

Response to consultation of RIN draft guidance on stewardship of digital data

DISC-UK (Data Information Specialists Committee - United Kingdom) is a forum for data professionals working in UKHE who specialise in supporting their institution's staff and students in the use of numeric and geo-spatial data (<http://www.disc-uk.org>). DISC-UK DataShare is a JISC repository enhancement project (March 2007 - March 2009) that aims to explore new pathways to assist academics wishing to share their data over the Internet. With four institutions taking part – Edinburgh, LSE, Oxford and Southampton – a range of exemplars will emerge from the establishment of institutional data repositories and related services.

We welcome RIN's attention to the issue of stewardship of digital research data and recognise the April 2007 draft paper as an important contribution to the much-needed discussion about how to move forward in this ill-defined area. The document itself is clear, well-informed, and covers the full range of important issues that need to be addressed. We hope the following specific comments contribute to the dialogue and may possibly inform the next version of the paper.

Background and Context

Why do we need a policy framework? (p 3)

It's useful for such high-level policies to be articulated, but more could be said here about the need to understand factors of resistance to this 'culture change' of sharing or preserving data. For example, in this section it states "The essential goals we are seeking to achieve are thus to facilitate the advancement of research and innovation, to enhance the efficiency and effectiveness of research, and to maximise the value of public and private investment in research." None of those goals would necessarily resonate with individual researchers (except altruistically). See for example the JISC StORe project survey for evidence about this, at:

<http://jiscstore.jot.com/WikiHome/SurveyPhase/SurveyFinalReport.doc> or

<http://www.emeraldinsight.com/Insight/viewContentItem.do?contentType=Article&contentId=1593463>

So if the approach is to deliver the policy 'from the top', perhaps more thought is needed about the motivation and drivers for researchers to participate. (Unless RIN views that as taken care of by mandates, and this paper as urging research councils or institutions to implement mandates. But evidence suggests even where there are mandates that is not enough in itself to change behaviour.)

Sensitivity to the requirements of different kinds of research data (p 4)

Useful categorisations here, and in line with paper by the National Science Board on Long-Lived Data Collections (footnoted on p 5). One class of data that could be mentioned explicitly (perhaps at "derived data, resulting from processing or combining 'raw' or other data") is proprietary sources, e.g. commercial databases. This is a demonstrated barrier for sharing derived data for researchers using Ordnance Survey data (see the EDINA GRADE project deliverables, or <http://technology.guardian.co.uk/weekly/story/0,,2049772,00.html>) and may also be for those using expensive proprietary databases, for example in the area of finance.

Principle 1: roles and responsibilities (p 7)

The phrase "creators and users of research data" is used. Note that early DCC work looked at user requirements of "creators, curators and users" which may be useful. (See <http://www.dcc.ac.uk/resource/interviews/>)

"In most cases, the main role of funders will be to set the broad policy framework, rather than detailed procedures, where the default responsibility rests with research institutions..."

This statement is reinforced throughout the document, where "(funders; research institutions)" as a pair are attributed for at least 20 actions. We are not sure this is an accurate level of placement of responsibility in many cases, particularly for establishing fairly detailed policies and procedures. Many research institutions, such as our own, are very large, and can only set a "broad policy framework" for the entire institution (similar to research funders). It would also likely be fairly impossible to set the same policies and procedures for the broad spectrum of disciplines and data types. Rather, the carrying out of policies and procedures for data curation is done at either the department level, or perhaps the research group. We are also surprised that none of the responsibilities go down to the level of the principal investigator, which is at the moment often the only locus of control that currently exists for data stewardship.

Principle 2: standards and quality assurance (p 8)

It may be useful to note that quality is often enhanced by making data available to users; as they discover errors or inconsistencies and report them back.

Principle 3: access, usage and credit (p 9)

Publishers, Data Services and Users (p 10)

33. In some cases, where authentication and authorisation is in place, it is not necessary to establish audit trails and to know who has had access, etc. Perhaps some of these questions err on the side of caution or perfection, where a more lightweight (and less resource-intensive) approach would suffice. It may be useful to study the infrastructure for authentication and authorisation services such as Athens and Shibboleth where the identity of the individual may be hidden to the service and only known by the authenticating institution (devolved). Of course it is possible to add a layer of registration on top, so the service can establish such use trails, but this should only be done where considered necessary (e.g. sensitive data).

Moreover, considering the policy objective mentioned at the start of the paper, that

"Ideas and knowledge derived from publicly-funded research should be made available and accessible for public use, interrogation, and scrutiny, as widely, rapidly and effectively as practicable,"

it is surprising that this section considers only "managed access" and not open access, especially considering the recent rise of an "open data" movement alongside the open access (to scholarly literature) movement. We understand that many of the stakeholders interviewed for this paper may have held strong views about this, but there are certainly options for publishing data on unrestricted websites as well as through managed services. Any requirement that a user must "confirm their acceptance of terms and conditions and access" will in effect mean that the resource in question will not be downloadable from a Google search, and is therefore a barrier to access. Arguably, this will reduce both discoverability and re-use, and therefore cost-effectiveness of funded research. We suggest that the RIN should further explore the model being established through Creative Commons and Science Commons, where resources are made available openly, with particular terms and conditions about re-use and attribution attached, but a record of use, registration, or audit trail is not required. For an open access repository, it is more realistic (and efficient) to provide tools for rights-holders to assign creative common licenses to their works than to be able to report to the rights-holder exactly who has downloaded the resources and for what purposes.

A suggested citation included with the resource can also help to encourage responsibility by users, reinforcing norms of academic behaviour such as proper attribution. This is suggested in the next section of the paper (Credit, Citation and Evaluation), but the "clearly understood arrangements for allocating credit to data creators" again seems to take a strong-arm stance rather than depending upon and encouraging existing norms of academic behaviour. Our experience working with researchers is that where data are not cited or not cited properly it is not due to lack of will to credit the source, but lack of understanding of how to cite them or lack of ease.)

Principle 4: benefits and cost effectiveness (p 11)

(37) "As data management becomes increasingly integral to the research process itself, all those with responsibilities for data – researchers, research institutions, library and data services, and research funders - need to ensure that their policies and practices operate cost-effectively."

Interestingly, libraries are only mentioned here and in the introduction. They are not given any explicit responsibilities for specific activities in the paper. This may be another reason to further break down "research institution" into component parts, and to explore what the responsibilities of libraries should or could be, though there is likely to be wide variation in the willingness of libraries to take on data curation responsibilities, as well as resistance by the research community if this is seen as interference with their work. Records managers may also play a role on behalf of the institution. In terms of training, we would point to the successful 'train the trainers' programme carried out in Canada as part of the Data Liberation Initiative as a successful effort between university libraries and Statistics Canada to increase librarian's skills for data support.

Principle 5: preservation and sustainability (p 12)

This section provides a broad sketch of important preservation issues including appraisal, security, format migration, provenance, version control, and sustainability.

Regarding the question:

"What arrangements are there to ensure that data of potentially high value over the very long term are transferred to a recognised specialist repository, that they are complete, and free from corruption?"

A paper by Ann Green and Myron P. Gutmann from the US may be helpful here in elucidating roles and responsibilities, and when in the data lifecycle important "hand-offs" may take place between researchers, institutional repositories, and domain-specific data archives/services:

'Building partnerships among social science researchers, institution-based repositories and domain specific data archives'. See

<http://deepblue.lib.umich.edu/handle/2027.42/41214> for an abstract and open access version of a paper published in OCLC Systems & Services (2007); the publisher's URL is

<http://www.emeraldinsight.com/Insight/viewContainer.do?containerType=Issue&containerId=24717>

Chris Rusbridge, Director of the Digital Curation Centre has also made the point that digital preservation is like a relay race, with different parties taking responsibility for a limited period and then 'passing the baton'.

Nevertheless, it may not always be clear which repository is the sustainable one, in terms of certainty of future funding: the institutional repository or a national repository, e.g. funded by a research council. (Witness the recent announcement by the AHRC about withdrawal of future AHDS funding.) Perhaps the LOCKSS strategy of "lots of copies keeps stuff safe" is worth bearing in mind in terms of redundancy between IRs and national data centres/archives. Institutional repositories have been criticised for not focusing enough on digital preservation, but it is still early days and they have time to improve before content degrades or becomes obsolete.

Robin Rice Edinburgh University Data Library

Stuart Macdonald Edinburgh University Data Library

Luis Martinez London School of Economics

Tanvi Desai London School of Economics

Jane Roberts University of Oxford

Harry Gibbs University of Southampton