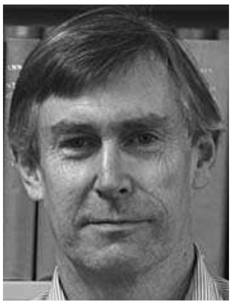
Journal of the European Association for Health Information and Libraries
(November 2008) vol.4(4) pp.4-6

Open Data: the elephant in the room?(*)



Peter Morgan
Cambridge University Medical Library
Addenbrooke's Hospital, Cambridge, UK
Contact: pbm2@cam.ac.uk

Abstract
The principles of the Open Access movement incorporate the need for open access to data, or Open Data. Research funding bodies are mandating the release and re-use of data, but small-scale research projects may lack the resources to implement Open Data management procedures. Libraries and institutional repositories, which have focused efforts on managing text resources rather than data, can assist in addressing this problem by collaborating with the research community.

Key words: information storage and retrieval; libraries, medical; open access to information; research.

Open Access and Open Data

The Open Access (OA) movement is now a well-established feature of the information landscape. Since its formative days interest has focused principally on research papers and the ways in which they might best be made freely available to researchers, but the original vision also included research data among its key elements. For example, the Berlin Declaration of October 2003 stated that *open access contributions include original scientific research results, raw data and metadata...* (1). Despite these early exhortations, it is only comparatively recently that the issue of research data has begun to receive serious and widespread attention from the Open Access community.

To give the issue a clearer identity its advocates have developed and promoted the concept of Open Data (OD), defined as *a philosophy and practice requiring that certain data are freely available to everyone, without restrictions from copyright, patents or other mechanisms of control* (2). While OD may be associated in some way with published OA papers, this is by no means essential, and the data in question might equally be associated with a non-OA paper published solely in a traditional subscription-based journal.

Establishing a policy

Scientific research is data-driven. Research projects generate vast quantities of data, which in turn provide the raw material on which further research is based. Increasingly, as the major research-funding bodies have embraced the principles of OA, they have come to recognise that if they concentrate their policies solely on research publications they will then be neglecting the potential value of the underlying data and failing to ensure the best possible return on their investment.

International bodies like the Organisation for Economic Co-operation and Development (OECD), the Commission of the European Communities (CEC) and

**The elephant in the room ... is an English idiom for an obvious truth that is being ignored or goes unaddressed. It is based on the idea that an elephant in a room would be impossible to overlook; thus, people in the room who pretend the elephant is not there might be concerning themselves with relatively small and even irrelevant matters, compared to the looming big one.* Wikipedia [updated 2008 Oct 15; cited 2008 Oct 16]. Elephant in the room; [about 3 screens]. Available from: http://en.wikipedia.org/wiki/Elephant_in_the_room the European Research Council (ERC), together with national public sector research funders like the Medical Research Council (UK) and research charities such as the Wellcome Trust, have thus responded by publishing statements and guidelines (3-7) that explicitly address the question of data, encouraging - and in some cases formally mandating - their researchers to make appropriate management arrangements for the data they produce, with the purpose of ensuring that the results are both accessible and re-usable. The Wellcome Trust's policy statement is a good example: *the Trust considers that the benefits gained from research data will be maximised when they are made widely available to the research community as soon as feasible, so that they can be verified, built upon and used to advance knowledge.*

This is not to say that all data should be openly accessible. There are circumstances in which access to research data must be restricted. On occasion a temporary embargo may be justified, where the researcher wishes to deny access pending publication of scientific papers based on the data in question. In the longer term there may be overriding legally binding reasons, such as pharmaceutical research that is commercially funded and where the data are subject to the funder's contractual ownership; or there may be issues of confidentiality and data protection, often a central concern with medical research projects. The importance of this latter consideration has been highlighted by recent events where the supposed anonymity of personal data has been shown to be compromised and led to the withdrawal of open access arrangements (8).

The question of data ownership, and thus the right to determine whether data should be openly accessible and on what terms, is not a trivial one. It can be complicated by uncertainties over the respective rights of individual researchers and their employing institutions. It can further be complicated in those cases where the data, while described as "open", are nonetheless dependent in some way (such as file formats or analytical programmes) on proprietary standards or software. The most important feature of Open Data, as with other aspects of OA, is not simply the ability to gain free access to a resource but also, crucially, the ability to re-use that resource with appropriate acknowledgement. Before other researchers can safely re-use data they need to be reassured that they have permission to do so and at the same time be made immediately aware of any restrictions that might need to be observed. Statements of ownership in metadata do not of themselves indicate what permissions or restrictions apply, and may usefully be supplemented by licences that convey additional information, such as those available from the Creative Commons organisation (9). In addition to these generic licences, other licences designed for more specific needs are now being developed, such as the Science Commons' Health Commons project (10).

The long tail of science

Research projects funded by bodies such as those listed above are likely to be large, generating a correspondingly substantial quantity of scientific data. In such circumstances, increasingly driven by funder mandates, it is usually the case that the research project will incorporate its own appropriately-funded data management procedures and technology, supervised by subject experts. It is equally true, however, that much research is conducted on a far more modest scale, still generating important data but in much smaller quantities and without the same level of resource being made available for data management. The discrepancy between these two types of research has been described as *big science versus little science*.

Big science functions within a well-funded infrastructure of major facilities, shared on a national or international basis and possibly including a purpose-built subject-based data repository. *Little science* - sometimes characterised as "the long tail of science" - embodies the realisation that much scientific research is conducted by a large number of small groups capable of producing significant results but lacking the benefit of a co-ordinated infrastructure and working in relative isolation. These groups suffer from the risk that their data outputs will not be readily accessible, partly because they have insufficient resources and skills to implement good data management practices and also because they are less likely to be governed by the mandatory requirements of data accessibility that the funding bodies impose on major grant recipients. As a result the datasets they create are vulnerable: they have no well-organised, sustainable home; their ownership - and therefore the right to determine how they can be released, and under what conditions - is uncertain; and little-science researchers lack the political, financial, and organisational muscle necessary to secure support within their parent institutions.

At the same time, few institutional repositories have yet begun to accept responsibility for offering a home to datasets. Their content is predominantly text-based (research papers and theses), and the task of persuading the local academic community to support Open Access self-archiving has tended to consume most of the manpower and time available for advocacy campaigns and strategic planning initiatives.

A role for librarians

Evidence is now emerging that this situation is changing. The problems associated with long-tail research data management have been attracting increasing attention recently, not least because they appear to offer scope for librarians, including repository managers, to co-operate with researchers in developing systems that will allow their data to be made both accessible and re-usable. Some university libraries have risen to the challenge by creating new posts designated as data librarians, demonstrating a commitment to the idea of sharing responsibility for data management with their local researcher community (11), while others are exploring ways in which their institutional repositories can acquire, manage, curate and expose research data.

For such co-operation to work, both parties - researchers and librarians - need to recognise that they have complementary skills and assets. Researchers bring domain expertise, familiarity with the workflows and protocols of scientific research, and an appreciation of the value of the data they produce; while librarians offer skills in organising knowledge and managing information technology. As libraries find their traditional roles increasingly under threat, so they need to identify and develop those areas of activity where they can provide services and advice that are currently unavailable. The management of research data, especially when derived from small-scale projects, and in particular the promotion of Open Data as a strategic objective, represent one such area.

Conclusion

The principles of Open Data are becoming better understood and are beginning to acquire a higher profile than before. However, discussions within the library and information services community on the concept and future of Open Access continue to perpetuate a widely held assumption that OA is essentially concerned with peer-reviewed research publications. To neglect the interests of the Open Data elephant in the Open Access room is to miss an opportunity for librarians to play a more active part in supporting the research process.

References

1. Conference on Open Access to Knowledge in the Sciences and Humanities, Berlin, 20-22 October 2003 [updated 2006 Dec 20; cited 2008 Oct 14]. Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities; [about 4 screens]. Available from: http://oa.mpg.de/openaccess-berlin/berlindeclaration.html
2. Wikipedia [updated 2008 Aug 23; cited 2008 Oct 14]. Open data; [about 9 screens]. Available from: http://en.wikipedia.org/wiki/Open_data
3. Organisation for Economic Co-operation and Development [cited 2008 Oct 14]. OECD Principles and Guidelines for Access to Research Data from Public Funding. 2007. 24 p. Available from: http://www.oecd.org/dataoecd/9/61/38500813.pdf
4. Commission of the European Communities [updated 2007 Feb 14; cited 2008 Oct 14]. Communication from the Commission to the European Parliament, the Council and the European Economic and Social Committee on Scientific Information in the Digital Age: Access, Dissemination and Preservation. SEC(2007)181. 10 p. Available from: http://ec.europa.eu/research/science-society/document_library/pdf_06/communication-022007_en.pdf
5. European Research Council [updated 2007 Dec 17; cited 2008 Oct 14]. Scientific Council. Guidelines for OpenAccess.2p.Availablefrom:http://erc.europa.eu/pdf/ScC_Guidelines_Open_Access_revised_Dec07_FINAL.pdf
6. Medical Research Council [updated 2007 Jun; cited 2008 Oct 14]. Policy & Guidance: Data Access [about 4 screens]. Available from: http://www.mrc.ac.uk/PolicyGuidance/EthicsAndGovernance/DataAccess/index.htm
7. Wellcome Trust [updated 2008 Feb; cited 2008 Oct 14]. Policy on Data Management and Sharing [about 2 screens]. Available from: http://www.wellcome.ac.uk/About-us/Policy/Policy-and-position-statements/WTD002766.htm
8. Los Angeles Times [updated 2008 Aug 29; cited 2008 Oct 14]. DNA databases blocked from the public [about 3 screens]. Available from: http://www.latimes.com/news/nationworld/nation/la-me-dna29-2008aug29,0,4364552.story
9. Creative Commons [home page on the Internet] [updated 2008 Oct 14; cited 2008 Oct 14]; [about 7 screens]. Available from: http://creativecommons.org/
10. Science Commons [cited 2008 Oct 14]. The Health Commons [about 4 screens]. Available from: http://sciencecommons.org/projects/healthcommons/
11. Data Information Specialists Committee - UK [home page on the Internet] [updated 2008 Oct 9; cited 2008 Oct 14]; [about 1 screen]. http://www.disc-uk.org/index.html