# To Share or not to Share: Publication and Quality Assurance of Research Data Outputs

Report commissioned by the Research Information Network (RIN)

## Executive summary

June 2008


www.rin.ac.uk

In association with:

JISC

NATURAL ENVIRONMENT RESEARCH COUNCIL

# Executive summary

The digital age has presented the research community with new opportunities. Research findings in digital form can be easily moved around, duplicated, handed to others, worked on with new tools, merged with other data, divided up in new ways, stored in vast volumes and manipulated by supercomputers if their nature so demands. There is now widespread recognition that data are a valuable long-term resource and that sharing them and making them publicly-available is essential if their potential value is to be realised.

There are two essential reasons for making research data publicly-available: first, to make them part of the scholarly record that can be validated and tested; second, so that they can be re-used by others in new research.

This report presents the findings from a study of whether or not researchers do in fact make their research data available to others, and the issues they encounter when doing so. The study is set in a context where the amount of digital data being created and gathered by researchers is increasing rapidly; and there is a growing recognition by researchers, their employers and their funders of the potential value in making new data available for sharing, and in curating them for re-use in the long term.

The last two years have seen the development of policies from funders both in the UK and internationally, seeking to optimise the value and the use of data produced during the course of research that they fund. Both policy and researchers' practice continue to evolve, and so this study should be seen as a picture of current activity that will change further in the future

We gathered information on researchers' attitudes and data-related practices in six discrete research areas – astronomy, chemical crystallography, classics, climate science, genomics, and social and public health sciences – and two interdisciplinary areas – systems biology and the UK's rural economy and land use programme. The primary methodology used was interviews with over 100 researchers, data managers and data experts. The report is in two parts, both available on the RIN website at [www.rin.ac.uk/data-publication](www.rin.ac.uk/data-publication) , along with this executive summary. The main report presents a synthesis of the overall findings, including recommendations for consideration by the relevant bodies. An Annex, presented as a separate document, reports in detail on the findings from each of the eight research areas.

## *Key findings*

### Data creation and care

1. Researchers create and collect many different kinds and categories of data during the course of their research, and datasets are generated for different purposes and through different processes. In determining which datasets should be made publicly-available, there are important distinctions to be made between those generated through
    a. scientific experiments;
    b. models or simulations; and
    c. observations of specific phenomena at a specific time or location.
2. There are significant variations – as well as commonalities - in researchers' attitudes, behaviours and needs, in the available infrastructure, and in the nature and effect of policy initiatives, in different disciplines and subject areas. We provide towards the end of this document a summary of the position in each of the eight areas that have been the focus of this study.
3. Data may undergo various stages of transformation in the course of the research process, and may be made available to other researchers at any of those stages. The convention in many fields is that derived or reduced data – as distinct from raw data - are what is made available to other researchers. Providing access to raw data is relatively rare, though it may be the most effective

means of ensuring that the research is reproducible. But there is discussion in some fields about the lack of access to raw data.

4. Many datasets of potential value to other researchers and users – particularly those arising from small-scale projects – are not managed effectively or made readily-accessible and re-usable. Many are stored by researchers themselves in a more or less haphazard manner on DVD or hard disk with little chance of effective retrieval; and those on websites are vulnerable in the long term especially if the website depends on project funding.

5. Many research funders are putting policies in place to ensure that datasets judged to be potentially useful to others are curated in ways that allow discovery, access and re-use. But there is not a perfect match between those policies and the norms and practices of researchers in a number of research disciplines.

6. Researchers in disciplines and subject areas which have large centralised data centres benefit from expertise and resources in data curation that cannot be provided consistently at local level. But such centres cannot accept all the data that is produced; and the recent closure of the Arts and Humanities Data Service shows that even apparently well-established centres cannot provide watertight guarantees for the long-term provision of accessible and usable data.

7. Distributed, local data storage may provide a more agile approach, with the advantage of closeness to researchers; but a key disadvantage is the current shortage of expertise and resources at local level.

8. The quality of metadata provided for research datasets is very variable, from the standardised, enhanced metadata of the large, professionally-curated data centres and databanks through semi-standardised schemes in smaller data collections to researcher's own *ad hoc* labelling.

9. Value may be added to data in a number of ways: by annotation, addition of additional datasets, and by curation, aggregation and enhancement. Researchers may do these things themselves to a degree. Data centres may carry out all these tasks as well as checking, verifying, and cleaning datasets and providing software tools for data access and manipulation.

## Motivations and constraints

10. Some researchers are motivated to publish their data by factors such as altruism, encouragement from peers, or hope of opening up opportunities for collaboration. But the lack of explicit career rewards, and in particular the perceived failure of the Research Assessment Exercise (RAE) explicitly to recognise and reward the creating and sharing of datasets – as distinct from the publication of papers - are major disincentives.

11. Many researchers wish to retain exclusive use of the data they have created until they have extracted all the publication value they can. When combined with the perceived lack of career rewards for data creation and sharing, this constitutes a major constraint on the publishing of data. Other disincentives include lack of time and resources; lack of experience and expertise in data management and in matters such as the provision of good metadata; legal and ethical constraints; lack of an appropriate archive service; and fear of exploitation or inappropriate use of the data.

## Discovery, access and usability

12. Some publishers are taking steps to underpin the scholarly record by creating persistent links from articles to relevant datasets; and this signposting is viewed positively by researchers.

13. Relatively few researchers have the expertise, resources and inclination to perform themselves all the tasks necessary to make their data not only available, but readily accessible and usable by others

14. Data centres invest heavily in ensuring that the datasets they hold are readily usable; but usability is an issue often overlooked by researchers who publish data themselves. Datasets on journal websites are commonly in PDF format which is unsuitable for meaningful re-use.

15. Other obstacles to locating and gaining access to datasets produced by researchers and other organisations include inadequate metadata, refusal to release the data; the need for licences (which may restrict how the data may be used or disseminated) and/or for the payment of fees; or the need to respect personal and other sensitivities.

16. Effective use of raw scientific data in particular may require access to sophisticated specialist tools and technologies, and high level programming skills.

### Quality assurance

17. Most researchers believe that data creators are best-placed to judge the quality of their own datasets, and they generally take other researchers' outputs on trust in terms of data quality and integrity.

18. There is no consistent approach to the peer review of either the content of datasets, or the technical aspects that facilitate usability.

19. Data centres apply rigorous procedures to ensure that the datasets they hold meet quality standards in relation to the structure and format of the data themselves, and of the associated metadata. But many researchers lack the skills lack the skills to meet those standards without substantial help from specialists.

## *Conclusions and recommendations*

### Data creation and care

1. In developing their policies, research funders and institutions need to take full account of the different kinds and categories of data that researchers create and collect in the course of their research, and of the significant variations in researchers' attitudes, behaviours and needs in different disciplines, sub-disciplines and subject areas; and to make clear the categories of data that they wish to see preserved and shared with others in each case.

2. Research funders and institutions should co-operate in seeking to ensure that long-term and sustainable arrangements are in place to preserve and make accessible the data that they deem to be of long-term value, and that such arrangements are not put at risk by short-term funding pressures.

### Motivations and Constraints

3. Research funders and institutions should seek more actively to facilitate and encourage data publishing and re-use by

    a. promoting more actively through the use of case studies the benefits and the value to researchers of data publishing

    b. providing visible top level support, and offering career-related rewards, to researchers who publish high-quality data

    c. providing expert support to enable researchers to produce sound data management plans, and closely reviewing the quality of those plans when they assess grant applications

    d. making clear to applicants for grants and to reviewers that including a budget to cover data management – including the provision of a dedicated data manager where appropriate - will not adversely affect a grant application

    e. providing better information about and access to sources of expert advice on how most effectively to publish and to re-use data.

f. developing strategies to address the current skills gaps in data management

g. promoting and providing better information about the mechanisms available to data creators to control access to and use of their data (e.g. embargoes, restricted access, licence conditions)

h. promoting improved access to research data through better discovery tools and metadata standards

i. identifying and documenting by subject area the barriers to effective re-use of data, and promoting guidance on good practice

j. promoting the "freeze and build" approach to dynamic datasets, where original data may be amended, added to, or replaced by newer data at a later date.

4. Learned societies should work with researchers, funders and other stakeholders to develop and promote standard methods for citing datasets,

## Discovery, Access and Usability

5. Publishers should wherever possible require their authors to provide links to the datasets upon which their articles are based, or the datasets themselves, for archiving on the journal's website. Datasets made available on the journal's website should wherever possible be in formats other than pdf, in order to facilitate re-use.

6. Researchers and publishers should seek to ensure that wherever possible, datasets cited in published papers are available free of charge, even if access to the paper itself depends on the payment of a subscription or other fee.

7. Funders, researchers and publishers should seek to clarify the current confusion with regard to publishers' policies with regard to allowing access for text-mining tools to their journal contents.

8. Researchers, funders, institutions, publishers and other stakeholders should monitor the development and take-up by researchers of Web 2.0 applications, and their implications for data publishing, sharing, and preservation.

## Quality Assurance

9. Funders should work with interested researchers, data centres and other stakeholders to consider further what approaches to the formal assessment of datasets – in terms of their scholarly and technical qualities – are most appropriate, acceptable to researchers, and effective across the disciplinary spectrum.