DICE project

Literature and resource review

DRAFT 3

Malcolm Raggett, February 2012

Contents

Summary

Although repositories for research *publications* are now well established in universities, the same cannot be said of repositories for research *data*. There are discipline-specific repositories at the UK national level but these set a high curatorial bar. There are very few institutional-level research data repositories, and this immaturity is reflected in the awareness and skills of many researchers in data management. The direction of current developments suggests that the DICE project should position the development of its training materials in the context of data management planning in order to future-proof the material but still cater for the immediate requirements of the taught Information Literacy programme and other needs.

Background

A key factor in the success of the LSE digital library[1] will be its ability to attract content from depositors, especially researchers in the LSE community. The DICE (digital communication enhancement) project[2] has been funded as part of the JISC Digital Preservation & Curation Programme[3] to allow us to analyse our potential depositors' understanding of and concerns about the preservation of their digital research data, and to develop appropriate materials to assist them in preserving their data.

Although the project is specifically about training for the preservation of research data, this is only one aspect of data management. There are several models used that cover the

[1] http://digital.library.lse.ac.uk/

[2] http://lsedice.wordpress.com/

[3] http://www.jisc.ac.uk/whatwedo/topics/digitalpreservation.aspx (accessed February 2012).

management of data, especially: SCONUL's Seven Pillars model[4], UKDA's 6-stage structure[5], and DCC's Curation Lifecycle model[6]. All include preservation of data within their models, however these models are principally aimed at professions working in libraries and archives. Since this project is about researchers and their needs, alternative models may also be valid, for example work at the University of Michigan has suggested that considering research projects as exercises in personal information management (PIM) may have more success[7].

It is LSE's intention to incorporate the material produced by this project into its Information Literacy Workshop (usually known locally by its course code MY592), though as Secker and Coonan point out, this is not a clearly defined term[8], and MY592 does have elements of personal information management in its content and presentation (though formally positioned within the SCONUL Seven Pillars model). The project outputs will therefore need to be appropriate for immediate consumption but adaptable to future researcher-specific requirements.

Other target audiences for the project's outputs are:

- the information professionals in the library, in order to to raise awareness and embed the vocabulary of digital preservation in everyday use;
- potential depositors of digital material to the LSE's Special Collections, particularly in deciding what to deposit, in what format and on what media;
- Records Management, where at least some of the same awareness-raising and thought processes will need to occur.

What is research data?

It is common to think of the outputs of research, usually publications, as being the valuable part of research. However, much of the time and cost of research is spent collecting raw data, both secondary and primary, and generating further data via analysis, on which to base the publications, but this primary, secondary and derived data is not the only possible definition of research data. In fact, what is perceived as "research data" can be hard to pin down: the American National Science Foundation (NSF), which has mandated a data management plan in all bids for funding[9], has declined to define what it means by data[10], leaving this to the

[4] SCONUL. The Sevel Pillars of Information Literacy. http://www.sconul.ac.uk/groups/information_literacy/seven_pillars.html (accessed February 2012).

[5] http://www.data-archive.ac.uk/create-manage/life-cycle (accessed February 2012).

[6] http://www.dcc.ac.uk/resources/curation-lifecycle-model (accessed February 2012).

[7] Fear, Kathleen. "You made it, you take care of it" Data Management as Personal Information Management. In International Journal of Data Curation. Issue 2, Volume 6, 2011.

[8] Secker, Jane and Coonan, Emma. 2011. A New Curriculum for Information Literacy. http://eprints.lse.ac.uk/37681/

[9] http://www.nsf.gov/bfa/dias/policy/dmp.jsp (accessed February 2012).

judgement of the bidder. Major UK funding bodies similarly require data plans and MANTRA (research data management training) explains that whether something counts as research data depends on what they (the data) are, and when and how they are used. The point is also made that "[data] can also be created by researchers for one purpose and used by another set of researchers at a later date for a completely different research agenda."[11]

The OECD defines data as: "...factual records (numerical scores, textual records, images and sounds) used as primary sources for scientific research, and that are commonly accepted in the scientific community as necessary to validate research findings. A research data set constitutes a systematic, partial representation of the subject being investigated."[12] It also explicitly excludes "...laboratory notebooks, preliminary analyses, and drafts of scientific papers, plans for future research, peer reviews, or personal communications with colleagues or physical objects... "[13]. Clearly this definition comes from a science-based approach to data and is likely to be inadequate for much cultural research in the arts and humanities. It could also prove limiting within the social sciences.

Not only is "research data" hard to define but recent work at the University of Michigan also shows that the concept of "data" tends to come from the sciences and social sciences. Many researchers in the arts and humanities do not consider themselves to work with data. "Moving the conversation away from 'data management' and toward 'working with information en route to creating a publication' might allow for a more inclusive discussion of data management."[14] This is certainly a point to bear in mind when designing training materials.

However, for the purposes of this paper the term "research data" will continue to be used but in the broad context of all disciplines; "data" are not just numerical or tabulated information but *all* information on which research outputs are based, whether primary input or first order results[15], and irrespective of the medium or format.

---

[10] http://www.lib.umich.edu/research-data-management-and-publishing-support/nsf-data-management-plans#data_included (accessed February 2012).

[11] http://datalib.edina.ac.uk/mantra/researchdataexplained.html (accessed February 2012).

[12] OECD Principles and Guidelines for Access to Research Data from Public Funding. 2007. http://www.oecd.org/dataoecd/9/61/38500813.pdf : p.13

[13] OECD Principles and Guidelines for Access to Research Data from Public Funding. 2007. http://www.oecd.org/dataoecd/9/61/38500813.pdf : p.14

[14] Fear, Kathleen. "You made it, you take care of it" Data Management as Personal Information Management. In International Journal of Data Curation. Issue 2, Volume 6, 2011: p.74

[15] Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information. Final Report of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access, February 2010, p.11. http://www.jisc.ac.uk/publications/reports/2010/blueribbontaskforcefinalreport.aspx

What is preservation?

Most major UK funding bodies require public access to and long-term curation of research outputs[16], and there is a growing expectation that data will also be made publicly available for at least 3 years to more than 10 years[17]. In some cases the preservation will be perpetual, though preservation without access will usually be perceived as of little value.

From a researcher's point-of-view, digital preservation normally means getting their information into a form that can be submitted to an archiving service. In the past this would have meant organising paper records in a systematic way. The same principle is true of digital records but with additional rules to cope with the nature of digital information and media. For example, the ESRC makes it an obligation on all funded projects to submit their research data to a "data provider", usually the Economic and Social Data Service (ESDS), within 3 months of the end of the project[18], however only [waiting for data from ESDS 15/2/12] are successfully accessioned.

The case for preservation of research data

The preservation of digital research data has been embedded at a national level for a number of years, with the Library of Congress (USA), the British Library and National Archives (UK) and the Australian National Data Service, to name only 4 organisations, all developing reasonably mature services. Several of the UK's HE funding councils have maintained data archives for decades but it has taken longer for a realisation of the growing importance of preserving digital research data to penetrate to the researchers. Nevertheless, it is the funding bodies that are now largely driving the need for data preservation along the lines articulated by the Organisation for Economic Cooperation and Development (OECD) in 2007[19]. These were adopted initially for the sciences[20] but it is easy to extrapolate the arguments to the social sciences, humanities and arts. This paper will not reiterate all of the arguments, but one example is worth quoting because of its general applicability:

> Research data can be valuable for many years after they are generated. Data that led to initial insights can sometimes be used to generate new findings in the same or entirely different research fields. Existing data can be reanalyzed or combined with new

[16] http://www.dcc.ac.uk/resources/policy-and-legal/overview-funders-data-policies (accessed February 2012).

[17] ibid

[18] ESRC Research Data Policy. 2010. p.6. http://www.esrc.ac.uk/_images/Research_Data_Policy_2010_tcm8-4595.pdf

[19] OECD Principles and Guidelines for Access to Research Data from Public Funding. 2007. http://www.oecd.org/dataoecd/9/61/38500813.pdf

[20] National Academy of Sciences, Committee on Ensuring the Utility and Integrity of Research Data in a Digital Age, 2009, Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age, (National Academies Press 2009).

data to verify published results or arrive at new conclusions.[21]

There is, of course, also a strong cost justification: recreating lost data is expensive and time-consuming, and sometimes impossible.

The current position

In the technologically developed, English-speaking countries of the UK, USA, Canada, Australia and New Zealand there is an acknowledgement of the need to preserve data at a national level since all of these countries have national libraries and archives that operate in the digital domain. Judging by the amount and quality of information about digital preservation on the Internet, the USA and UK are leading in the development of research data preservation, closely followed by Australia, with New Zealand and Canada some way behind.

Interestingly though, this national lead has not cascaded evenly to the academic level of universities or research funding bodies. In the UK the (then) Social Science Research Council established the SSRC Data Bank, now renamed the UK Data Archive, in 1967. Even at this early stage in digital developments it was recognised that data were being lost or duplicated due to a lack of coordination[22]. Thankfully the (now) Economic and Social Research Council (ESRC) can be applauded for giving the service continuous support, which is vital to its work. In 1996 the Arts and Humanities Data Service (AHDS) was established by the (now) AHRC but funding for it was withdrawn in 2008 and the various parts distributed to alternative hosting and funding arrangements (there is no evidence of data loss but at the time it did feel like a breach of trust between researchers and funders). The AHDS has therefore had a somewhat more chequered history, and this highlights the need for a long-term strategic commitment to any attempt to preserve data.

It is no surprise that these UK-wide services curate their data to a high standard: a researcher cannot simply store all their stuff in them in the hope that someone might find it useful. It is also likely that some researchers are unaware of these services or don't think to use them as a repository for non-Research Council-funded research. So is there a need for a level of preservation service below the national level but still curated professionally? If so, this is where an individual university's repository would prove useful. It is also where most work needs to be done both at a senior management level and at the individual researcher level: the former to ensure an on-going and long-term commitment, and the latter to ensure good practice in data management. There is the generic "Lots Of Copies Keeps Stuff Safe" argument to justify a local repository but this alone seems inadequate for the provision of a potentially costly and long-term commitment. Undertaking the work to answer this question specifically for LSE is

[21] National Academy of Sciences, Committee on Ensuring the Utility and Integrity of Research Data in a Digital Age, 2009, Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age. Executive summary. p.7. Available online at www.nap.edu/html/12615/12615_EXS.pdf

[22] http://www.data-archive.ac.uk/media/54761/ukda-40thanniversary.pdf

outside the scope of the DICE project however recent work at the University of Oxford did seek to justify the development of local research data repositories, and gave four requirements which are worth reproducing here, although it is the second and third requirements that are most interesting in answering the above question. The requirements were for the provision of:

1. Advice on practical issues related to managing data across their life cycle. This help would range from assistance in producing a data management/sharing plan; advice on best formats for data creation and options for storing and sharing data securely; to guidance on publishing and preserving these research data.
2. A secure and user-friendly solution that allows storage of large volume [sic] of data and sharing of these in a controlled fashion way [sic] allowing fine grain access control mechanisms.
3. A sustainable infrastructure that allows publication and long-term preservation of research data for those disciplines not currently served by domain specific services such as the UK Data Archive, NERC Data Centres, European Bioinformatics Institute and others.
4. Funding that could help address some of the departmental challenges to manage the research data that are being produced.[23]

About half of the correspondents in the Oxford survey were from the humanities or social sciences so although these requirements can only be treated with caution in their applicability to LSE, there is at least some similarity to give grounds for encouraging an LSE-based research data archive. It may well be worth undertaking a similar study at LSE, perhaps as part of the recently-formed Research Data Management Working Group's investigations.

For its part, JISC certainly foresees a mixture of repositories: "...There is likely to be a hybrid of institutional, national and international research data centres as well as subject and disciplinary data centres."[24] Work done by the UK Research Data Service (UKRDS) project examined the feasibility of a framework of standards and policies so that an interlinked grid of national and institutional repositories could be established. Its final report in May 2010 proposed a pathfinder project based in 3 universities together with a national support organisation[25]. Although HEFCE allocated some funding for 2010/2011[26], this appears to have funded Cloud development and no outputs seem to have been produced or further recommendations made regarding research data services.

Although data repository services were developed and still exist as stand-alone entities, today no one except a technical specialist would deal with the preservation of research data in

[23] Martinez-Uribe, Luis. 2008. Findings of the Scoping Study Interviews and the Research Data Management Workshop. p.11. http://www.ict.ox.ac.uk/odit/projects/digitalrepository/docs/ScopingStudyInterviews-Workshop%20Findings.pdf

[24] Heery, Rachel. 2010. Digital Repositories Roadmap Review: towards a vision for research and learning in 2013. p.30. www.jisc.ac.uk/media/documents/themes/infoenvironment/reproadmapreviewfinal.doc

[25] UK Research Data Service. 2010. Proposal and Business Plan for the Initial Pathfinder Development Phase. http://www.ukrds.ac.uk/resources/download/id/16 (accessed February 2012)

[26] JISC Executive. 2010. Annex A, JISC(10)25.

isolation: it is now viewed in the context of research data management, and the need for a data management plan, which would include preservation, is becoming increasingly popular with funders. The NSF in America and most of the funding councils in the UK have made a data management plan an integral part of the funding bid. This gives the DICE project a certain difficulty as it is scoped to deal only with "preservation", not "planning". However, since data management planning is important, at least as a framework, for the DICE project, it will be dealt with further in the next section of this review.

The requirement to preserve research *outputs* is quite well established within UK universities, with many operating their own institutional repositories (OpenDOAR suggests that around 90% of UK HEIs now operate one or more open access repositories[27]). However the use of these repositories does vary: some contain full-text, others only abstracts, but most are a mixture of abstract, published and pre-published content. So it seems that there is now strong support at the institutional level for provision of research *publications* that are free at the point of access. Unfortunately the policies regulating these repositories are very variable, for example, of 206 UK institutional repositories, only 23.3% have a preservation policy[28]. It is as if the impetus to set up these repositories has past and their development has reached a plateau while the repositories are populated with content by researchers.

The need to preserve research *data* is less mature at the institutional level: only 7.7% of the UK repositories contain datasets[29] (although the latter category is only a narrow definition of "research data", no other recent information has been found to demonstrate commitment to preserving research data at the institutional level). The LSE repository, eprints.lse.ac.uk, contains over 22,000 items with an accession rate of just over 2,000 per year[30]. A few of these items contain associated data but the vast majority are publications.

Data management planning

As previously stated, research data preservation is now universally regarded in the context of an overall data management plan for a research project. The variable nature of research across all disciplines means that it has not been possible to devise a single template or formulaic approach to developing a data management plan. The Digital Curation Centre (DCC) has produced a 3-tier model for a data management plan[31]: minimal, core and full:

- a minimal data management plan includes only the questions required by the funder or institution at the application stage;

[27] http://www.opendoar.org/index.html (accessed February 2012).

[28] ibid

[29] ibid

[30] http://eprints.lse.ac.uk/view/year/ (accessed February 2012).

[31] https://dmponline.dcc.ac.uk/system/attachments/8/original/DCC_Checklist_DMP_v3_md_sj.pdf?1300724157

- a core data management plan includes all necessary information for in-project data management;
- a full data management plan adds long-term preservation and data management to the core plan.

The DCC has done an excellent job turning this into a set of questions that in turn has become an on-line form for (initially) bid authors to develop their plans[32]. Although there are advantages to this model (for example, why go to the trouble of developing a full plan at the application stage?), it does mean that preservation considerations are left until a full plan is developed, which may be quite late in a project if it happens at all. It may also mean that mistakes may have been made early in the project that will be time-consuming or impossible to correct later.

As a model, the full data management plan as structured by the DCC offers a useful framework within which to position the DICE project's work on data preservation. Some of the questions are prerequisites for the preservation-related questions. Extracting the prerequisite and preservation-related questions gives the following list:

1. give a short description of the data being generated or reused in the research;
2. are there any funding-body, ethical or legal issues (including IPR) to consider for preservation of the data?
3. which file formats will you use and why?
4. are the datasets you will be capturing and/or creating self-explanatory or understandable in isolation? If not:
    a. what documentation and/or metadata will you create, how and in what form?
    b. and why have you chosen particular standards and approaches for documentation and metadata?
5. how will you backup data during the lifetime of the project? What is the backup schedule and who is responsible for backing-up?
6. will or should the data be kept beyond the lifetime of the project? If so:
    a. will it be all data or just a selection (what is the justification for the selection)?
    b. how long should the data be kept and what is the long-term strategy (i.e. beyond the life of the project) for maintaining, curating and archiving the data?
    c. what data centre/repository/archive will be the long-term place of deposit?
7. If the dataset contains sensitive data how will this be managed over the long-term?
8. will transformations be necessary to prepare data for preservation and/or sharing? If so:
    a. describe these.
9. will the documentation or metadata need to be transformed to meet the needs of the data centre/repository/archive? If so:
    a. describe these.
10. how will the issue of persistent citation be addressed?
11. who will have responsibility over time for decisions about the data once the original personnel have gone?

[32] https://dmponline.dcc.ac.uk/

12. in the event of the long-term place of deposit closing, what is the formal process for transferring responsibility for the data/documentation/metadata?

Focusing on this set of questions would keep the DICE project within its scope but position it within a recognised framework so that training relating to other aspects of the data management plan can be organised in a complementary way in the future. It is worth noting that JISC is currently funding a set of projects on research data management[33] that are due to finish in July 2013 and which may supply more detail and maturity to the planning framework.

Survey of training available

A number of organisations, for example, the DCC[34] and ULCC[35], provide training and training materials that include aspects of digital preservation. These courses and materials are aimed more at librarians and archivists than researchers. Although information professionals are one target audience for DICE, the DCC/ULCC materials will not be appropriate for training LSE researchers. Similarly, National Archives and Libraries provide some training but this is not aimed at the academic researcher.

Several of the national repository services offer training to researchers though at least some of these link back to the UK Data Archive as the training provider. The ESDS certainly perceives the lack of training available: "There is a deficit in the training of early career researchers in the areas of handling and management of the research data they produce."[36] The UK Data Archive provides some excellent notes for guidance available on its Web site and makes its training material for researchers available for download as a single zip file[37]. The material is a well-written distillation of their extensive experience in data preservation aimed firmly at the researcher who is reasonably IT literate. Even so, the materials would be best used as part of a teacher-guided course. The materials cover the whole of the data life cycle but could be cherry-picked and adapted to LSE's requirements.

MANTRA[38], a JISC-funded project to develop online training for research data management, ran from September 2010 until August 2011. The course is online[39] but it appears that the course is

[33] Research Data Management Infrastructure Projects. http://www.jisc.ac.uk/whatwedo/programmes/di_researchmanagement/managingresearchdata/infrastructure.aspx (accessed February 2012).

[34] http://www.dcc.ac.uk/resources/curation-lifecycle-model (accessed February 2012).

[35] http://www.ulcc.ac.uk/content/digital-preservation-training (accessed February 2012).

[36] Looking After and Managing your Research Data. 2012. http://www.esds.ac.uk/news/eventdetail.asp?id=3116

[37] http://www.data-archive.ac.uk/media/335419/trainingresources.zip

[38] http://www.ed.ac.uk/schools-departments/information-services/about/organisation/edl/data-library-projects/mantra

not complete: the module on “Preservation, Sharing and Licensing” - the most interesting from DICE’s point-of-view - is shown as “in development”. Although the JISC funding has now finished, there is still some development taking place and this module is expected to be ready “in the near future”[40]. In the meantime there is little for DICE here except to note the existence of other modules that may be of some use in future developments at LSE.

The Digital Preservation Outreach and Education (DPOE) section of the Library of Congress conducted a user survey in 2010, the results of which are intended to inform its development of training material including data preservation. Unfortunately (for DICE’s purposes) 90% of the respondents were from libraries, archives or museums; very few were from researchers or research groups, so it is unlikely that the DPOE’s materials will be of much use to LSE. In any case, the DPOE survey does not appear to have resulted in any online resources yet: the resources shown on their Web page[41] are from external providers and none of these are directed towards the university researcher.

Jorum[42] contains several researcher-targeted training packages about data management directed at specific academic areas (archaeology and health care) or they are peripheral to data preservation, e.g. “selecting data to keep” and “working with digital media files”.

It would seem then, that in the areas of social sciences and humanities, there is a dearth of training material directed at researchers to raise their awareness and skill level in data preservation. This is good, since to find otherwise would have shortened the DICE project considerably.

Recommendations

For the DICE project

There are several models that could be used to position the project outcomes within a recognised framework. Although the Seven Pillars model is used for the MY592 Information Literacy course at LSE, for entirely practical reasons (i.e. a structure in the form of a set of questions already exists) DICE should work within the DCC’s framework for data management planning. This will give the lowest risk of the project’s outputs becoming redundant in the future as other data management planning training is developed. Nevertheless, the outputs should still fulfill the immediate requirements of the MY592 course. If the outputs can also be made useful for Special Collections depositors and for Records Management purposes, then this

[39] http://datalib.edina.ac.uk/mantra/ (accessed February 2012).

[40] Donnelly, Anne. Feb. 2012. Per. Comm.

[41] http://digitalpreservation.gov/education/courses/

[42] http://www.jorum.ac.uk/

opportunity should also be taken, though this should not compromise the value to the principle audience of researchers.

The use of the term “data” can be anathema to some researchers, particularly in arts and humanities. In order to engage these people it will be necessary to use alternative language such as “information”, which is a more comfortable term for them.

It is not necessary or desirable to define “data/information” in order to work in the context of preserving it. It is more important that the researcher should be able to decide which of their research data/information is important in the long-term, and the actions needed to preserve them. It is therefore recommended that we adopt the position of the NSF: avoid definitions and expect researchers to develop the skills within a framework.

For LSE

Although there is an implicit acceptance of the need to archive research data centrally within LSE, the requirements have not yet been drawn up and costed. This should be done, perhaps using the Oxford model (see p.6), but should certainly involve researcher input to the specification.

Training researchers in data preservation will raise the questions and expectations that can be predicted as: “What file formats should I use?” And “where should I store my data for long-term preservation?” LSE will need to develop the infrastructure, policies and guidelines to address these issues.

There is a need to bid for funding for further developments to complete a training package for research data management. This can build on the work of DICE but care should be taken not to duplicate the recent MANTRA project or the current JISC-funded projects on research data management infrastructure.