Les métiers liés aux données de la recherche: Data Librarian

Alex Ball
Digital Curation Centre (DCC)/UKOLN Informatics, University of Bath

2013-09-19

——— *Slide 1* ———

Title slide

Hello, my name is Alex Ball and I work for the Digital Curation Centre (DCC).[1] We are a centre of expertise in research data management primarily serving the UK Higher Education sector, but we do a lot of work internationally. For instance, we are currently involved in the European 4C Project: 'the Collaboration to Clarify the Costs of Curation'.[2] Since we began work in 2004 we have seen research data management grow enormously in importance, and I am pleased to be able to talk to you today about the effect this is having on the role of information professionals.

There are some fields of research where datasets are highly valued. Usually this is because they record unrepeatable observations, or are expensive to collect. In these fields, there has always been a culture of data sharing and good data management. But over the past decade there has been a push to make sure all research data is properly managed.

——— *Slide 2* ———

OECD Declaration

In 2004, the OECD issued its *Declaration on Access to Data from Public Funding* with support from 34 countries including the UK and France.[3] This called for publicly funded research data to be made open as far as possible, so that the results could be better scrutinised and to increase the transparency and efficiency of the research enterprise.

——— *Slide 3* ———

UNESCO Policy Guidelines

At the same time, UNESCO was arguing that governmental information, including data from publicly funded research, should be better organised, disseminated, and placed in the public domain.[4] Again, the argument centred on accountability, transparency and scrutiny.

——— *Slide 4* ———

1. DCC: `http://www.dcc.ac.uk/`
2. 4C Project: `http://4cproject.eu/`
3. OECD Declaration: `http://acts.oecd.org/Instruments/ShowInstrumentView.aspx?InstrumentID=157&InstrumentPID=153`
4. UNESCO Policy Guidelines: `www.fas.org/sgp/library/unesco_govinfo.pdf`

Open data in the UK

In the UK…

- In 2008, the Research Information Network published a policy framework that called for clarity on the roles and responsibilities of research data management.[5] It argued that data should be quality assured, preserved, made accessible, and routinely reused and cited.
- In 2012, the Royal Society report *Science as an Open Enterprise* argued that research data should be 'intelligently open'; that is, easy to find, intelligible, open to assessment, and in a reusable form.[6] It also argued that disseminating scientific conclusions without the supporting evidence is a form of malpractice.

——— *Slide 5* ———

This year, there have been two major developments for UK government data.

- In May, an independent review of public sector information by Stephan Shakespeare recommended the UK Government should start releasing publicly funded data as rapidly as possible, but also publish a National Core Reference Data set to which data is added after full quality assurance.[7]
- In June, the G8 leaders signed an Open Data Charter which says all government data should be published openly by default, in order to increase accountability as well as to support innovation and provide economic opportunities.[8]

——— *Slide 6* ———

UK research data management

Our major funding bodies have also come around to the benefits of open data. In 2005, RCUK (which represents all our research funding councils) issued a position statement promoting open access to research outputs. Between 2007 and 2011 five major funders made data sharing a condition of funding. But perhaps the strongest requirements came from the Engineering and Physical Sciences Research Council (EPSRC), which threatened to stop funding universities that did not provide a research data management infrastructure for their researchers.

——— *Slide 7* ———

EPSRC expectations

The EPSRC expects universities to do things that have traditionally been done by data centres:[9]

- preserving data,
- making sure it is identifiable and citable,
- making sure it can be accessed long into the future,
- keeping track of any restrictions on access or use.

5. RIN Policy Framework: `http://www.rin.ac.uk/our-work/data-management-and-curation/stewardship-digital-research-data-principles-and-guidelines`
6. Royal Society Report: `http://royalsociety.org/policy/projects/science-public-enterprise/report/`
7. Shakespeare Review: `https://www.gov.uk/government/publications/shakespeare-review-of-public-sector-information`
8. Open Data Charter: `https://www.gov.uk/government/publications/open-data-charter`
9. EPSRC Expectations: `http://www.epsrc.ac.uk/about/standards/researchdata/Pages/expectations.aspx`

So in effect, UK universities will have to start running their own data centres. But getting to that point will be a lot of hard work, nd they will need help.

——— *Slide 8* ———

DCC Institutional Engagements

For the past two years, the DCC has been running a programme of Institutional Engagements. In an Institutional Engagement, we offer a university 60 days of effort, where we help them to improve how they manage their data. So far, we've helped 21 different UK universities in this way.

We would normally start an engagement by getting agreement from the Pro-Vice Chancellor for research, but most of our work would be with a working group or committee set up to look at data issues. This group might include senior researchers and representatives from the Research Office, IT Services, the Records Office, and so on, but it would normally be dominated by the Library. Why is that?

——— *Slide 9* ———

Librarians: existing skills & knowledge

Libraries are taking the lead because

- they have a highly relevant skill set:
  - they already know about how to organise and document information to make it easier to find and put into context, so it is not a big leap to transfer those skills to data;
  - they already teach information literacy to students, so are well placed to teach research data management;
  - their experience of running publication repositories can be transferred to running data repositories;
  - they already know how to liaise with departments, and negotiate with publishers;
- they already have good relationships with researchers, so it is easier for them to provide support;
- they have been leading the way towards open access to research publications, so it seems natural to extend that to data as well.

——— *Slide 10* ———

Librarians: new skills & knowledge

So librarians are the right people, but they need new skills. The sorts of things that universities ask us to help with, reveal the range of new or enhanced skills that data librarians need to have.

- They need to become detectives to uncover research data scattered across personal hard drives and private departmental servers.
- They need to become consultants, interviewing researchers about their current workflows and practices, and identifying where improvements can be made.
- They need to bring their negotiation skills to bear on senior managers, and know where to find convincing evidence of the need for and benefits of data management.
- They need to become expert advisers to policy makers, perhaps even drafting university data policy themselves.

- They need to adapt their existing skills with repositories to the specific demands of data. They may even need to learn how to be data publishers so the data can be reliably cited.
- And of course, they will need to know how to support and train researchers in a range of data management issues.

We help librarians and others to gain these new skills by

- providing them with tools and training,
- by being at their side when they try things out for the first time, and
- by answering their questions and checking their work for them.

Here are a few examples.

——— *Slide 11* ———

Data Asset Framework

If a librarian needs to set up and populate a registry of all the data held by a university, we would guide them through how to run an audit based on the Data Asset Framework.[10] This is a highly flexible method for answering questions like

- how much data are researchers generating?
- where is it being stored?
- who is looking after it?
- how are they doing that? and
- what might its long term value be?

——— *Slide 12* ———

University of Bath Research Data Survey

Having said that, at Bath we used an online survey of principal investigators, research officers and postgraduates to get a rough idea of those answers.[11] And they turned out to be typical for a UK university.

- There were legal and contractual barriers to much of the data being shared. Could any of it have been made more open if researchers had negotiated harder at the outset, or used a different consent form?
- Data was being stored all over the place. Some in the shared storage provided by the university, but other data was kept in more risky locations like personal hard drives and USB sticks.
- No wonder data was being lost.

——— *Slide 13* ———

10. DAF: `http://data-audit.eu/`
11. Research360 requirements report: `http://opus.bath.ac.uk/36361/`

CARDIO

To evaluate current data management practice and work out what the next steps should be, we have another tool called CARDIO.[12] We think of it as an institutional health check for data management. Using it is a collaborative process: researchers, academics and support staff all contribute to a consensus brokered, more often than not, by a librarian.

The results can be used

- to build a business case for investing in data management;
- as a baseline for future improvements; and
- to decide what those improvements should be, and the order in which they should be applied.

——— *Slide 14* ———

Roadmaps

That is exactly the kind of information we at Bath used to create our Research Data Management Roadmap.[13] We needed one of these in order to continue receiving funding from the EPSRC. We had to go through each of the EPSRC's Expectations and say

- how we measured up currently;
- what we would need to do to achieve compliance;
- how we would do it;
- when;
- and who would be responsible for getting it done.

——— *Slide 15* ———

Policies

This provided us with a mandate for subsequent activity. Other universities achieved this by first adopting a Research Data Management Policy. There are several ways of going about introducing such a policy.

- The University of Oxford decided to do theirs in two stages. They started by adopting a statement of commitment, then set up the necessary infrastructure, then adopted a policy that meant researchers had to use it.
- The University of Edinburgh decided to adopt an aspirational policy, that is, one with which they could not comply immediately. They would work towards full compliance over the coming years.
- The University of Hertfordshire already had a policy for administrative data, so it extended it to cover research data as well, accompanied by support materials for researchers

Policies and roadmaps may provide a mandate, but to release the money needed, one also needs to produce a business case.

——— *Slide 16* ———

12. CARDIO: `http://cardio.dcc.ac.uk/`
13. UoB Roadmap: `http://www.bath.ac.uk/rdso/University-of-Bath-Roadmap-for-EPSRC.pdf`

Business Cases

Here is the one for Bath. In it we contrasted the benefits of investing in Research Data Management against the real dangers of not investing. It was not an all-or-nothing proposal: we offered several different options, each with their own costs and benefits.

It was worth doing. The Vice Chancellor's Group approved the recommendations and created two new posts within the Library specifically to deal with Research Data Management issues: one with a more human focus and one with a more technical focus.

One of the things they will be doing is setting up our institutional data repository.

——— *Slide 17* ———

Institutions as data publishers

While there are a lot of similarities between running a research paper repository and a research data repository, there are some significant differences.

Data is rather harder to look after than a paper:

- It is less easy to spot if there are bits missing, and
- each discipline has different standards for how data should be documented.
- Furthermore, while publishers have the peer review system for guaranteeing the quality of papers, they rarely submit research data to the same level of quality control. So that task often falls back on data repositories.

Subject-based repositories like the ICSU World Data Centres add a great deal of value to data by cleaning it up, normalising it and combining it into data products. They have a solid claim to be data publishers and to hold the version of record. So what does that mean for universities setting up their own data repositories?

- They need to have conversations with subject-based data centres about sharing metadata, and about who hosts the data.
- If a researcher deposits data with a data centre, the university may need to be ready to take it back and re-home it should that centre shut down.

This is because setting up a data repository and publishing data and metadata brings with it a lot of commitments: not just regarding data quality, but to making the data persistently available and usable, and uniquely identifiable, so that researchers can cite it with confidence.

——— *Slide 18* ———

So what does that mean in practical terms? The university needs a data repository for storing and preserving the data assets it holds. There are standards such as the Data Seal of Approval which describe how these should be run. It also needs a data catalogue that makes the assets visible and easy to find; the catalogue can also hold metadata for assets that its researchers have deposited in external data centres.

To make data assets uniquely identifiable and citable, there are several solutions available. One could set up a Handle server, acquire a Name Assigning Authority Number for assigning ARKs, or get an account with a DataCite member for assigning DOIs. The University of Bristol recently became the first UK university to mint a DOI for one of its datasets.

Once a dataset has an identifier it can be cited, and those citations tracked. Thomson Reuters' Data Citation Index was launched in 2012, but one could also track impact through free services like ImpactStory or DataCite's statistics service.

We are also beginning to see new interactions between published data and traditional publications. Data journals provide a highly readable form of documentation for datasets,

and also a more traditional style of citation for those disciplines not ready for citing data held by repositories. There are also hybrid papers as exemplified by Elsevier's Article of the Future, where data is loaded from a repository and visualised as part of the article's online text.

The technology is available to do all this. The real barrier that needs to be addressed is the skill set of the data librarians running the service. In particular, how will data librarians handle different subject-specific requirements? Might we see data librarians specialising in a particular subject, and trading their expertise with other institutions? In the UK, data repositories have been dealing with this issue by working closely with researchers on meeting their metadata requirements. And, indeed, working with researchers on data issues is probably the most important part of the data librarian's job.

——— *Slide 19* ———

Research Office vs Library vs IT Services

When researchers have a problem with their data, where might they turn for advice? No doubt many will turn to colleagues, but as far as central support goes, I suspect where they go will depend on the problem. On the slide I have set out a likely mapping based on my experience. And I think that gives a very good idea of the kinds of areas in which data librarians will need to be expert.

So, for the next few slides I'll look at some of them in more detail.

——— *Slide 20* ———

Data Management Plans

One of the first things researchers will need help with is writing a data management plan (DMP). UK funders started asking for these in the mid-2000s, and in her 2007 report *Dealing with Data*, Dr Liz Lyon recommended that all bids for funding should include a DMP. This is because if researchers write and follow a good DMP,

- everyone knows what they ought to be doing,
- all necessary tasks are accomplished,
- the data is looked after properly while active,
- it can be preserved more easily, and
- it can be shared as openly as possible, in a timely fashion.

We find that researchers need a lot of help when they come to write a DMP. There are a couple of tools available that can simplify the task:

- DMP Tool specifically caters for applicants to the National Science Foundation and the National Institutes of Health in the US,[14] while
- the DCC has developed one called DMPonline, which caters for the requirements of a wide variety of funders.[15]

Both of them provide step-by-step guidance for filling out a DMP.

——— *Slide 21* ———

14. DMPTool: `https://dmp.cdlib.org/`
15. DMPonline: `http://dmponline.dcc.ac.uk/`

DMPonline v4

In version 4 of DMPonline, due out later this year, the guidance it offers has been packaged up into 40 different themes. Each question in each funder template has been classified according to these themes, so that the right guidance is shown wherever it is needed. CASRAI, the Consortia Advancing Standards in Research Administration Information, is looking at this taxonomy of themes as the possible basis for an international CERIF-compliant DMP profile.

The tool includes a template we have written for researchers who do not have to follow a particular funder template.

We also encourage universities provide their own templates and tailored guidance. At the moment we invite them to do this by adding to our guidance spreadsheet, and we load what they write into the system. In future they'll be able to add to the system directly, but the spreadsheet might also be helpful for librarians who want an overview of possible DMP guidance.

——— *Slide 22* ———

Institutional DMP guidance

Many libraries are providing DMP guidance outside these tools as well. On the slide are a DMP template and a guide to writing DMPs from the University of Bath. Some universities also provide examples of successful DMPs that researchers can use for inspiration. You see that a lot on American university library websites. In our training courses we sometimes like to use a spoof DMP to demonstrate what not to write.

——— *Slide 23* ———

Institutional RDM guidance

There's more to research data management than just the planning. There are all those issues I showed you before, such as storage allocations and metadata standards. Universities are starting to provide research data management portals as a resource for their researchers. Here are two examples from the Universities of Leicester and Bath. The Bath page, although it is hosted by the Research Office, is in fact maintained by the data team in the Library.

——— *Slide 24* ———

Giving librarians the skills they need

Where will data librarians learn the skills and knowledge they need to support research data management?

We are seeing research data management courses being taught by Library and Information Schools now. The US is leading the way here, but in Europe we're beginning to catch up.

——— *Slide 25* ———

For practising librarians moving into the data area, there are several good courses that have been developed.

- Mantra, the do-it-yourself research data management training kit for librarians, was developed by EDINA and the University of Edinburgh in association with the UK Data Archive, the DCC, and the Distributed Data Curation Center at the Purdue University Libraries. It consists of presentations, podcasts and assignments through which librarians can work.[16]

16. MANTRA: `http://datalib.edina.ac.uk/mantra/libtraining.html`

- RDMRose is an Open Educational Resource that was developed by the University of Sheffield iSchool and tested out on librarians from the Universities of Leeds, Sheffield and York.[17]
- Data Intelligence 4 Librarians is a four-module course developed by 3TU.Datacentrum and DANS in the Netherlands.[18]

——— *Slide 26* ———

At Bath we're currently developing a course of our own in collaboration with the University of Melbourne. It is called ImmersiveInformatics, and as the name suggests, its distinctive feature is that the participants use their new skills straight away.[19] They shadow a practising researcher and work with them to curate their real working data. They document their progress by keeping an electronic data diary.

——— *Slide 27* ———

In addition to training courses, there is a wealth of guidance material available online. The DCC produces both briefing papers and how-to guides on a range of practical research data management topics.[20] We also maintain a catalogue of useful tools and disciplinary metadata schemas. The Australian National Data Service also has a selection of guides on research data management at both the awareness and working level.[21]

——— *Slide 28* ———

The future data librarian?

So, to conclude, what might the data librarian of the future be doing?

- We might see data librarians from different universities collaborating to provide shared data infrastructure services. That might be one way of solving the problem of providing discipline-specific informatics support to a diverse range of researchers.
- We will almost certainly see data librarians involved in providing statistics that demonstrate their worth to the senior management of the university.
- It would be good to see information professionals fully integrated into research workflows, in order to raise and maintain standards of data management and provide support all the way from the initial bid to long-term preservation.
- Once we are reaping the benefits of data sharing and good data management in the wider research community, might we see the role of data librarians properly acknowledged, with co-authorship citations for datasets, or even the associated research papers?
- Data-related courses are becoming more mainstream in Library and Information Science courses, but with so much library work tied up in the digital realm, it is not such a great leap to imagine that data management might be seen as a core librarianship skill in future.

17. RDMRose: `http://www.sheffield.ac.uk/is/research/projects/rdmrose`
18. Data Intelligence 4 Librarians: `http://dataintelligence.3tu.nl/en/home/`
19. ImmersiveInformatics: `http://immersiveinformatics.org/`
20. DCC resources: `http://www.dcc.ac.uk/resources`
21. ANDS guidance: `http://www.ands.org.au/guides/`

- Our institutional data librarian in Bath was trained as a biochemist. I am confident we will see more specialists with scientific, technical, engineering and medical backgrounds being brought into the library profession to deal with specialist data issues.

Whatever the future holds, one thing is certain. The academic community needs its data librarians, and so the library community must respond by recruiting and training them.

——— *Slide 29* ———

Acknowledgements

Some slides were based on those by Liz Lyon (http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/presentations.html#munich-liber-2013-06), and Sarah Jones, Marieke Guy and Miggie Pickton (http://www.dcc.ac.uk/training/rdm-librarians).

The DCC is supported by Jisc.