, and how this can be quantified.

RIN's **Stéphane Goldstein** then unveiled a draft scoping document produced jointly by RIN and JISC for a project which will make a business case for improving the data-sharing environment by considering factors such as usage patterns, impact, evidence bases, timeliness and collaboration. Contributing to a larger and wider debate, this will need to benefit all types of stakeholder: funders, HEIs, researchers, data managers, government and other policy makers, although Stéphane was at pains to stress that the benefit need not necessarily be expressed in terms of financial value. The project is expected to run from June 2009 to March 2010, with a budget of £150,000.

**Jenny Fry** of Loughborough University then spoke about her co-authored JISC-commissioned report, "Identifying benefits arising from the curation and open sharing of research data produced by UK Higher Education and research institutes." (Fry, Houghton, Lockyer, Oppenheim & Rasmussen, 2008). This report addresses the problem of time lags between expenditure on research and the measurement of its subsequent impact. Fry and her co-investigators took an embedded case study approach, which found that while costs are easier to quantify than benefits, both costs and benefits can be either direct or diffuse. There may also be tricky ownership issues attached to datasets; these vary between domains, and can be particularly complex in the social sciences.

**Matthew Woollard** of the UK Data Archive (UKDA) began his talk on "Collecting the metrics: social science as an exemplar discipline" with an observation that attaching monetary value is difficult when it comes to datasets: there are no auctions to help establish this!

But there are more than just monetary values which we can attach to data, and Matthew listed a few varieties from a social science perspective. First and foremost is

<sup>4</sup> DCC SCARP Project <http://www.dcc.ac.uk/scarp/>

the data's value to the original research: even if the data are not subsequently reused (or reusable), they may yet have justified the investment in their collection or creation. And it may take a long time for a dataset to be reused. Matthew gave an example of a dataset that was not downloaded for 19 years, and then downloaded 15 times in quick succession: if the archive's retention policy had been set at less than 19 years, these long-awaited reuses might not have been possible. We can use number of accesses as one indicator of value, while remaining mindful that audiences and trends will vary over time, but it is important to remember that accesses are not the same as uses, and therefore not ideal for measuring impact.<sup>5</sup> Data managers and funders can look at citations and web analytics to inform their judgements, but – as is the case with all statistics – they need to be interpreted sensibly and responsibly.

Matthew ended with a warning similar to Astrid's yesterday, namely that we need to be sensitive to the costs of impact assessment, and thus careful that we do not spend more money and effort in evaluating datasets than we do in creating, curating and preserving them.

**Jenny Walsby** of the British Geological Survey then spoke on “Benefits beyond the sector: repurposing data with industrial partners”, focusing her talk on data collection, research, modelling, and visualisation. Jenny cited numerous partnerships with the public and private sectors,<sup>6</sup> and various ways and means of accessing and repurposing geological data via discovery metadata and geographical information systems, in addition to third-party data delivery via NORA, eMapSite, and Wikipedia.

The British Library perspective was provided by **Adam Farquhar**, who drew attention to the gap between published research and the datasets that underlie them. Published work tends to be held by libraries, and the data by data centres, so there is traditionally a disconnect between the two. Around 45% of journals currently provide access to data related to papers, but there are no accepted rules on how to publish, present, cite or catalogue datasets; so if papers and data are to gain an equal standing, we need a method to identify datasets persistently. Digital Object Identifiers (DOIs) can be used to do this via the Joint Registration Agency: a European and global infrastructure which works with publishers and research councils to assign and issue identifiers for datasets.

The last discursive talk of the morning came from **Phillip Lord** of Newcastle University, who gave an entertaining view on “Why everybody else should share their data.” The basic need from a researcher's point of view is for metrics which demonstrate that better-annotated data lead to increased numbers of citations, leading to more published papers, and therefore greater recognition and esteem. The problem is that these metrics are not readily available.

Referring to Adam's presentation, Phillip spoke of the need to share code as well as data, lamenting that there is no great tradition of this in neuroscience. Data are considered as artefacts, while code is merely “a snapshot of a development process.”<sup>7</sup>

<sup>5</sup> Additionally, Matthew reported that for the UKDA, downloads of documentation tend to outstrip the number of downloads of the datasets themselves by approximately a factor of ten.

<sup>6</sup> Insurance companies in particular are interested in swelling or shrinking clays which can cause damage to property.

<sup>7</sup> We should note, of course, that data are often also dynamic, so a dataset in turn becomes a snapshot of a different development process.

Since no one cites software, it follows that there is no parity of esteem between doing science and writing code, hence code sits below both publications and datasets on the scale of esteem.

The last item of the morning was an update on actions from the last RDMF event provided by DCC Deputy Director **Graham Pryor**, including a diagram which aligns a number of core skills and competencies to particular categories of role in the data management process,<sup>8</sup> and a forthcoming white paper which will tackle the issue of training, development and certification for data professionals.

After lunch, the delegates split into three breakout groups, which addressed metrics, quality management, and outcomes. It is difficult to do justice to the breadth of these discussions in this format, but issues addressed included the risk of misinterpretation of metrics, the need to specify achievable metrics that are aware of the limitations of data, and some blue-sky thinking on an automated mechanism which would ensure that the cost of collecting the metric is justified in terms of its evidential benefit.

With regard to data quality, the delegates urged the Research Councils and other major funders to continue to collaborate in developing coherent data policies that seek to maximise the opportunities for publicly funded data to be shared and accessed, but which are at the same time sensitive to domain differences.

The Forum's fourth meeting is expected to be held in Autumn or Winter 2009. Full details will be released via the RDMF blog<sup>9</sup> and the JISC Research Data Management mailing list<sup>10</sup>, so do subscribe to these to be kept in the loop.

## References

- Corbyn, Z. (2009, April 30). Bang for your bucks. *Times Higher Education*. Retrieved April 30, 2009, from <http://www.timeshighereducation.co.uk/story.asp?sectioncode=26&storycode=406311>
- Esler, P. (2009, April 30). We inhabit a changed world, and it is our duty to engage with it. *Times Higher Education*. Retrieved April 30, 2009, from <http://www.timeshighereducation.co.uk/story.asp?sectioncode=26&storycode=406338>
- Fry, J., Lockyer, S., Oppenheim, C., Houghton, J., & Rasmussen, B. (2008). *Identifying benefits arising from the curation and open sharing of research data produced by UK Higher Education and research institutes*. Report commissioned by JISC.

<sup>8</sup> Research Data Management Forum: RDMF2: Core Skills Diagram <http://data-forum.blogspot.com/2008/12/rdmf2-core-skills-diagram.html>

<sup>9</sup> Research Data Management Forum <http://data-forum.blogspot.com/>

<sup>10</sup> Contact: [research-dataman@jiscmail.ac.uk](mailto:research-dataman@jiscmail.ac.uk)

Research Councils UK. (2007). Excellence with impact: Progress in implementing the recommendations of the Warry Report on the economic impact of the Research Councils. Retrieved May 28, 2009, from <http://www.rcuk.ac.uk/innovation/impact/default.htm>