Data librarianship – a gap in the market

What happens to ‘orphan datasets’ – the raw data collected by researchers to support a PhD thesis? The answer, is not necessarily very much.

Leaving aside ‘big’ e-science and the computing middleware to support it (the issues are different), there has been an international network dedicated to data sharing and data archiving in the social sciences since the 1960s.

Yet, in spite of talk of sharing and re-use of data, and of open access repositories for eprints (electronic copies of peer-reviewed journal articles deposited in the institutional repository), the truth is we hardly exploit as yet the capacity of digital technology for sharing. That is why Jisc set up Disc-UK DataShare, a project to develop ‘new models, workflows and tools for academic data sharing’.

The policy and technology environment is complex. There is new emphasis in the HE community on ‘stewardship of knowledge assets of all types’. There are new technologies for doing e-research. The research councils have policies and mandates for the recipients of research funding. There is also much talk of open access and open data. So why has progress on sharing data been so slow?

A primary reason is that although there always has been some sharing between a few collaborators, usually through informal networks, on the whole, academics are cautious about placing their data in the public domain.

And there are many other issues, to do with formats, standards, metadata, politics and policy, including how national data archives should complement institutional repositories, and whether you need one or two archives for each set, one for ongoing collaboration, one for the ‘finished’ outcome.

On top of all this is the vexed matter of copyright. If a social scientist re-works a publisher’s financial data for research purposes, part of the resulting data subset belongs to the publisher. So, even if the research councils mandate public access to, and re-use of, publicly financed research as a condition of funding, as Stuart says, translating policy into

Stuart Macdonald and Luis Martinez-Uribe have humorously described themselves as ‘accidental’ data librarians, but they have a serious message for the research community. They talked to Elspeth Hyams about data sharing and the increasingly important role of librarians in harvesting, curating and facilitating access to datasets.
practice is not simple. ‘There are licensing issues. It’s a whole can of worms.’

From this discussion it might seem that being a data librarian is primarily a policy, strategy and advocacy job with some technical work on preservation and metadata thrown in, and that providing guidance on how to deposit lies at the core of the role. In fact, quite simply, says Luis, it is about ‘supporting research’.

Nowadays, managing the life-cycle of research data provides a land of opportunity. The role of the data librarian is so fluid you can almost invent the job as you go along. ‘Watch out for more on emerging roles in a special supplement for Jisc in the October Update.’

Data librarians are not just there to save datasets from the scrapheap, though. They deal with the ‘selection, acquisition and management of a multi-disciplinary collection of electronic data resources’.

In the social sciences these include: micro-data (government surveys, population censuses, election studies), macro-data (country-level economic time-series, e.g. IMF, OECD, World Bank, UN data products), GIS (geospatial data such as the Ordnance Survey suite of data products, digitised boundary data, satellite and aerial imagery), and financial data (such as Datastream, Bureau Van Dijk, Amadeus, Bach).

In addition, data professionals support researchers in finding and using data resources via a whole range of activities such as:

- running training courses
- one-to-one reference interviews
- preparation and compilation of user guides
- cataloguing of resources
- subssetting/matching users’ data with recognised data sources
- troubleshooting data-related problems
- teaching
- interpreting codebook/questionnaire/documentation
- data management
- current awareness, via mailing lists, conferences, seminars and workshops
- institutional representatives for national data centres.

It isn’t a new job, then. It is due to political pressures and the ‘open data’ movements of the last few years that data librarians and other related players are finding themselves in the spotlight now.

**Huge volume of data**

As Stuart and Luis pointed out in a recent article, data is more than just statistics. Primary research data drives academic research across all disciplines.

Recent research carried out by the Australian Department of Education, Science and Training has indicated that the amount of data generated in the next five years will surpass the volume of data ever created, and in a recent IDC White Paper it was reported that, between 2006 and 2010, the information added annually to the digital universe will grow at a compound rate of 57% per year for data and 55% for voice, thus moving the total exabyte fold from 161 exabytes to 988 exabytes.

That implies a more prominent role within the research lifecycle for data management, and involves a range of functions.

**Researchers, librarians, technologists, publishers and policymakers will have to adapt their practices in order to deal with this new landscape’, they said.**

Meanwhile, secondary analysis of data itself leads to new data output which can feed back into the research lifecycle.

Apart from the complex policy issues that exist with datasets for work in progress and for archiving (which Harry Gibbs in her ‘State-of-the Art Review’ covers well), a lot of social sciences data is already collected by a number of agencies, including Edina, Mimas, and the UK Data Archive (UKDA). In the UK, the National Datacentre for Social Science Data is established by Edunserv, the agency which negotiates licences to access the data on behalf of the HE community.

National data centres play a similar role, either digitising much-requested non-digital research material, or building up a portfolio of products for a specific subject area.

It is part of the job of data professionals to advise researchers on the datasets available through these agencies, and through national statistical agencies, private companies, government departments, NGOs, etc. Many data providers offer academia special purchase/subscription rates.

Data professionals are in touch with both national providers such as the UKDA and the Economic and Social Data Service and international bodies like Iassist (the International Association for Social Science Information Service and Technology). Contact is useful for access to their networks of practitioners, for their expertise in evaluation of the resources and technological developments, and for their acquisition mechanisms.

Data libraries also sometimes purchase data on request for researchers.

**Training for the role**

Are librarians getting enough training to cope in this new role? As far as we know, no library school in Britain offers classes in data librarianship/curation and/or archiving.

The received wisdom within institutions is that, still that specific subject domain experts are the best qualified to deal with specialist datasets. But this topic is much discussed. Luis and Stuart and others at Disc-UK DataShare think there is a strong case for a librarian role in harvesting and curating datasets (though you do need at least a rudimentary knowledge of the dataset and any associated production and analysis tools).

UK data librarians come from a range of backgrounds including maths (Luis) and biochemistry (Stuart). Some, like them, have a qualification in information science, but others have stumbled into this line of work. Some have research skills, some library skills, some are statisticians and some are subject experts. Each of the UK’s main data libraries addresses different academic audiences and institutional biases, and local staff have developed different skill-sets accordingly.

Disc-UK, the UK’s small group of specialist data librarians, hopes to set up a programme with a data provider to train more ‘trainers’ and on-site role it is to support individual institutions’ own data providers. There are also opportunities with commercial data service providers, many of whom offer on-site and online training materials for their services. They offer training workshops to the academic community. Individual data librarians within institutions already conduct one-to-one training on the use of a particular interface, data product etc. There are also well-known learning and teaching resources (from the relevant ILT subject specialists) tailored for particular subject areas.

**Pressure from funding councils**

Although there is not yet agreement on how much data should be archived or provided by national services and how much looked after locally, there is in any case increasing pressure from the funding councils for progress. If researchers must now deposit their datasets as part of their research outcomes, it is probably only a matter of time before the UK’s library schools adapt to the new political climate and follow the Universities of California at Berkeley and Illinois at Urbana Champagne in including data management modules in LIS courses.

The one thing that is certain is that if more training were funded, the number of data libraries, the infrastructure to support dataset sharing, and the available dataset management expertise in the community, would grow. Canada, which had just six data libraries before it set up a Data Liberation Initiative with funding from Statistics Canada and the Canadian government, now has more than 70. There appears to be a huge gap in the UK market – and data librarians like Stuart and Luis think Disc-UK could help meet that latent demand.

Looking to the future? They would like to see information professionals who work in academic libraries enhance their support and organisational skills. They should explore relationships with the research and computing communities. It is there that a mixed economy of institutional and national data management capabilities is emerging.

Library schools could help, by adding data curation to their curricula. This would nurture a workforce capable of supporting research in a dynamic and technology-driven environment where an inter-disciplinary community uses the new web technologies to share expertise and knowledge. All hands to the pump for the data deluge?

**References and notes**

1. www.disc-uk.org/datashare.html. This project is funded by the Jisc Repositories and Preservation Programme. Disc-UK is the Data Information Specialist Committee (UK).


6. The University of Glasgow’s Humanities Advanced Technology and Information Institute does run a MSc in Information Management Preservation (www.hatlit.gla.ac.uk/impl/index.html).

Stuart Macdonald is from the Disc-UK DataShare & Edina National Data Centre and Luis Martinez-Uribe works at the Oxford e-Research Centre.