Research data and libraries: who does what

Based on a paper presented at the 35th UKSG Conference, Glasgow, March 2012

A range of external pressures are causing research data management (RDM) to be an increasing concern at senior level in universities and other research institutions. But as well as external pressures, there are also good reasons for establishing effective research data management services within institutions which can bring benefits to researchers, their institutions and those who publish their research. In this article some of these motivating factors, both positive and negative, are described. Ways in which libraries can play a role – or even lead – in the development of RDM services that work within the institution and as part of a national and international research data infrastructure are also set out.

The last few years have seen significant attention and funding devoted to the effective management of research data, by government, research funders, researchers, their host institutions and a range of other stakeholders. What roles (if any) do librarians and publishers have to play? This has been the subject of some debate, not all of which is settled. I will try first to explain why this issue is assuming such importance, particularly for the universities where the research takes place. Then I will put forward a view on some of the roles libraries should be taking on – and some they should avoid. One thing that I hope will become apparent is that many of the motivating factors about proper research data management (RDM) are not specific to research data.

KEVIN ASHLEY
Director
Digital Curation Centre

Why care?

The primary motivating factor, the one that would cause us to care about this issue even without any other motivation, is data re-use. Data is expensive to collect and therefore represents an investment of some sort. Much data is observational in nature and cannot be replicated. This is true whether it comes from a satellite observing the earth's weather, a survey of people's voting intentions in Glasgow in early May 2012, or a collection of reports from the public of the birds seen in their garden on a particular day. Each of those data sets involves an instrument (a satellite, a survey tool, citizen science data collection) that can be used repeatedly, but the data that results will typically be different. The value of such data is thus arguably represented by more than the simple cost of acquisition. All data, whether observational or not, has some potential for re-use. The potential varies, as does the cost of retention of data, so this is not an argument that everything needs to be kept forever. Rather, intelligent decisions need to be taken to balance the cost of data generation, the cost of retention and the potential value of re-use. In some cases, it really is cheaper to throw the data away and generate it again when needed.

> " … intelligent decisions need to be taken to balance the cost of data generation, the cost of retention and the potential value of re-use. In some cases, it really is cheaper to throw the data away and generate it again when needed."

Re-use must also be considered outside the strict context of research in the discipline where the data is generated. Many of the most interesting re-use cases come from cross-disciplinary research. Data re-use can also occur outside academic research; re-use in learning and teaching is one example which is still of primary importance to universities. Instilling good practice in the collection, description, choice and analysis of data in the students of today is often best done with real examples of data – both good and bad – collected for research purposes, rather than with artificially-constructed examples. Data can also have re-use potential outside the academy, to inform public policy and to support commercial activities. Similarly, data generated within these non-academic domains is often of value to academic research. There is a striking parallel here with the re-use of administrative, bibliographic and activity data from libraries for other purposes, not all of which are of direct benefit to the libraries themselves.

There are other reasons why good data management is important. Statutory and regulatory requirements can force us to manage data properly[1]. In the UK they include laws relating to data protection and freedom of information (FoI), and requirements imposed on us by research funders; The Engineering and Physical Sciences Research Council (EPSRC), one of the largest UK funding bodies, will require data from its funded projects to be retained for ten years after its last use – for popular data, this could mean forever. There are also specialist requirements that impinge on specific areas of study such as clinical trials.

The good conduct of research and the reputation of the institutions where it is conducted are also strong drivers for effective data management. Where data supports the conclusions of published research it is important that it is available to those who wish to question or validate that research, and many publishers and disciplines now require that the data behind a publication is located somewhere that can give it a permanent home and make it available to others. Many cases of research misconduct in the past would have been exposed much earlier if this had been done. Some journals, such as those of the International Union of Crystallography[2], take this even further with a single submitted file being both the published article and the data which supports it. This allows the online publication to provide for direct exploration of the data.

Finally, studies have shown that the impact of research is enhanced by the availability of the data behind it[3,4] and by effective linkages being put in place between publication and data, which typically reside in different custodial regimes. It results in increased citations to the research and to the data and increased re-use. This is a benefit that accrues to researchers, their institutions and to publishers and it is in all their interests to collaborate to make data citation mechanisms more widespread, straightforward and effective.

We can summarize the motivating factors as being in two classes: some mean that bad things happen if we don't practise good RDM; others mean that good things will come to those who do practise good RDM.

What to do?

Assuming that you are now convinced that this is something for you to care about, I hope you are now asking what you or your institution can do about it. That is one of the areas where my organization, the Digital Curation Centre (DCC), aims to help. We provide guidance and tools to help with all aspects of the data management lifecycle and the development of capability and processes for effective RDM within institutions. We have identified a number of roles which we think libraries are often well-suited to undertake. Some relate well to existing skills already present in research libraries, and the task of appraisal and selection is one of those. As already mentioned, not everything can be kept forever, so we need to make informed choices about what needs to be retained and for how long. Although some of the factors involved require domain knowledge that only the data creators will have, many general principles are also involved and libraries should be in a good position to develop these as part of institutional policy.

> " … most institutions, even most research departments, do not have a good understanding of what data they have, who is responsible for it now and who should be responsible for it in the future."

One other task librarians can undertake is in the initial audit of what data already exists and who is responsible for it. The DCC has developed a tool called DAF (the Data Audit Framework) that can help structure this process. Experience with using it shows that most institutions, even most research departments, do not have a good understanding of what data they have, who is responsible for it now and who should be responsible for it in the future. Not having such basic knowledge can hamper forward capacity planning and can also lead to serious legal consequences if subject access or FoI enquiries are made relating to this data. To deal with these properly in the longer term really requires an institution to develop a data registry, which can also help to expose knowledge about its research data outputs to others and be a target for links to data from publications. Operating such a registry is another area where libraries can take a leading role, but they would need to work effectively with research offices which will have an increasing requirement to link data sets to the research grants which funded their creation. In maintaining data registries which fit into national and international data discovery infrastructures, libraries can play a key role in ensuring that data is preserved, citable, findable and reusable.

> " … libraries can play a key role in ensuring that data is preserved, citable, findable and reusable."

Libraries can also act in a leadership role in galvanising institutional action on RDM, responding to external pressures such as the EPSRC requirements. Libraries cannot solve these issues alone, but they can be the catalyst for bringing together those in IT services, research computing, research administration and the researchers themselves to find institution-wide solutions. The DCC is already working with a number of universities to do this and will produce case studies to help others understand how to replicate good practice. In the meantime, the DCC's CARDIO tool can help you scope the problem, identify what is already in place and the appropriate next steps to take.

There are some things you should avoid doing. In particular, a number of research disciplines already have good data centres available which do an excellent job of all the tasks I outlined above and many more. Where they exist, researchers will want to use them and you should not stand in their way, although you might be able to streamline the process. You'll probably still want a metadata record for your institutional data registry which can point to the data centre copy. You also should not try to work alone, in particular without the support of your researchers, from Pro-Vice-Chancellor level downwards.

Remember, do this job well and you and your institution will benefit; do it badly or not at all and your reputation and future funding could be at risk. All the material described above and much more can be found at the DCC website (www.dcc.ac.uk).

References

1. DCC Policy & Legal resources home:
   http://www.dcc.ac.uk/resources/policy-and-legal (accessed 28 May 2012).
2. Example of an International Union of Crystallography journal: *Crystal Structure Communications Online*:
   http://journals.iucr.org/c/ (accessed 28 May 2012).
3. Piwowar, H, Day, R and Fridsma, D, Sharing Detailed Research Data Is Associated with Increased Citation Rate, *PLoS ONE*, 2007, 23.:e308 (accessed 28 May 2012).
4. Pienta, A M, Alter, G C and Lyle, J A, *The Enduring Value of Social Science Research: The Use and Reuse of Primary Research Data:*
   http://hdl.handle.net/2027.42/78307 (accessed 28 May 2012).

Kevin Ashley, Director, Digital Curation Centre
E-mail: kevin.ashley@ed.ac.uk

To cite this article:

Ashley, K, Research data and libraries: who does what, *Insights*, 2012, 25(2), 155–157, doi: 10.1629/2048-7754.25.2.155

To link to this article:

http://dx.doi.org/10.1629/2048-7754.25.2.155