1st LEARN Workshop, Embedding Research Data as part of the research cycle

London, 29th January 2016

Breakout Session Group 1

Chair: **Paul Ayris** (UCL)

Rapporteurs: **Myriam Fellous-Sigrist, assisted by Martin Moyle and Matt Mahon (UCL)**

Breakout group initial participants list:

Alea Lopez de San Román	League of European Research Universities
Allison McCaig	University of Exeter
Andrea Meyer Ludowisky	Senate House Library London
Andrew Gray	University of London
Ash Barnes	Imperial
Caroline Williams	University of Nottingham
Claire Sewell	University of Cambridge
Verena Weigert	JISC
David Carr	Wellcome Trust
Federica Fina	University of St Andrews
Helen McEvoy	University of Salford
Kerry Miller	University of Edinburgh
Lesley Thompson	Elsevier
Mandy Thomas	De Montfort University
Matt Mahon	University College London
Myriam Fellous-Sigrist	University College London
Paul Ayris	University College London
Sarah Stewart	University of London
Thomas Shaw	University of East London

This breakout group tackled two main questions: the barriers and the drivers to move towards a research environment where research data are increasingly shared and re-used. During the discussion three areas of needs were identified: finding a clearer leadership and authority to mandate and guide data management; being discipline-sensitive enough to avoid a "one size fits all" model of data management and sharing; and, finally, developing skills of data managers, researchers and publications reviewers.

The Group was chaired by Dr Paul Ayris. It was mostly composed of British participants but was diverse professionally-speaking (researcher, data scientist, data manager, digital curator, journalist, research funder officer). Five of the data managers present (out of the 19 participants) had taken up their post very recently, i.e. between a few weeks and a few months ago.

What barriers are there to moving towards a research environment with more data sharing and re-use?

A gap in leadership

Despite mandatory funders' policies on research data, no real change has been observed by participants over the last years. These policies are known by researchers but they are not prescriptive and the disparity of expectations makes compliance look like a box-ticking exercise. For instance, not all funders require or even recommend Data Management Plans and researchers only write them if they are compulsory in grant applications, not with the idea of using and updating them during the project. Research funders could also develop more discipline-specific repositories to get the money invested in projects back; for example, ESRC and NERC have already adopted that strategy.

Research communities are much more likely to be regarded as the main source of standards and codes than researchers' institutions. These communities expand beyond institutions and national boundaries. Learned societies might also help advocate for better data management practices.

Data management can also be advocated by researchers themselves ("local champions") via emails to colleagues, meetings in their research departments or by becoming the local expert in data management. Even if participants report a success in using such solution (research colleagues do come and ask questions of the expert), all agreed that this strategy is not scalable.

Regarding Research Data Policies, it is crucial that they are based on the research lifecycle so as to reflect the researchers' (especially the Principal Investigators') activities and decision-making processes.

The need for a data publication model

Infrastructure and standards to publish data are still lacking: data could be treated separately from the "narrative" of the publication; "data papers" should have their own publication date, ORCID number, peer-review model and should be interoperable with the linked "narrative" paper or book if there is one.

An evaluation model of the quality of data published has yet to be invented, along with a coherent system to select and structure data to disseminate, and training materials for the future data reviewers. Publishers will have to decide whether publications containing data already published in a separate data repository would be acceptable.

Any data publication model will have to be secure and transparent enough to reassure researchers fearing that their data might be stolen by potential reviewers.

A lack of confidence & skills

Academic support staff struggle with the variety of enquiries sent by researchers regarding data management; such enquiries vary greatly in terms of disciplines but also types of questions (technical, legal, ethical, policies, costs, etc.). Case studies on how data are managed and shared would be welcome. There is a need for an authority body to concentrate expertise in data management training and advocacy; this would enable the community to avoid duplication of work and to provide a source of up-to-date information to re-use for support staff. The group suggested that such role could be taken up by funders or institutions such as the Digital Curation Centre or JISC.

Support staff sometimes have to face the "wilful ignorance" of some researchers who hope that funders' requirements are not set in stone. Researchers (at all career stages) should also be trained regularly in data management and new funders' requirements; this should start with first-year undergraduates (e.g. data awareness training at Imperial College London). Doctoral students in particular should benefit from data management and requirements training; they could be encouraged and trained to submit their thesis along with their data and create bi-directional links between the two.

Metadata standards (both general and subject-specific), anonymisation and basic IT skills are some of the key skills to enable data sharing; they will be essential to train data managers and researchers.

Several participants have found that quick, topic-specific and hands-on workshops within departments are an efficient way to disseminate data management skills; such workshops can be based on the model of Data Carpentry workshops or the LERU Doctoral Summer Schools.

Developing and testing costs models

A movement of funds from grant-holder to library or IT services is required. Crude charging models (£/GB or £/dataset) could be used but they can lead to bad practice, as researchers try to control costs. An established, central funding stream would be preferable and would require funders to clarify eligible costs, then Principal Investigators would identify costs and finally, on award, monies to cover data management costs would be transferred to a central fund.

Alternatively (or as a complement), discipline-specific research funders or RCUK could provide funds for national research data services. For instance, JISC is currently piloting a shared RDM-service project. Such projects are still in their early phases and the group was sceptical as to their long-term utility. These questions will need to be raised again in a couple of years.

What incentives are there to encourage researchers to share their data?

The interest and limits of an Open Access-like driver

Looking at the wide adoption of Open Access (OA) practices should help us understand what mandates and what benefits drove different disciplines to adopt them. Most researchers now see the benefits of OA to publications or comply with the OA policies because of the REF requirements.

However the REF cannot be used to require data sharing and Open Data as these models are not appropriate or relevant for all disciplines. Any appropriate driver would have to be discipline-sensitive enough to avoid the "one size fits all" model, allow for the diversity of ethos and cultures and rely on recognised bodies of peers rather than institutional policies to set the limits of what "open" means and when it is appropriate.

Data metrics

Use of metrics should be encouraged on the basis of increased citation and visibility of data produced by researchers, not as another ranking tool. Digital Object Identifiers (DOIs) are increasingly accepted as useful by researchers.

The group discussed whether actual re-use of data should be put forward as an argument for sharing more data: it is not clear whether most researchers actually want to re-use data and given the low readership of academic publications, it seems too early to rely on figures of re-use.

Data metrics cannot be produced in all disciplines and limits to data visibility also have to be identified here. Open Science is not an appropriate model for all fields, let alone for interdisciplinary projects where different research cultures have to be accommodated. Some disciplines are historically used to sharing as a result of the cost of data production (e.g. crystallography, genetics, astrophysics); others also have that tradition, such as chemistry, psychology (after anonymisation of data) or acoustics.

Showing the benefits of data sharing

Data Access Statements could be encouraged more widely, not only by a few research funders (such as EPSRC or NERC); publishers, funders and institutions could expect and train researchers to use them in their publications. Several other ideas of incentives were discussed:

- giving the example of Google Scholar where datasets are indexed like articles,

- explaining how datasets are as valuable and respectable as publications,

- showing data during paper presentations at conferences,

- having prizes for PhD candidates who share their data (under embargo or not), like the Open Chemistry Prize already awarded by Imperial College London,

- data visualisations of how often research datasets are accessed and downloaded.

Good relationships between research data advocacy teams (be it the library, Research Office or IT department) and research departments are critical to ensure a long-term dialogue on these issues.

Training authors to better cite their own data and others’ data used in publications is also crucial. Clear recommendations should be given to them and a consistent model of citation (which should include DOIs) needs to be more widely shared.