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

Report commissioned by the Research Information Network (RIN)

Main report

June 2008

www.rin.ac.uk

In association with:

JISC

NATURAL ENVIRONMENT RESEARCH COUNCIL

Contents

A separate Annex, describing the detailed findings for the eight research areas, is also available on the RIN website at www.rin.ac.uk/data-publication .

Foreword

The last four years have seen a growing interest from governments and other agencies across the world in measures to encourage more effective dissemination and sharing of research findings and outputs of all kinds. One sign of that interest has been the number of reports published on a range of issues relating to the preservation of and the provision of access to the research data that is now being produced in increasing quantities in digital form. Reports from agencies in many countries, in the English-speaking world in particular, have made recommendations to funders, researchers and others as to why and how they should make the data that researchers gather and create accessible to other researchers and users. And research funders have begun to develop policies in this area.

What we lack, however, is a clear picture of how researchers are responding to these challenges: whether they are in fact making their data available and accessible to others, and the issues that they are encountering when and if they do so. In 2007, therefore, the Research Information Network (RIN), in association with the Joint Information Systems Committee (JISC) and the Natural Environment Research Council (NERC), commissioned a study of researchers' attitudes and their practices in relation to the data they create and gather in the course of their research.

We commissioned Key Perspectives Ltd to undertake the study; and we are most grateful to Alma Swan and Sheridan Brown for all the work they have done in gathering and analysing the evidence and in preparing successive drafts of their findings and conclusions. We believe that this report presents a valuable view of the current state of play in researchers' approaches to the range of issues that arise in relation to managing, sharing and publishing research data. On the basis of this evidence, we shall pursue our discussions with researchers, funders, and institutions as to how the process of data publishing and quality assurance might be enhanced for the future.

Michael Jubb
Research Information Network

Acknowledgments

Over one hundred people were interviewed during the course of this project. All of them gave their time willingly and graciously. In many cases interviews lasted an hour or even two, testing the generosity of spirit of these individuals rather severely.

They helped us in the interests of research and laying the groundwork for the best practices of the future. We promised anonymity but they know who they are. We thank them for all they did to aid us in exploring and reporting on the current situation with respect to research data in the United Kingdom.

Alma Swan and Sheridan Brown
Key Perspectives Ltd

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 . This 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.

1. Overview

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 are ways to ensure that their potential value is realised. The technology that supports these new opportunities continues to evolve rapidly, though researchers' attitudes to data creation and dissemination are not keeping pace in all disciplines.

In this context, the study reported here was primarily designed to investigate three key areas:

- the nature and range of arrangements for making research data as widely available as possible (referred to as "data publishing");
- the role that data outputs currently play alongside or as an alternative to conventional publications in the research communication process; and
- current practice for ensuring the quality of such data.

This report is based primarily on the results of more than 100 detailed interviews with researchers across eight subject areas. A more detailed description of the goals and methodology is presented in Section 6.

There is now potential for researchers and their funders to reap payoffs that were beyond the imagination of the print-on-paper age. The value of making data available for discovery, access and re-use is not yet quantified and this may be the focus of future work. It is clear, however, that there are new opportunities for science and scholarship: where even finding out about other people's results was once difficult, now it is simple; where accessing their findings took time and effort, now it can be almost instantaneous; where directly incorporating those data into one's own research was virtually impossible, now it needs a just little persistence and the right tools. The future is full of promise, but extracting the full potential of data will depend upon further progress in data management policies and practice – a melding of researchers' desire to share the fruits of their scholarly endeavours; effective work by researchers, data managers and curators to ensure data can be found and re-used by others; and funders' developing appropriate guidelines and policies to maximise the return on their investment. In seeking to progress in all these ways, however, it must be stressed from the outset that making research data available does not necessarily mean that they are accessible, and that making them accessible does not necessarily mean that they are readily usable.

Effective exploitation of data depends upon proper data management and curation. The Research Information Network recently published a set of principles and guidelines regarding research data[1]. The five main principles are:

1. The roles and responsibilities of researchers, research institutions and funders should be defined as clearly as possible, and they should collaboratively establish a framework of codes of practice to ensure that creators and users of research data are aware of and fulfil their responsibilities in accordance with these principles.
2. Digital research data should be created and collected in accordance with applicable international standards, and the processes for selecting those to be made available to others should include proper quality assurance.
3. Digital research data should be easy to find, and access should be provided in an environment which maximises ease of use; provides credit for and protects the rights of those who have gathered or created data; and protects the rights of those who have legitimate interests in how data are made accessible and used.

[1] Stewardship of Digital Research Data : Principles and guidelines (2008) Research Information Network. http://www.rin.ac.uk/data-principles

4. The models and mechanisms for managing and providing access to digital research data must be both efficient and cost-effective in the use of public and other funds.

5. Digital research data of long term value arising from current and future research should be preserved and remain accessible for current and future generations.

There are two main ways of storing and curating data – using large, centralised national or international data centres; or using a distributed array of local data stores (based on or in research institutions, researchers' own resources, or formal publication outlets such as journals). There are advantages and disadvantages to both routes and practice continues to evolve in each of them.

Centralised data centres provide expertise in data curation and archiving that cannot be provided consistently at local level, together with significant storage capacity. But they are selective in what they will accept for curation and storage, since they lack the capacity to take responsibility for everything that is produced in their disciplines or subject areas. Hence the curation of datasets that are rejected by the centres is left to researchers' inclinations and abilities.

Distributed, local data storage may be a more 'agile' approach and has the advantage of being 'close to the laboratory bench'; but a key disadvantage is the current shortage of expertise and resources at a local level. Relatively few universities, for example, have experience and expertise available in all that is involved in data curation and preservation. The role of journals presents interesting issues, since there are two ways for them to make data available. The first is to publish the dataset on the journal's website (or insist that the author deposits it in a recognised public databank); the second, a newer development, is when the traditional journal article format is eschewed in favour of publishing datasets instead. Thus a journal can contain just a series of datasets, providing a formal way of citing them and ensuring that they are preserved at least in the medium term. Examples are Acta Crystallographica E from the International Union of Crystallography and, in project phase, OJIMS (the Overlay Journal Infrastructure for Meteorological Sciences)[2].

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 cultural norms in some research disciplines and funder requirements. Some disciplines are well ahead of funding bodies in that they have had a culture of sharing data for a long time and have developed the infrastructures and methods for doing this. In other disciplines, data sharing is not commonplace and therefore funder policies may imply significant modifications to researchers' attitudes and behaviour. In the United Kingdom, five of the seven research councils have data sharing policies in place.

- The Arts and Humanities Research Council's (AHRC) policy came into effect from April 2008[3].
- The Biotechnology and Biological Sciences Research Council (BBSRC) takes a devolved approach and its policy came into effect in April 2007
- The Economic and Social Research Council (ESRC) requires data to be offered to its national data centres (UK Data Archive and the Economic and Social Data Service)[4].
- The Engineering and Physical Sciences Research Council (EPSRC) has no policy on data as yet
- The Medical Research Council (MRC) has no data centres but adopted a data sharing policy with effect from April 2006[5].
- The Natural Environment Research Council (NERC), which has a detailed data policy handbook and guidelines for grant-holders, has seven designated data centres where grant-holders can deposit their data[6].

---

[2] http://www.jisc.ac.uk/whatwedo/programmes/programme_rep_pres/repositories_sue/ojims.aspx

[3] But AHRC funding for the Arts and Humanities Data Service ceased on 31 March 2008

[4] http://www.esrcsocietytoday.ac.uk/ESRCInfoCentre/Support/access/

[5] http://www.mrc.ac.uk/PolicyGuidance/EthicsAndGovernance/DataSharing/PolicyonDataSharingandPreservation/index.htm

- The Science and Technology Facilities Council (STFC), formed in 2007 by a merger of the CCLRC (Council for the Central Laboratory of the Research Councils) and PPARC (Particle Physics and Astronomy Research Council), has yet to develop its formal data sharing policy, though its facilities have well-developed individual policies.

In other countries, too, data sharing polices are being developed by research funders. To document and clarify these, the SHERPA Juliet[7] directory service now covers the data archiving policies of a wide range of funders, providing a look-up service for those who wish to familiarise themselves with them.

The roles and responsibilities of the different parties involved in data publishing – researchers, research institutions and research funders – were addressed in a recent report for the JISC[8] covering strategy and policy, technical practice, legal issues, sustainability and training and skills with respect to research data. Various recommendations were made in that report and they are supported by findings from this present study.

Here, although we address the roles of research funders and report on the effect of their policies, we focus primarily on researchers. Researcher behaviour is at the centre of this report, an approach that differentiates this study from others that have taken a more top-down approach. The methodology employed was to go out into the UK research community and talk to people about how they work with data, how they produce data, how they manage their data, what constraints or regulations they work under, and what problems they encounter. In particular, the project sponsors wished to discover what motivates researchers to publish their data and, for those who choose not to, what factors inhibit them. Allied to this, we aimed to investigate issues of quality with respect to the scholarly content and the usability of published.

Eight research disciplines or areas were selected for study. Six of these are discrete disciplines or subject areas: astronomy, chemical crystallography, classics, genomics, systems biology, social and public health science. Two are interdisciplinary areas – the cross-Research Council programme on rural economy and land use and systems biology – selected because they exemplify new ways of doing research and because each represents a large and well-funded area of research in which considerable amounts of data are produced.

One of the strongest messages to be drawn from this study is the lack of uniformity across different research disciplines in terms of behaviour, policies or needs. Any solutions to the problems we identify, therefore, will need to be tailored to the requirements and practices of each individual research discipline. Interdisciplinary research needs especially careful consideration in this light.

There are, however, some commonalities that can be identified and this first part of the report deals with the issues in a broad way, highlighting those commonalities as well as the contrasts as part of the overall story. Its main purpose is to provide an overview of researchers' attitudes to data creation and publishing in the UK, to reflect the problems, and to point to some possible solutions. A summary table appears at the end of each main topic, presenting the main points to aid assimilation. In the second part of this report, presented in a separate document, we provide the detailed reports of how data are gathered, manipulated, managed and shared in each of the eight different disciplines and subject areas that we studied.

---

[6] http://www.nerc.ac.uk/research/sites/data/policy.asp

[7] http://www.sherpa.ac.uk/juliet/

[8] Lyon, EJ (2007) Dealing with data: Roles, rights, responsibilities and relationships. http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/reports/dealing_with_data_report-final.doc

2. Data creation and care

2.1 Forms and varieties of data

Researchers produce an array of forms and varieties of data, depending on the field in which they work. Often, however, key players in the scholarly communications field use the term "data" in a generic, imprecise fashion which can lead to confusion. As noted in the RIN's recent guidance on *Stewardship of Digital Research Data*[9] research data come in many varied forms. They are generated for different purposes and through different processes:

- scientific experiments, which may in principle be reproduced, although it may in practice prove difficult, or not cost-effective, to do so;
- models or simulations, where it may be more important to preserve the model and associated metadata than the computational data arising from the model; or
- observations – from the astronomical to the zoological – of specific phenomena at a specific time or location, where the data will usually constitute a unique and irreplaceable record.

Similarly, they can be generated or collected together for different reasons:

- for the benefit of those engaged in a specific project, where some of the data may have little value beyond the life of that project;
- for the benefit of a wider group within a discipline, or across disciplines, to provide reference information; or
- for the benefit of a very broad community of researchers and users who need to use canonical or reference data relating, for example, to gene sequences, chemical structures, or literary texts.

Finally, data may be produced at different stages in the research process, with variations in the status attached to them:

- raw data created or gathered in the course of experiments or observations; or
- derived data, resulting from processing or combining "raw" or other data.

In all fields of research, data may undergo various stages of transformation from the initial raw format – that collected from research instruments, by observation, by survey and so forth – to a final highly-processed form. The processing involves different procedures in different fields, with summarising and analysis taking various forms, including reduction (for example, image processing in astronomy), annotation (for example. assigning a biological function to a particular gene sequence), or curation (for example, formatting machine data into a community-developed standard such as the crystallographic community's CIF format). Data may be made available to other researchers at any and all of these stages or there may be conventions within a research field about the formats of data that are to be shared.

Researchers produce data in the normal course of their work, typically as part of a process, particularly in the natural and life sciences, towards publication in the form of journal articles. These datasets may be highly specific and limited in their potential re-use value. There is a subset of the research community, however, whose job it is to build and maintain datasets which themselves are the bases for many other researchers' work. Important longitudinal cohort studies and national survey datasets have been developed in the fields of social and health sciences, for example. This is an important distinction: whereas the datasets that are produced in the normal course of research may not be published for a variety of reasons, reference datasets that are produced with the intention that they should be used as a basis for further research are normally professionally curated with long term viability, usability and quality in mind.

---

[9] Stewardship of digital research data: a framework of principles and guidelines, Research Information Network, January 2008 http://www.rin.ac.uk/data-principles

With regard to data produced in the normal course of research, the convention is often that derived or reduced data are the type that are available to people other than the data creators. But there is considerable discussion in some communities about the lack of access to raw data. There may be practical reasons as to why data in their rawest form cannot be provided, such as:

- datasets may be too large and unwieldy in raw form for most people to use; or
- machines may produce data in proprietary formats and data must be derived from these in more standardised formats so that others can access them.

There may also be cultural reasons why the rawest data are not provided, such as:

- researchers may wish to keep the raw data to themselves to use in future work; or
- a community may have settled on a certain format as a standard and be content to work with that (even though it may not be optimal).

Nonetheless, the availability of raw data does mean that checks and balances can operate at the most fundamental level. It can also be the most effective way of ensuring that the research is reproducible, a cornerstone of the scientific method. If there is anything about a dataset that must be left to trust or interpretation, reproducibility is lost.

There is therefore something of a trade-off. Derived data (those that have been reduced or processed in some way) are generally easier to work with by those who wish to build on previous findings. Something has been lost in the processing, however, and so reproducibility is compromised. In addition, derived data properly have descriptive metadata that document the processing that has taken place to produce the derived product: in a sense a second trade-off has been reached, where new metadata describe the derivation process but may lose details of the original provenance.

The raw data *vs* derived data issue is very much alive. Even in disciplines that have traditionally had a convention of sharing only derived data, such as chemical crystallography where the convention is to share data in the CIF format, there is now discussion on about the merits of sharing raw data as well, in the form of the diffraction patterns produced by the machines used to analyse the crystals.

2.2 Adding value to data

For the purpose of this project we use the term “adding value” to convey what researchers do to make data easier to discover, more accessible, richer in content and easier to re-use. Adding value in this way may constitute little more than providing a brief annotation about context; but it may involve the incorporation of additional data from disparate sources to aid analysis and give additional meaning, such as archaeologists incorporating geo-spatial data into the analysis of the data they have gathered. These are some examples of how the researchers to whom we spoke add value to their data:

- annotating data by adding descriptors or other contextual information to a dataset;
- adding additional data: for example, combining data from other sources or from further experimentation;
- aggregating and linking to other types of data, such as from a gene sequence to the published literature, or bringing together dispersed information to produce a new corpus or test bed for analysis. Such linking is common practice in some disciplines, done sometimes by researchers themselves but more frequently by professional databanks that aim to include as much contextual information as possible to give their products optimal value;
- providing metadata to make it easier to discover, access, use and curate research datasets;
- providing tools for manipulating and using the data: this is usually left to data centres, but we spoke to researchers who themselves develop tools to visualise their model data; or

- curating and preserving datasets: again, this is usually the task of data centres, but a few researchers themselves take responsibility for looking after data for the long term.

Established data centres and large databanks add considerable value to the data they curate. Depending on the condition of the data deposited by researchers, they will clean, verify, organise, document and look after the data they have received. Relatively few researchers have the skills and resources necessary to perform all these tasks themselves. Researchers working in the scientific disciplines that are catered for by good data facilities and services will often have received some instruction in how to use them, both for data retrieval and for data deposit, as part of their research training. And researchers working on the larger and better-funded projects in arts and humanities are likely to have sought advice from the AHDS or their own institution's computing centre about best ways to manage the data that the project will produce. In other fields, even closely related ones, the story can be very different, with *ad hoc*, sometimes very temporary, arrangements in place for keeping and sharing data. Some researchers store data on their own computers with little or no idea of what will happen to them in the future and with only rudimentary metadata.

2.3 Metadata

The term "metadata" is unknown to many researchers, and to others it is a source of confusion. There are many definitions of metadata but the simplest is that metadata provide information about an information resource. There have been many attempts at metadata typologies which typically describe the potential uses of metadata for discovery, use, management and preservation of information resources; but many researchers perceive the mechanics of metadata to be complex or even baffling.

Baffling or not, metadata are essential. Good metadata enable efficient curation, management and re-use of data, and they are critically important for discovery and access. Conversely, where metadata are lacking or of poor quality, datasets are difficult to discover and access; they are effectively consigned to obscurity. The extent to which effective metadata schemes have been adopted varies considerably between fields and also within fields according to the type of experiment or study being executed. In some fields there is a pronounced degree of standardisation. Largely, this derives from datasets being stored and curated in professional or semi-professional databanks where depositors must comply with a set of rules which include the provision of a structured and sometimes very detailed set of metadata.

In astronomy and crystallography, for example, there are community-developed file formats for datasets, and the embedded metadata contain all the relevant provenance details (machine used, creator's name, date of creation, experimental conditions and so on) that enable other researchers to use the dataset with confidence and complete understanding of its integrity, and thus how they can incorporate the data in their own scientific work. The same degree of professionalism pertains in areas of research that were covered by the Arts and Humanities Data Service (AHDS), which stipulated detailed and strictly-controlled conditions of metadata creation and data organisation. Such expertly-curated databanks have systems to make sure that datasets are readily discoverable and integrable wherever possible.

On the other hand, some researchers take responsibility for looking after data themselves, but provide only rudimentary metadata. In these cases, even finding the data is difficult. The lack of informative metadata, and possibly file format inconsistencies, mean that such datasets are to all intents and purposes lost to the community.

2.4 Long term viability of datasets

Long term viability of data is a thorny issue. The value of some kinds of datasets, such as those created through the longitudinal studies carried out in the social and biomedical sciences, increases over time. And there is an argument for keeping the observational data that is collected in many disciplines ad infinitum, since an event may happen only once, and at some perhaps unpredictable point in the future that event may be of interest to someone. Clearly, however, there is a very big cost to curating data forever, or even for the foreseeable future. It must also be recognised that other kinds of data may decrease in value as the focus of

research moves on, or as the earlier data are built upon and newer data products become more important. In climate science, for example, most model run data is widely assumed to have a useful life of five years.

At bench or desk level, long term care of all kinds of data – even those that have long-term potential value - is subject to various kinds of funding and other pressures. Data are very prone to becoming unusable if they are not expertly curated. Many researchers keep their data on computers that need upgrading from time to time. They have neither the time nor the incentive nor, in many cases, the skills to migrate ten-year-old datasets stored on floppy disk to up-to-date formats on DVD or hard disk. Transportable storage media (such as DVDs, CDs) are commonly used by individual researchers or small teams for keeping data, and the problems that go with that practice in terms of care, locating and access in the longer term are obvious.

The national data centres take a long term view of their remit to look after data, and in general provide a firmer guarantee of preservation and long-term access to data. But the guarantee is not watertight. The vulnerability of some of the large publicly-funded data facilities was revealed during this study. Many appear to be soundly funded and are anticipating no change in circumstance, with an implied guarantee of long term provision of accessible and usable data. The giant space science centres such as NASA would appear to be in this category. But the recent decision to close the Arts and Humanities Data Service has shown that no service, even a Research Council-funded data centre, can be assumed to exist indefinitely. In other cases, money to pay for reformatting or restructuring work, or even day-to-day archiving activities, is being sequestered from funds for research projects. International-level databanks, such as those serving the molecular life sciences, may also be vulnerable in some instances, and those that operate on a subscription-based business model face an uncertain future in an open access world.

A further problem is that the existing network of data centres does not have the capacity to accept all the datasets that are potentially valuable or useful. In our discussions with researchers and others some suggested that higher education institutions are well-positioned to curate datasets. Many of the technical and organisational aspects of this distributed approach are currently being investigated by the JISC-funded, DISC-UK Datashare project and by the feasibility study for a UK Research Data Service funded primarily by HEFCE[10]. Although some librarians are wary about taking on a data curation role, mainly because of the likely staffing and financial resource implications, a study published by the RIN[11] in 2007 showed that nearly two thirds of researchers thought that librarians do have a role to play in looking after research datasets, a position shared by many librarians.

A final issue for long term viability is the software that is often needed for access, manipulation and analysis of data. In some fields data formats are proprietary or complex, and using them requires very specialised tools that a whole field of research may depend upon. It may even be necessary to develop specialist software and regularly upgrade it to maintain its functionality as data formats alter. This presents little problem if some funder, somewhere, is continuing to fund software development, but we learned of one case – in astronomy – where such funding has ceased, leaving only one data centre in the US responsible for maintaining the software (and that centre intends to do this only for a finite period).

Recommendations

In developing their policies, research funders and institutions need to take full account of the different kinds 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.

[10] http://www.jisc.ac.uk/whatwedo/programmes/programme_rep_pres/repositories_sue/datashare.aspx

[11] *Researchers' Use of Academic Libraries and their Services*, Research Information Network (2007): http://www.rin.ac.uk/files/libraries-report-2007.pdf

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.

DATA CREATION AND CARE			
Forms and varieties of data	**Metadata**	**Adding value to data**	**Long term viability of datasets**
Astronomy			
Observational data (images, spectra and light curves) from X-ray, infrared, light or gamma-ray telescopes. Theoretical data from computational analysis of observational datasets	Embedded in the standardised data format (FITS). May be enhanced by data curation centres	Data are processed from their very raw form by the facilities into derived data that are made available to the community	Data are curated and preserved at data centres. There is a substantial cost to this activity, in terms of storage space (though the cost of this is decreasing) and in terms of training and labour for looking after data for the long term. Preservation at project level, and even in data centres in some cases, is often paid for from project funding Individual research groups or astronomers store datasets locally on hard disks or CDs, but there is little or no reformatting or migration work done on data from these sources. Accessibility of older data is an issue since software to enable this is no longer adequately supported.
Chemical crystallography			
Diffraction images (raw data). Derived datasets in the form of the CIF file	Embedded in the standardised data format (CIF). Enhanced by data curation centres, which standardise on consistent, professionally-developed schemes	Considerable value is added by the professional databanks, such as chemical structure searching, sub-structure searching, specialised software tools for analysing data	The professional databanks guarantee long term preservation so long as they are in business. They are funded by public money (in the US) and by subscriptions in the case of the Cambridge Crystallography Data Centre in the UK New, open access collections are also curating and preserving data for the long term Researchers tend to keep experimental data locally, on hard disks or on CD/DVD

Forms and varieties of data	Metadata	Adding value to data	Long term viability of datasets
Genomics			
Genomic sequences as a trace and as a sequence of 4 letters corresponding to nucleic acid base pairs. New high-throughput technologies are increasing the amount of data produced Microarray data, which show which genes are being expressed by covalent bonding to mRNA (or cDNA) labelled with a fluorescent probe. Recorded as a pattern of luminous fluorescent dots Amino acid sequences Photomicrographs from fluorescence *in situ* hybridisation (FISH) experiments ('chromosome painting')	GenBank database family has strict and demanding metadata requirements, though not all fields must be completed Main problems are inconsistent vocabularies, inconsistent annotations, lack of standardisation of gene names and, in the case of microarray data, of sample annotation	The main value added to genomics data is annotation, i.e. information regarding the function, provenance, sample details, experimental conditions etc	The large public databanks (the GenBank family) preserve data for the long term and their public funding base does not appear to be threatened Curation of metadata is good but genomic sequences themselves are not manually curated Smaller databases, such as those dedicated to one species, generally have less strict standards Funding for smaller databases is more problematic and their long term viability is more precarious

Forms and varieties of data	Metadata	Adding value to data	Long term viability of datasets
Systems biology			
As for Genomics, plus: • Protein structure data • Metabolic pathway data • Images from confocal microscopy • Mathematical models • Computer models • 3-D simulations	As for Genomics, plus: • Variable quality outside the professional databanks • Metadata provided by journals are standardised to some extent	As for Genomics, plus: • Annotation is the main added value in systems biology and of critical importance in relating information across multiple experiments, a core activity in the field • Annotation standards are not yet fully established	As for Genomics, plus: • Funder requirements to preserve data mean a major commitment by systems biology groups to preserving very large volumes of data • Larger groups use a data warehousing solution • Smaller groups may struggle to manage the large volumes of data being produced, and these will grow even more in future as high-throughput technologies and larger confocal imaging data become commonplace • Software tools for manipulating data also need to be preserved
Classics			
Catalogues; lists; lexica; annotated texts Numerical and statistical data	Purpose of metadata widely understood Expert advice on metadata schemes from AHDS Big projects invariably have good metadata	Classics researchers rarely publish raw data Published datasets tend to have significant value added Examples include editing, interpretative analysis and links to other electronic resources	Strong desire among classicists for long term viability of datasets Dismay and uncertainty about the closure of the AHDS Not all funders make financial provision for curation Many think their university should be responsible for curation of datasets

Forms and varieties of data	Metadata	Adding value to data	Long term viability of datasets
Social & Public Health Sciences			
Many of the datasets in this field are based on responses to questionnaires, interviews and focus groups The national datasets are invariably large, complex and difficult to use	National cohort studies and surveys have professional-standard metadata Individual researchers' awareness of the importance of metadata is low, as is their propensity to publish or share datasets	There is enormous value associated with the national collections of data. Data are collected properly, cleaned, verified, organised, documented and curated Other researchers add value by deriving new variables, performing new analyses and creating new datasets (which tend not to be shared with others)	The national cohort studies and surveys are designed for long-term viability (though continued funding is not always guaranteed) Some smaller datasets are looked after by the UK Data Archive But many smaller datasets have no life beyond the end of a research project
RELU			
A wide variety of data are produced spanning the biological, environmental and social sciences The range includes field observations, monitoring, lab experiments and qualitative interviews Data types include numeric, tabular, GIS, qualitative, audio and image data	There is a good general awareness of the role of metadata because award holders sign up to a Data Management Policy, and they have guidance from the Data Support Service The effective implementation of good quality metadata is, though, variable	Researchers add value to raw data in terms of data cleaning, coding and deriving higher order data These activities are usually specific to the goals of the project; it is rare to find researchers manipulating data with the aim of making it more usable to third parties	RELU researchers are perhaps better positioned than most to ensure that the datasets they produce are long-lived due to the expert support available to them (from the Data Support Service) and the opportunity to have their datasets curated by either the ESDC or one of the NERC's designated data centres But datasets are not always presented to the data centres in an acceptable form

Forms and varieties of data	Metadata	Adding value to data	Long term viability of datasets
Climate science			
Large volumes of model run data Smaller volumes of observational data Historical data Commonly used data formats include: • Hierarchical Data Format (HDF) • Network Common Data Format (NetCDF) • NASA Ames format	Sound metadata standards exist but their use by researchers is limited	Raw model run data have little value other than to the creator; processed model run data is deemed to have had useful value added Conversely, observational data have value in their raw state	Climate scientists benefit from NERC's investment in a network of designated data centres Model run data is thought to have a maximum life of 5 years (though expensive, high resolution model data is viable for a longer period) Many believe that raw data in other areas of climate science (that born of, for example, observational techniques) should be curated for the long-term

3. Publishing data: motivations and constraints

3.1 "Publishing" datasets

The concept of "publishing" datasets means different things to different researchers. Some see a natural analogue with the traditional means by which scholarly papers are published; that is, the information is "fixed" in a particular form at a particular point in time. There is, for some, also an implication that the information has been through a quality control process. The point of publication may be perceived, therefore, as a line in the sand. But researchers are also making their data available at a pre-publication stage. There may be, therefore, a need for terminology that differentiates datasets that are effectively "work in progress" from those datasets that are in their final stage of evolution as far as their creators are concerned – at which point they might be published in the formal, traditional sense. There is no commonly accepted understanding among researchers for data "sharing", though among other stakeholders the term appears to be used interchangeably with data "publishing". In some cases, however, it conveys the notion of an informal – sometimes peer-to-peer – mode of making data available.

Definitions of this kind do not currently exercise researchers to any discernible degree and so, for the time being and for the purpose of this report, it is sensible to adopt a pragmatic definition of "publication of datasets" as "making datasets publicly available". The act of putting a dataset in a data centre or other type of repository, or on a departmental or a personal website, for third parties to discover, access and re-use, is a form of dissemination. In time, a focused vocabulary might develop in much the same way as terms such as "preprints" and "postprints" are accepted and understood by many – though by no means all - researchers.

It should be noted that in some cases, what is deposited may be a selective representation of a broader mass of data that will not be made public, because researchers may have the resources to make public only a selected portion of the data they produce, choosing the most representative result from a batch of samples, for instance. It should also be noted that depositing a dataset in a collection or service does not necessarily mean that all others have ready access to it. Access may be restricted for legal or ethical reasons, for example, and often the creators of datasets impose an embargo period. There may be additional restrictions in those – relatively few- disciplines where some data centres levy a charge for access.

Each discipline has its own norms for making data public. Many have national or even international data centres or services that curate data according to the highest standards. Such bodies impose structure and quality on the data they curate, often adding considerable value to the data in the process. They may not, however, accept all the datasets offered to them, since they cannot accommodate all data being produced, and so as the volumes of data grow, the role and policies of these data centres will need to be reviewed in the light of changing circumstances.

Many datasets are never offered to professional data centres in the first place. The estimate made by the Cambridge Crystallography Data Centre, for example, is that well over half the crystallography data files generated in the UK are never deposited in its centralised service, nor see the light of day elsewhere. On the other hand, researchers may wish to make available data that represent every experiment or investigation they carried out, but have the resources to deposit publicly only one or a few representative samples of a many-dataset experiment. This means that data that could be or used by another researcher are effectively lost; although they could in theory be requested from the data creator, in practice this would not often happen.

There is a real problem, then, in relation to the large amounts of data that are not being made available. Some of these data are of potential value, either as part of the scholarly record or for re-use, and this is something that should be addressed urgently by funders.

3.2 Response to requests for datasets

Although some researchers will not share their data with others, most told us that they try their best to respond positively to requests for their data. There are positive reasons for doing so: it may lead to co-authorship of a journal paper, or the data may be cited and thus drive readers to the data creator's work.

The process of responding to requests can, however, be time-consuming, and ignoring or responding negatively may simply be expedient. Often it is not simply a case of finding an archived file and sending it via electronic means; rather, in the absence of decent metadata, there needs to be an explanation of the nature of the data, the methodology used to produce them, and the uses to which they should or should not be put. There is a limit to how much time researchers can devote to fulfilling such requests.

More often, the reason why a request is not fulfilled is simply inability to locate the data. This is common for data stored on DVD or CD and residing in cupboards, or at home, or in another institution. In such cases, the attitude to sharing can be summed up as 'willing but unable'.

3.3 Motivations to publish datasets

Researchers' behaviour in terms of how they produce and disseminate their research outputs is conditioned primarily by the Research Assessment Exercise (RAE). Because the RAE is perceived to value above all else the publication of papers in high-impact journals, most researchers focus their efforts on such publication. The pressure to publish, to compete for and win grant funding, and to repeat the cycle, is strong and persistent. Researchers' career trajectories largely depend on their success in these activities.

There are no such career-related rewards for sharing or publishing datasets. The RAE's perceived failure explicitly to recognise and reward the creating and sharing of datasets effectively stifles this activity. Many researchers report that they have neither the time nor the inclination to publish data when the value of doing so is not recognised at the top level of the funding structure. Many Research Councils have developed data management policies designed raise the profile of data management and sharing. But even where data management has become a mandatory part of the grant application process, researchers' behaviour is still primarily governed by the perceived strictures of the RAE.

One of the objectives of this project was to discover what role data publishing plays alongside conventional publishing. To the extent that it is possible to generalise, data publishing comes a very poor second to publishing papers in scholarly journals, books, or grey literature. Nevertheless, researchers do publish and share their data and, we discus in the rest of this section what motivates them to do so as well as the constraints.

It is important to stress at the outset that researchers' attitudes to making their datasets available for others to re-use vary widely according to discipline and sub-discipline. In areas such as astronomy, genomics and classics there is a tradition of sharing data, and infrastructure that facilitates such sharing. In other areas, such as climate modelling, data sharing is not the norm, re-using other researchers' model run data is not common practice, and hence researchers see little point in make data available for re-use. As we show in the rest of this section, there are various factors that encourage or discourage researchers from publishing their data.

3.4 Benefits and incentives

This project has provided valuable insights into the reasons why some researchers choose to publish their datasets, particularly in disciplines where it is not the norm to do so. Understanding the motivations of these researchers may help funders and other agencies to encourage others to also publish their datasets. Positive motivations include:

- altruism and acting for the good of scholarship;
- data sharing culture within subject or niche

- researchers who share their data are not only more likely to have the favour reciprocated, but tend to feel uninhibited about asking peers for access to their datasets;

- greater visibility for research group and institution
- opportunities for co-authorship of papers
  - in many of the disciplines covered by this study, researchers who publish and share datasets are often asked to be co-authors of papers for which re-used data is the basis;
- opportunities for collaboration with others in and beyond subject niche
  - the relationships formed through the process of publishing and sharing datasets often lead to collaborations that may not otherwise have been conceived –with other researchers either in the same or in different fields;
- esteem factors and positive feedback to funding body
  - researchers who share their data also tend to receive acknowledgements (or in some cases direct citations to the datasets themselves). This recognition, while not valued by the RAE, may be included in subsequent grant applications, especially when the funding organisation is known to encourage data sharing;
- encouragement from peers; and
- own expertise and interest in data-related issues.

3.5 Incentives

When asked what might encourage them to devote more attention to publishing or sharing their data, researchers typically point to one or more of the following incentives:

- evidence that there are benefits to be had from publishing datasets (e.g. through case studies)
- standard, workable mechanisms for citing datasets
- more explicit rewards in terms of career progression, with funding bodies and research institutions
  - taking account of formal assessments of data sharing/publishing
  - closing the gap between reward for publishing papers and for publishing data
  - taking account of past data sharing/publishing record when considering new grant applications

3.6 Constraints on data publication and use

Many of the reasons are given by researchers as to why they do not publish their datasets are the obverse of the benefits and incentives outlined above:

Lack of time and resources

Time is the governing factor in so many aspects of a researcher's working life. Some researchers perceive data management to be time-consuming and not central to their research project – so it is not done. There is a widely-held view that if Research Councils want researchers to treat data management with due seriousness, then they must provide appropriate funds. But even when Research Councils do this, some researchers feel that the funds could be better used on the research itself. Some researchers also report themselves unwilling to apply for funds for data management because it makes their grant application look more expensive and, in their view, lessens the chance of winning the grant.

Lack of time to deal with requests for information

As well as not having enough time (or funding) to deal personally with disseminating datasets, researchers worry that if they do "publish" datasets, they will have to spend scarce time dealing with requests. They may also have to provide explanations, analytical tools, metadata, further data and so forth, all of which take time to gather and transmit.

Lack of experience or expertise in data management

There are many researchers for whom data management is an unfamiliar and daunting prospect. Many of the researchers to whom we spoke for example, know little or nothing about the concept of metadata. In theory help is at hand. For example, NERC award-holders can call on experts at NERC's designated data centres for advice. In the case of RELU, a programme-specific Data Support Service (DSS) was set up to advise award-holders on data management. Yet the use of the service has been partial, and this demonstrates how, even when expert support is available, researchers will not necessarily avail themselves of it. The RELU DSS is addressing this issue through a process of personal visits, a strategy that is reported to be more successful than the passive approach.

Availability, accessibility, usability

Researchers who wish to publish data themselves – via an institutional or departmental website, for instance – rather than via a data centre have to tackle issues to do with accessibility and usability. To make a dataset available is relatively simple. To make it accessible (by, for example, providing proper structure and enough metadata so that general search engines can index the resource) requires some data-related expertise. But to make a dataset both accessible and usable by others requires expertise that is not yet widely distributed among to researchers. To be fully usable datasets must have good metadata together with comprehensive supporting and contextual information – including a description of the methodology and at least a reference to the tools and technologies used to create and analyse the data, if not the syntax itself.

Legal or ethical constraints

It is not always clear to researchers whether or not they have the rights to make datasets publicly available. This rarely appears to prevent researchers sharing datasets on a one-to-one level, but gives pause for thought when it comes to publishing the data more widely. This is especially so when a dataset is the product of a collaborative effort involving researchers from different organisations, or when a dataset has been created using data from third parties whose creators' licenses forbid the sharing or re-use of their data by people other than the licence holder. In areas of research where personal data are collected, issues of confidentiality and data protection come to the fore. There are anonymisation techniques available to mask the identity of survey participants – though the ESRC has identified a shortage of skills in this respect – but many researchers appear reluctant to obtain permission from interviewees to share the project's data, fearing that to do so might diminish the likelihood of interviewees continuing to participate in the study. Often consent is only sought only or the purposes of the original project, precluding re-use of those data for other projects.

Do not know where to archive the data

If relevant data centres decides not to accept a dataset because it falls outside their selection criteria, researchers will often not have a fallback position. They may not have the technical infrastructure or skills to publish the data themselves, and grants do not necessarily include monies to pay other people or organisations to do this work. Researchers may also not know where the best place for depositing their data might be. Although some researchers do publish and look after datasets themselves, many funders recognise that this is not an ideal use of researchers' own time and resources. Development of the technical

and human capacity to look after important datasets is an important issue for funders to address, with long-term funding implications.

Competitive factors

The role of professional competition in limiting researchers' desire to publish datasets must not be underplayed. It is a significant factor across most subject areas. Many researchers wish to retain exclusive use of the data they have created until they have extracted all the publication value they can. Many funders (and, in the case of astronomy, facilities) allow researchers a "reasonable" period of exclusive access to the datasets they have created, but the period of exclusive access is not fixed and depends on the nature of each project and the resulting datasets. Nevertheless, some researchers simply do not want to share the data they have created whatever the timeframe; and some researchers simply wish to control who has access to their data.

Fear of exploitation or misuse

Researchers feel an affinity with their data: models have been carefully crafted and observations painstakingly recorded. Some researchers express a genuine fear that their data might be "hijacked" by someone altering the data to a minimal extent and then claiming intellectual property rights. This fear is exacerbated in some subject areas where commercial organisations might conceivably seek to benefit from researchers' published datasets. Some researchers also fear that their data may be misrepresented, or that conclusions may be drawn that are not warranted.

Will anyone want the data?

Although Research Councils may give the impression that all data are unique and potentially valuable, researchers themselves don't necessarily take this view. In fact many find it difficult to believe anyone else will want access to their datasets – particularly with some data types such as model run data or those deriving from small-scale projects. This supposition appears to be confirmed when requests for access to researchers' data are few in number. This may be, of course, partly because data sharing is relatively uncommon in some disciplines, or because datasets are hard to discover

Limited or no specific reward

In an environment where researchers are measured primarily according to their publication record, there are few explicit incentives to publish their datasets. Research Councils' requirements with respect to data management do not always have the desired effect in the absence of effective monitoring and enforcement processes. Whereas the citation process is important in the world of publishing papers in journals, it is very limited with respect to datasets. Researchers will cite well-known datasets within their subject area (though there is not always an accepted format for doing so), but for less-recognised datasets the default behaviour is to cite one or more articles based on them. So unless researchers are motivated for one or more of the reasons presented earlier in this section, the default position is to do nothing in terms of data publishing.

3.6 Ownership of data

Researchers are not always clear who owns the datasets they create during the course of their work. Indeed we found that researchers are rarely clear on this issue, except in circumstances where they create datasets for a fee, in which case the data belong to the fee-paying client. This is not the norm, but it does occur, for example, in the applied sciences.

When pushed, most researchers name their employing organisation or their funder as the data owner, although they recognise that where they are jointly working on a project with people from other

organisations, the issue of data ownership becomes confusing and obscure. Data ownership is not a central issue for most researchers and they spend little time thinking about it. Whatever the legal position, researchers who have physical control over access to a dataset appear able to choose to permit or deny access to it. Ownership is potentially an important issue for funders, however, not least because it can provide a lever for them to encourage researchers to publish datasets that have a potential value for re-use.

3.7 Policies and enablers

Many Research Councils have introduced measures to encourage data publishing and sharing because they believe that datasets produced with public money should be available to other members of the research community – and indeed more widely - for the benefit of scholarship and the wider economy. They want to see data publishing on a bigger scale, and to a standard that facilitates re-use by other researchers. Effective policies and other measures to achieve such goals need to take full account of the motivations and constraints outlined earlier in this section.

Promoting data publishing and sharing is particularly challenging in a context where researchers tend not to think much about such issues as a matter of course. Data publishing to a standard that facilitates re-use requires effective planning and management of data through the life-cycle of a project. Research Councils' approach tends towards encouragement rather than enforcement, but the officers of some Councils are concerned that progress toward effective data management and sharing is too slow. Even in the case of RELU - a programme which put data management at the forefront - the outcomes in terms of data publishing are variable.

Positive measures that have been suggested to facilitate and encourage data publishing include

- promoting more actively through the use of case studies the benefits and the value to researchers of data publishing;
- providing visible top level support from employers, and offering career-related rewards to researchers who publish high-quality data;
- making clear to applicants for grants and to reviewers that including a budget to cover data management will not adversely affect a grant application (for sizeable projects in particular, providing funds to employ a data manager may help to optimise the value of the outputs from projects in the form of high-quality data);
- providing better information about and access to sources of expert advice on how to publish data[12];
- developing strategies to address the current skills gaps in data management and in quantitative analysis;
- promoting and providing better information about the control mechanisms available to data creators (e.g. embargoes, restricted access, license conditions);
- ensuring that there is an adequate physical infrastructure of data centres and services where researchers can readily deposit their data;
- promoting improved access to other researchers' data through better discovery tools and metadata standards.

Monitoring as distinct from encouraging and helping researchers to meet their data management obligations is difficult and, at present, is not routinely undertaken except in some long-term projects that are subject to interim review. Such reviews offer an opportunity to evaluate the project's data management plan, and continued funding may depend on satisfactory performance in this respect. The extent to which Research Councils should monitor researchers' performance is a matter of continuous debate, and any monitoring inevitably places burdens on both researchers and funders at a time when there is pressure to

[12] The Medical Research Council has recently launched an online resource to help researchers understand the data management process

reduce costs. Those burdens have to weighed, however, against the benefits that can accrue from data sharing, and the desire to maximise the impact of funders' investment in the research process.

Hence some funders, and indeed researchers, suggested to us that Research Councils and other funders may need to take a firmer line in monitoring as well as encouraging award-holders to meet the obligations they agree to when they receive grants for their projects. The measures suggested include:

- withholding the final tranche of an award until a dataset has been submitted to an appropriate data centre in an acceptable format (that is, in a form which can be re-used, which implies the provision of metadata and relevant contextual information) [13];
- using information about non-compliance to inform future grant applications;
- feeding back details of non-compliance to researchers' institutional employers.

3.8 The role of publishers

Most journals publish the data that accompanies journal articles in pdf format. This can be useful as a representation of the data that supports the findings presented in the article; but it does not facilitate re-use, since even when the data are linear or text-based, pdf files are difficult to work with, requiring "scraping" into another format that can be manipulated more easily. For many other data types, representation in pdf format is at best contrived and in come cases impossible.

In a statement published in June 2006 two of the scholarly publishing industry's leading representative bodies, STM and ALPSP, said

> *The associations recommend that raw research data should in general be made freely available. When data sets are submitted along with a paper for consideration in a scholarly journal, the publisher should not claim intellectual property rights in those data sets, and best practice would be to encourage or even require that the underlying research data be publicly posted for free access.*14

The message from publishers is that they have no desire to stake a claim on researchers' datasets. While scholarly publishers are maintaining their focus on journal publishing, some are already taking steps to underpin the scholarly record by creating persistent links from articles to relevant datasets, even if it is not always clear exactly what data they regard as relevant. This signposting is viewed positively by researchers and, if these links could be harvested, there exists the possibility of creating alerting services for datasets.

Currently only a relatively small number of journals require their authors to provide either links to the datasets upon which their articles are based, or the datasets themselves, for archiving on the journal's website. Where linking is used, some journals specify which repositories are acceptable to them (Genbank, for instance); others do not. For datasets archived on the journal website there is a preference on the part of the publishers for them to be in the form of pdf files. Since such files cannot easily be processed, their role is more for readers to *verify* the data underpinning an article, rather than providing *access* to the data themselves.

[13] The ESRC has adopted this approach; it will not pay the final tranche of an award until the UK Data Archive is offered a dataset for archiving. The UKDA is not obliged to accept the dataset for curation and, in practice, the act of offering a dataset "ticks the box", regardless of the quality of the dataset or whether it is presented in a form where it can be archived and re-used by others.

[14] See http://www.alpsp.org/ngen_public/article.asp?id=0&did=0&aid=1331&st=data&oaid=0

Recommendations

- ***Research funders and institutions should seek more actively to facilitate and encourage data publishing 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 gap 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); and***
  - ***h. promoting improved access to reserach data through better discovery tools and metadata standards.***
- ***Learned societies should work with researchers, funders and other stakeholders to develop and promote standard mechanisms for citing datasets.***
- ***Publishers should wherever possible require their authors to provide links to the datasets upon which their articles are based, or to provide 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.***
- ***Researchers and publishers should seek to ensure that wherever possible, the 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.***

PUBLISHING DATA: MOTIVATIONS AND CONSTRAINTS				
Publishing datasets	**Responses to requests for access to datasets**	**Motivations and constraints for publishing datasets**	**Policy and enablers**	**Ownership of data and constraints on use**
Astronomy				
Data are made public by the big facilities (telescopes) after 12 months. Researchers publish datasets to accompany journal articles, on their websites	In almost all cases astronomers will try to provide data if requested Requests to share unpublished data are usually not complied with	In the absence of mandatory policy, the main impetus to share comes from the career reward associated with publishing journal articles There is tacit community recognition and acknowledgment of data sharing Discovery and use of datasets can drive people to a journal article and thus boost citations (the real career reward)	There is no policy on data sharing from the main UK funder of astronomy research, STFC	Ownership of datasets from publicly-funded work is unclear to most astronomers Some facilities or agencies claim ownership and do not share data

Publishing datasets	Responses to requests for access to datasets	Motivations and constraints for publishing datasets	Policy and enablers	Ownership of data and constraints on use
Chemical crystallography				
Datasets are published alongside a journal article on the insistence of the publisher (and may also be deposited in the CCDC) CIF files that are not supporting an article can be deposited in the CCDC but a substantial proportion are not	Requests for data are uncommon since the assumption is that all datasets that are intended for public access are deposited in the CCDC (though this is not actually the case)	In the absence of mandatory policy, the main impetus to share comes from the career reward associated with publishing journal articles, and most crystallography journals insist on having datasets to support articles There is tacit community recognition and acknowledgment of data sharing The CCDC publishes a list highlighting the Top 200 authors reflecting the number of datasets they deposit *Acta Crystallographica E* publishes 'articles' that are really just datasets, and these can be cited like normal journal articles, enabling citations to be gained	There is no policy on data sharing from the main UK funder of crystallography research, EPSRC	Ownership of datasets from publicly-funded work is unclear to most crystallographers With respect to privately-funded work ownership resides with the funder

Publishing datasets	Responses to requests for access to datasets	Motivations and constraints for publishing datasets	Policy and enablers	Ownership of data and constraints on use
Genomics				
Genomics sequences are submitted to GenBank as a norm Best practice is to submit data from all repeats runs from a sample but this is not always adhered to There is an increasing tendency to publish data direct from the machine to a blog site Full methodologies (in contrast to the abbreviated version published in journal articles) are frequently published on project websites	Where datasets are not in public databanks researchers will usually endeavour to fulfil requests as the community norm is to share Failure to supply datasets on request may be because they cannot be found or because they have been discarded ('willing but unable')	The main impetus to share comes from the career reward associated with publishing journal articles and most genomics journals require sequence data to have been deposited in GenBank or similar public databanks There is tacit community recognition and acknowledgment of data sharing Researchers do commonly list dataset-publishing as an item on their CVs	The main funders as well as the main journals, have policies on data sharing The community norm is to share data and thus to comply with requirements; but compliance is not universal	Ownership of datasets from publicly-funded work is unclear to most genome scientists

Publishing datasets	Responses to requests for access to datasets	Motivations and constraints for publishing datasets	Policy and enablers	Ownership of data and constraints on use
Systems biology				
As for Genomics, plus: • Journal policies and standards similar to those in the Genomics area are becoming common in proteomics and metabonomics • Systems biologists generating microscopy data and 3-D images generally publish a representative selection but are looking at grid technology for visualisation and sharing of data • Some groups publish datasets on project websites: best practice is freeze-and-build but this is not always put into effect	As for Genomics, plus: • Some systems biology data are generated under commercial contract conditions in which case the entity which has commissioned the work owns the data	As for Genomics	As for Genomics	As for Genomics, plus: • Some systems biology data are generated under commercial contract conditions in which case the body which has commissioned the work owns the data

Publishing datasets	Responses to requests for access to datasets	Motivations and constraints for publishing datasets	Policy and enablers	Ownership of data and constraints on use
Classics				
Most don't publish because they consider their datasets too small There's no lack of understanding: there's a long tradition of electronic data dissemination in classics Classicists tend to have a positive attitude towards data sharing	Classics researchers are generally content to share Sharing of unpublished data is by request, though requests are usually granted Once datasets have been published other researchers are expected to download them themselves	Attachment to their subject area and a desire to disseminate information about it as widely as possible To enhance their own reputation and that of their institution As a teaching resource	Classics has a tradition of electronic dissemination of data A number of key funders of classics research favour data sharing Applications for AHRC funding have required the completion of a plan for data management The AHDS has been available hitherto to provide guidance and a place to deposit classics datasets	Ownership of datasets is not always clearly defined and a source of confusion Copyright issues can constrain sharing and re-use since classics datasets can often include data from third party sources

Publishing datasets	Responses to requests for access to datasets	Motivations and constraints for publishing datasets	Policy and enablers	Ownership of data and constraints on use
Social and Public Health Sciences				
In general social scientists are more likely to share their datasets than biomedical scientists. The UK Data Archive provides guidance and archiving facilities But some do not publish because they believe no-one would be interested in their datasets	Access to national data collections is far from guaranteed, and if granted, access is governed by licences In some cases researchers may send requests for data but are not allowed direct access to the datasets Researchers' requests for access to other researchers' unpublished data are a very hit and miss affair	Although the national datasets are published, that doesn't necessarily mean researchers can use the data at will. Access is often controlled to varying degrees and the data is often difficult to use Many researchers do not think anyone will want access to their datasets so they tend not to go out of their way to publish them Researchers in this field are judged primarily on their publication record and until this changes they don't perceive any career benefit from making their data publicly available	Researchers employed directly by the units looking after the big national datasets publish data because that's their job, but they also analyse the data and produce publications. The quality of the publications is often one of the key indicators of the value of the big longitudinal datasets. They tend to be closely monitored by their funders, and long-term funding is not necessarily guaranteed Among other researchers, many are rarely asked for their datasets and few feel the need to publish them. The ESRC and MRC are trying to change these attitudes through various means, including clear data management policies and appropriate infrastructure to support and encourage researchers to share their data	Data collected for the cohort studies or national surveys are owned by their funders Custodians of the cohort studies and national surveys control use and protect the scientific integrity of the data using licences and, in some cases, limiting direct access to the data Often access is limited or denied for legal or ethical reasons Some researchers believe there to be an imbalance between control and access of these national data resources

Publishing datasets	Responses to requests for access to datasets	Motivations and constraints for publishing datasets	Policy and enablers	Ownership of data and constraints on use
RELU				
Many researchers think the datasets they produce will not have a life beyond the end of their project, and that they are not likely to be useful to other researchers RELU award-holders are relieved of the need to worry about long-term viability since they are required to offer their datasets to a data centre (which may or may not accept it)	Most researchers are reluctant to grant access to their data at least until their project ends, mainly for competitive reasons In fact the Data Management Policy that applies to RELU award-holders permits data creators exclusive access to their datasets for up to one year	Some award holders are persuaded of the merits of publishing and re-using datasets but others are not persuaded Few of the researchers to whom we spoke see the benefit to themselves of publishing or otherwise sharing data, and cultural precedents are weak In many cases researchers simply think their data will not be of use to anybody else For most, the main purpose of their project dataset is as a basis for producing articles for publication in journals	RELU award-holders are required to produce a data management plan at the beginning of their projects, and to sign up to the RELU Data Management Policy. Award-holders also have access to a dedicated Data Support Service which provides expert advice on data management issues. Award holders are required to offer the datasets produced to a data centre which may choose to adopt them for long term curation This system of encouragement and facilitation has had a positive impact on some but by no means all researchers; compliance is variable	Most researchers think their employers own the datasets they produce They rarely dwell on issues of ownership since the data outputs rarely have commercial potential Some RELU projects produce or re-use data that is confidential in nature, or has licence-based restrictions

Publishing datasets	Responses to requests for access to datasets	Motivations and constraints for publishing datasets	Policy and enablers	Ownership of data and constraints on use
Climate science				
The datasets produced by bigger projects are often destined for curation by one of the NERC data centres But for many smaller projects little thought is given to publishing datasets	Researchers are inclined to respond positively to requests for access to their datasets, though demand for raw model run data is low Researchers report problems when requesting access to sample materials and observational data since only a proportion of these resources are held by the data centres	Many climate science researchers perceive there to be few, if any, explicit rewards for publishing data They tend to be more aware of the costs (particularly in time) of publishing data than potential benefits Climate modellers tend not to publish data, mainly because they think others have no need of the raw data, but sharing data is more common in ocean modelling Researchers producing and using observational datasets do tend to publish data There are often data sharing collaborations between researchers who produce modelling data and those who produce observational data	There is no strong culture of data sharing or publication There is a network of designated data centres, and NERC award-holders are asked to offer their datasets for curation In the near future researchers seeking NERC funding will need to develop a data management plan	NERC's Data Policy gives clear guidance on ownership and the implications for licensing, but climate scientists based at institutions are sometimes unclear about who owns the data they produce Datasets produced at NERC-funded centres belong to NERC

4. Discovery, access and usability of datasets

4.1 Different kinds of users and their needs

This study focuses on the process of creating research datasets and then making those data available in some form for others to discover, access and potentially re-use. In reality, the distinction between data creators and users is blurred, since it depends on the researcher's position in a project life cycle at a particular time: a data creator is thus likely also to be a data user. In many disciplines, researchers use other researchers' datasets as building blocks for their own work, sometimes by extrapolating key facts and integrating these into a new body of evidence, sometimes by using informatics tools and approaches to derive new information from a variety of datasets. The key point is that researchers as creators and publishers of datasets should bear in mind the needs of different kinds of users. With this in mind, we offer a three-tier typology of data users who may wish to find, access and use datasets produced by other researchers. This typology provides a reminder that datasets may conceivably be used by people outside the research community where the data were produced, and that this should be a consideration when data management plans are being developed.

- The first group comprises researchers in the particular discipline or subject area. They are familiar with the data relevant to their field of study and with any infrastructure that exists to curate datasets. They understand the processes required to generate the data in the first place, and generally have access to the tools required to transform and analyse the data.
- The second group comprises researchers from other disciplines or subject areas. They may require data as part of an inter-disciplinary study or as input to their own subject-specific models. They rely more heavily on good discovery services and need comprehensive documentation and metadata if they are to make proper sense of the datasets they find. Their lack of direct experience in how particular data are generated, gathered or processed may mean that they need support in dealing with the data appropriately and in securing access to the tools required to transform and analyse the datasets efficiently or accurately.
- The third group work outside the research community and include people and organisations in the commercial sector, central or local government. They commonly encounter real difficulties in finding, accessing and using research datasets, and they require high levels of support from data creators and others if they are to find and use data effectively.

4.2 Discovering relevant datasets

The first challenge for all data creators and publishers is to ensure that their data is discoverable. In theory, so long as data creators have provided good metadata, their datasets should be readily discoverable. As we have noted already in this report, many researchers who seek to self-publish their data know little about how to provide good metadata, and so their datasets are less likely to be found by search engines.

Nevertheless, most researchers in the first group of users report that they have no problem discovering the datasets most relevant to their work. They are well acquainted with the key sources, although many acknowledge that there may be datasets produced by small research groups or individuals they don't know about, but which may be useful. In general, researchers' discovery routines are the product of habit and are not necessarily comprehensive or especially effective; nor are they used in a systematic fashion. Few researchers consider their routines for searching for relevant (but non-core) datasets to be comprehensive, with time and searching expertise being limiting factors. Within those constraints, the main types of approaches taken by researchers in the first group are:

- searching sources with which they are closely acquainted, including the established data centres where those are available;
- turning to peers or colleagues for advice and pointers as to where relevant data might be located;

- using published articles as signposts to datasets, finding the data either as supplementary material or contacting the authors to ask for them;
- using specialised data discovery tools provided by data centres or funders, such as those provided by the NERC; or
- using a generic search engine such as Google (almost invariably a suboptimal approach, since there is usually little contextual information present to allow adequate discrimination about whether to pursue a line of enquiry or not).

Users in the second and groups find it more difficult to discover research datasets that are anything other than mainstream because they do not normally have access to a discipline-specific peer network, nor are they familiar with the relevant specialist discovery tools. Perhaps most important, because they are not closely acquainted with the subject area, they may not recognise datasets that are relevant to their information needs.

4.3 From discovery to access

The second key challenge for data publishers is to ensure that, once discovered, their data are accessible. In a number of fields of scholarly endeavour, notably in the social sciences and fields such as bioinformatics, researchers want access to data produced by other researchers as building blocks for their own work. Nevertheless, researchers may face several obstacles in their quest for access to datasets produced by other researchers or organisations. The most common obstacles are that:

- the data creator will not release the data
- negotiations are required to license certain types of data, including conditions which restrict how data can be used and disseminated;
- the data are in a form that requires tools not available to the user;
- there is a charge for re-using the data: for many databases – including some published by the Ordnance Survey, the Environment Agency and the Meteorological Office - there are fees that can stretch the budgets of even the best-funded research project;
- the dataset is too large to transfer electronically for local processing;
- there are confidentiality issues that must be respected;
- access is physically restricted where data is deemed to be sensitive; users may need to travel to a secure location in order to access data, or direct access may be prohibited, in which case users may have to send their queries to the data creators; or
- for researchers who wish to use articles as sources of data, the process of data mining may be blocked by publishers.

All these constraint on access need to be addressed or taken into account as funders develop their policies and guidance to researchers both as creators and users of data.

Looking to the future, a few researchers are already aware of and beginning to exploit the potential of Web 2.0 applications in relation to publishing datasets. In addition to blogs, wikis and commentaries, the process of tagging should aid the discovery process; network services such as Swivel and Google Palimpset will augment data hosting capacity and may thus help promote data publishing; and data mashups offer interesting ways to process data. Re-use of data may be stimulated by initiatives such as the myExperiment Virtual Research Environment[15], which offers a structured approach to re-using datasets in publicly shared workflows. Awareness of such developments is as yet low across the research community as a whole; but again, policy-makers as well as researchers will need to pay careful attention to such developments and their implications for the future.

[15] http://www.myexperiment.org/

4.4 Use and usability of datasets

Assuming that a dataset can be found and that access can be obtained, perhaps the biggest challenge is being able to use it: the usability of datasets is central to enabling effective data sharing and data publication but it is an issue often overlooked by researchers publishing data themselves. Data centres, on the other hand, invest heavily in ensuring that the datasets they choose to look after are readily usable.

Commonly, datasets are insufficient in themselves to enable other researchers to use them effectively. Data files in pdf format are especially problematic, since it may be impossible to manipulate them: in some disciplines the practice of making files available only in pdf format is known as "protecting by pdf". Even when the file format is satisfactory, however, users require contextual information about, for example, how the data were collected and what tools or syntax were used to derive new variables or produce particular analyses. Providing the information necessary to render a dataset usable is thus of critical importance. It can also provide a means to alleviate data creators' concerns about their data being misrepresented or used inappropriately.

Particular issues arise with dynamic datasets where original data may be amended, added to, or replaced by newer data at a later date. The optimal approach in such cases – not always adopted – is to 'freeze and build', so that original datasets are preserved and made available alongside, rather than being replaced by, the newer datasets. This means that the original dataset remains available for analysis and re-use, that changes can be identified, and that any erroneous annotations or amendments can more readily be rectified. Even with 'freeze and build' procedures, however, it is not always clear whether multiple validations have taken place, or whether earlier data have simply been incorporated and assumed to be correct without further validation.

4.5 Tools and technologies for analysis

For this report we have made the assumption – which is substantiated by our interviews – that researchers normally have the tools and technologies they need to do their research and create their datasets. This section, therefore, focuses on the tools necessary for accessing, manipulating and analysing data. They vary from the simplest possible – reading a paper-based report – to writing machine code almost every time a new dataset is to be used, as can happen in astronomy. Although tools and technologies can be very sophisticated, we have not found that this presents a significant problem for most researchers. Considerable training may be necessary to develop the relevant skills, and often significant time and effort must be spent in solving problems that arise. But generally speaking, within each community the tools to enable access and re-use of data do exist.

Raw scientific data, where available, often pose the greatest problem since there may be very little or no exchangeability in raw data from machines. In these cases high levels of programming skills may be required to allow access and re-use. Larger research teams facing such problems may use the services of software engineers or mathematicians to facilitate access and analysis. Specialised data experts may also be employed fulltime on larger programmes where extensive data manipulation and database skills are required. As we noted in our discussion of long-term viability, specialist software tools may need to be developed and regularly upgraded to ensure that datasets of long-term value remain accessible and usable.

The emphasis of this study has been on discrete datasets of one form or another, but there is an additional source of data that is of growing interest to some researchers: the data that reside within the text of published articles. Experimentation with text-mining is becoming more common in many areas of research. The UK National Centre for Text Mining provides text mining services, information, help and training[16]. The basic premises are that:

- information resides in text;
- pieces of information in disparate texts may be related; and

[16] http://www.nactem.ac.uk/

- if those relationships are examined and exploited, new information may be learned or created

Text-mining is still a relatively young technology but it has huge promise and it is likely to become increasingly important. Some confusion has arisen, however, as to publishers' policies with regard to allowing access for text-mining tools to their journal contents. Current uncertainties need to be resolved if the potential of this technology is to be realised.

Recommendations

- ***Research funders and institutions should seek more actively to facilitate and encourage data publishing and re-use by***
  - ***a. promoting improved access to other researchers' data through better discovery tools and metadata standards***
  - ***b. identifying and documenting by subject area barriers to effective re-use of data, and promoting guidance on good practice***
  - ***c. 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***
- ***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***
- ***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.***

DISCOVERY, ACCESS AND USE OF DATASETS			
Discovering relevant datasets	**Access to datasets**	**Use of datasets**	**Tools and technologies for analysis**
Astronomy			
Good metadata standards prevail Every observation is assigned an ID number that can be used to track data Journal articles cite the accession number of datasets in the archive where they reside Large facilities and national data centres provide professionally curated databases Discovering data other than those deposited in large public databanks can be problematic	Curated data can be accessed at the facility where they were collected (made publicly available after twelve months) or in data centres Derived data can be accessed from journal websites or by requesting them from the creator There is often the need for software to be written to access others' data but this is normally within the skillset of astronomers	Re-use of data is a commonplace in this community, though the development of software tools to permit this may be needed Data sharing is a norm	Researchers access FITS files using standard software packages or their own versions Computer scientists may be employed on larger teams to write software for accessing or analysing datasets produced by others
Chemical crystallography			
Data are found in the public databanks, in *Acta Crystallographica E* and as supplementary data in the main crystallography journals Discovering data other than those sets deposited in large public databanks can be problematic	Access is normally through public databanks New open access collections are growing	Re-use of CIFs in the CCDC is permitted New open access collections present content for re-use	Analysis of CIF files is simple and the CCDC provides a range of software tools for this purpose (for free) Access to data in rawer form is limited but increasingly considered desirable to enable more thorough validation and to exploit the data more fully

Discovering relevant datasets	Access to datasets	Use of datasets	Tools and technologies for analysis
Genomics			
Data are found in GenBank or in smaller public databases used by specific communities Data, especially DNA primer sequences, may also be discovered using Google Discovering data other than those sets deposited in large public databanks can be problematic	Access is normally through public databanks Some databanks may levy a charge but these are rarely considered useful or central to a research project	Re-use of sequences in public databanks is permitted and is a fundamental part of genomics work Re-use of privately held data may be by negotiation	Software tools for manipulating data are generally available in large laboratories and may also be provided by databanks Researchers write their own scripts for accessing non-standard data, or enlist the help of bioinformaticians or computer scientists for this task
Systems biology			
As for Genomics, plus: • Systems biologists frequently trawl other groups' project websites for data • Discoverability can be poor via this route and it is very time-consuming	As for Genomics, plus: • Project websites are used quite extensively as data sources • Accessibility via this route can be poor, requiring specialised software to be written • Freeze-and-build is not a ubiquitous practice, thus precluding access to older datasets or very raw data	As for Genomics, plus: • The importance of software for access and manipulation is high	As for Genomics, plus: • Systems biology data require mathematical or computing technologies for manipulation and analysis • Larger groups have one or more data managers, programmers or bioinformaticians in their team who write software and construct databases • Commercial software is available for working with systems biology data but is expensive • Open source software solutions are also available but this is not a preferred solution for many systems biologists

Discovering relevant datasets	Access to datasets	Use of datasets	Tools and technologies for analysis
Classics			
No general problems The discipline is of a size where people normally know what resources exist in their own field There is some low level interest in Web 2.0 technologies for data discovery	There are few cost barriers in the way of accessing third party datasets Where charges exist they are normally very low, and are designed to enable data creators recoup the direct costs of providing the data		Not much automation other than spreadsheets and off-the-shelf relational database packages
Social and Public Health Science			
It is a straightforward matter to find the big datasets produced at a national or regional level; the key ones are well known to researchers It can be very difficult to find the results of smaller projects, mainly because people tend not to publish them. This means discovery via printed publications and access by direct request to the creator	Access to the big national or regional datasets can be straightforward if they are requested from the UKDA Access to the big datasets directly from the creators can be more difficult, sometimes very difficult Important large scale datasets can be behind high toll and other control barriers, though these are the exception	Many researchers feel a moral obligation to extract the maximum research value from the big national datasets that have been created with public funds There is therefore some resentment on those occasions where barriers are put in the way of researchers trying to obtain access to important datasets	Social scientists tend to spend a lot of effort analysing data, deriving new variables and incorporating data from third party sources. They often write their own syntax for data analysis. Biomedicine researchers in the field have a reputation for excellent data collection techniques

Discovering relevant datasets	Access to datasets	Use of datasets	Tools and technologies for analysis
RELU			
In the subject themes covered by the RELU programme, the large relevant datasets are well-known to researchers Other discovery mechanisms mentioned often were the use of peer networks, and attendance at conferences and other meetings In many cases researchers were producing primary data and had little need of discovery tools	For those researchers who required access to third party datasets, they tended to be major datasets with a spatial basis; OS and Environment Agency datasets for example Gaining access to these can be very expensive, running to tens of thousands of pounds	Some RELU award-holders had a need to use datasets produced by third parties as building blocks for their own work Researchers normally want to use processed rather than raw datasets Sometimes researchers purchase data produced by commercial organisations, such as those who regularly collect data from farmers	The RELU programme is interdisciplinary by design, so an array of discipline-specific analytical tools is used Researchers tend not to question the tools or techniques used by their project collaborators whose training is in a different discipline; there is a high level of trust required
Climate science			
In some sub-disciplines there are only a few relevant datasets and these are well-known to researchers Researchers will use their peer network to find datasets They will use published papers as signposts to datasets Some use the facilities offered to search the datasets curated by the NERC data centres Researchers also use generic tools to search the web These discovery methods tend not to be used in a systematic fashion	The NERC data centres offer good access to a large number of datasets Researchers can also access data produced by key national or international organisations such as NOCS, ECMWF and PCMDI. A very few research groups make their data available via websites Access to many datasets is possible only by direct negotiation with the creators	The use of raw model run data is very limited except in the case of high resolution models; there is more demand for processed datasets There is consistent demand for other types of climate data born of observation, remote sensing or physical sampling (ice cores, for instance) Some researchers – such as those for whom palaeo data is important – would like to see better arrangements for the curation, access and use of this type of data	Climate scientists use a variety of tools and technologies for data analysis In the modelling community the emphasis is on building and refining computer models and then processing the raw data

5. Quality assurance

5.1 Quality assurance in the data creation process

The term "quality" is conventionally associated with the notion of being "fit for purpose". With regard to creating, publishing and sharing datasets we identified three key purposes: first, the datasets must meet the purpose of fulfilling the goals of' the data creators' original work; second, they must provide an appropriate record of the work that has been undertaken, so that it can be checked and validated by other researchers; third, they should ideally be discoverable, accessible and re-usable by others. Fulfilling the first and second of these purposes implies a focus on scholarly method and content; the third implies an additional focus on the technical aspects of how data are created and curated.

Researchers are interested in producing the best data they can in order to answer the research questions they are posing, but also to provide a sound basis for producing papers that pass the scrutiny of their peers and are published in reputable journals – this being one of the key drivers of career progression. Most researchers also believe that data creators are best placed to judge the worth of their own datasets and that, on the whole, these judgments fairly reflect their scholarly value.

The requirement to uphold reasonable quality assurance standards is partly met in the sciences by machine efficacy: most machines that create data (such as telescopes, spectrometers, gene sequencers) have inbuilt data checking and verification steps. Manual checking is usually added, and in those disciplines where data are collected by other means manual verification may involve very detailed work. There is pride in turning out datasets of good quality and shame in exposure as a creator of flawed or incorrect data. Research communities are thus to a large extent self-regulating in respect of data quality assurance. As a result, most researchers reported to us that they generally take other researchers' outputs on trust in terms of data quality and integrity, and we received no reports of dissatisfaction with this state of affairs.

5.2 Data management planning

In all disciplines, larger projects that are receiving or expecting to receive substantial levels of funding engage in a data planning process. A formal data plan is included in grant applications – written by the data manager if the team has one – and this covers the kinds of data that will be created and how; how they will be manipulated; where they will be stored; and how they will be made available for sharing.

Smaller projects typically do not have this degree of formality or concern about data management. Researchers will acknowledge that their funder has a policy or guidelines on data management, and if this includes a requirement to produce a data plan then they will write something in their grant application, usually at the last instance and quite often with little care. There is, therefore, a big difference between the professional and detailed approach taken to data management by some researchers and the cursory approach taken by others. In summary, data management planning is at present highly variable in quality.

5.3 Quality assessment of datasets

At present, the scholarly merit of data is assessed by the peer community by comment, re-use, and building upon data outputs. This is done at two stages: first, when an article or monograph is peer-reviewed by other researchers in the field prior to acceptance of the work for publication; and second, when the community accesses and uses the published work.

Peer review may involve checking supporting data in a more or less detailed way. In some disciplines, reviewers check data extremely thoroughly and are capable of unearthing flaws or inconsistencies at this point. In other cases, checking is less than thorough, partly because reviewers may not be able to judge the data satisfactorily, partly because datasets may be too large to review in their entirety, and partly because the data may be too complex to be judged in this way. Reviewers may check that the data are present and in

the format and of the type that the work warrants, and leave it at that. Overall the approach is uneven. There is a concern also that even if peers have the skills to review the scholarly content, they may not be able to judge the technical aspects of a dataset that facilitate usability.

It might be possible to have a two-stage review process focusing on content and technical merit separately, though many argue that in a digital environment the two are interdependent. A report commissioned by the Arts and Humanities Research Council17 and published in 2006 made a number of recommendations which included consideration of a two-stage process at grant application stage, focusing separately on scholarly content and technical issues, along with an open post-completion review where paid reviewers' comments would be attributable and data creators have a right of reply

Variability in and concern about the quality of peers' assessment of the content of the datasets that underpin publications is one of the key reasons why many researchers to whom we spoke do not discount the idea of instituting a formal process for assessing the quality of datasets. Some researchers are mildly enthusiastic but – and it is a big but – no-one can see it working effectively in practice. Several concerns were expressed:

- that it would be difficult to find reviewers with sufficient expertise in highly-specialised fields to understand the data, let alone appraise it;
- that the pool of researchers willing to take the time to review journal articles is diminishing, and that hard-pressed reviewers would be even more unlikely to want to take on further work in assessing datasets[18];
- the costs of the review process, who would pay, and whether the money might not be better spent on research; and
- that having to have a dataset reviewed would add time to the research process at a point in the project life cycle where researchers want to be writing papers for publication.

In summary, there is some sympathy with the concept of expert assessments of the quality of datasets, but researchers don't see how it might work in practice and, given that they are not unhappy with the present situation, there is no grass-roots pressure to introduce a formal assessment process. That is not to say that, in time, research funders themselves might wish to see a more rigorous and consistent quality assurance process for datasets, particularly if they, along with other organisations, are investing heavily in the infrastructure required to support their publication.

There is, however, more to quality assessment than just the consideration of the scholarly merit of a dataset. If the process of data sharing is to become more effective and useful, much more consideration needs to be given to making datasets accessible (through the effective use of metadata) and usable (by providing the information and possibly software tools necessary for others to re-use the data without the help of the data creator). Data centres and databanks come into their own in this regard, applying stringent rules and checks that ensure that datasets deposited meet quality standards, both of the structure and format of data themselves and of the metadata. Where a dataset has significant scholarly merit but is lacking other respects, data centres will normally work with data creators to ensure that the dataset is discoverable, accessible, and usable.

Whilst datasets that are accepted by data centres must conform to such standards, there is no such imperative for researchers who look after their own datasets. They may believe this to be outside the boundaries of their research function but, perhaps more importantly, our study indicates that many researchers do not have the skills to publish their data such that it can be discovered, accessed and re-used by the scholarly community at large and beyond. Whether the creators of datasets should be encouraged to gain such skills through education, persuasion, grant conditions or other means is an issue for Research

---

[17] *Peer review and evaluation of digital resources for the arts and humanities*, Arts and Humanities Research Council, September 2006 http://www.britac.ac.uk/reports/peer-review/index.html

[18] This echoes concerns reported in a new study of researchers' attitudes to peer review where 40% of reviewers and 45% of journal editors said it was unrealistic to expect peer reviewers to review authors' data. Ware, M (2008), Peer Review: benefits, perceptions and alternatives. http://www.publishingresearch.net/documents/PRCPeerReviewSummaryReport-final-e-version.pdf

Councils, other funders and the data centres to consider. An alternative approach would be to train and recognise the value of data scientists – whether from a research or information background – whose role would be to work alongside researchers, helping them devise and achieve the goals of effective data management plans.

Recommendation

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.

QUALITY ASSURANCE		
Quality assurance in data creation	**Data management planning**	**Quality assessment of datasets**
Astronomy		
Large facilities provide some level of quality assurance at data creation stage Researchers check the quality of data before use or publishing Data centres carry out extensive and detailed data checking procedures	Grant applicants usually write a description of what data will be collected during the project and what will be done with them STFC does not have a mandatory policy on data so there is no requirement to produce a data plan, though in practice this is always present in some form and is expected of bigger projects Such projects view a data plan that emphasises sharing of data as something that will enhance their chance of gaining funding	Datasets may be peer-reviewed if they are offered as supporting data for a journal article, but these are always derived data not original raw data Peer review may be more or less careful and detailed and sometimes cursory
Chemical crystallography		
Quality assurance is implicit in the data creation phase as machines have some degree of quality checking at the creation stage Data integrity is checked manually by data creators before storage	No formal data plan is required by the main UK public funder of crystallography work, EPSRC Most people give some thought to what they will do with their data and generally store derived data on hard disk or DVD	Datasets may be peer-reviewed if they are offered as supporting data for a journal article, but these are always derived data not original raw data Peer review may be more or less careful and detailed and sometimes cursory The CCDC carries out extensive checks on data integrity of CIF files deposited in its databank The IUCr provides a tool, CHECKCIF, for checking the integrity of CIF files (requires training in its use)

Quality assurance in data creation	Data management planning	Quality assessment of datasets
Genomics		
Sequences are manually checked before deposition in public databanks in most cases In smaller projects, one sequence representative of all repeat runs may be checked and deposited There is a generally-accepted error rate in all sequence data	Data plans are required by the funders BBSRC, MRC, NERC, the Wellcome Trust and the European Union Larger teams running big projects may submit elaborate data plans Small teams may regard this as the least important element of a grant proposal	GenBank operates a careful manual metadata-checking system though the sequences themselves are not checked Post-deposit checks are by use and reporting of errors by the community Journals play some role in quality control through peer review but this it patchy and limited
Systems biology		
As for Genomics, plus: • There is some level of quality assurance given by the machines used to generate data	As for Genomics, plus: • Large systems biology teams have a data manager to prepare formal data plans and to manage their implementation	As for Genomics
Classics		
There is a tradition of careful editing and proof reading of data Classicists tend to trust the quality of their peers' datasets	AHRC award holders have been required to complete a Technical Appendix to their grant applications Expert advice has been available from the AHDS, though this is due to close There are some concerns about version management	Classicists would not mind additional quality assessment, but it would be more work It is thought datasets are already adequately assessed via the funding application process and peer review of publications that are based on datasets

Quality assurance in data creation	Data management planning	Quality assessment of datasets
Social and Public Health Sciences		
For the big datasets quality assurance is said to be very good and there are formal procedures to ensure high quality data creation Researchers tend to trust each others' datasets, though where people don't know of the data creator they will do some simple tests on the data	Data management planning is an integral part of creating the big cohort studies and national survey datasets It tends not to preoccupy the thoughts of many individual researchers Research councils such as the MRC are trying to change this by requiring consideration of data management planning in grant applications	The big national datasets are normally subject to external assessment and many are required to re-apply for funding periodically Among researchers generally few see the immediate necessity for quality assessment, though many see the practical drawbacks. People need subject-specific expertise and time to provide reliable assessments
RELU		
There tends to be no cross-disciplinary checking; the expectation is that project collaborators will produce data to the accepted quality standards that prevail in their discipline The different disciplinary strands of a project are brought together at meetings and workshops, but the quality of peers work would not normally be questioned	RELU award holders are required to submit a data management plan early in their project, but this is no guarantee of quality. Many researchers are reported to be unconvinced of the case for data management planning In the later stages of the RELU programme, substandard plans have had to be revised and re-submitted. Early indications are that the funder's insistence on high quality management plans is beginning to produce dividends	There was little appetitite among our interviewees for external quality assessment There is a general concern that any external assessment process would divert money from research funding There is also a concern that reviewers would not have time to review datasets which, in the view of some, are unlikely to be used by others
Climate science		
Climate science researchers in the UK believe there are few if any problems with quality assurance in the data creation process	Researchers are able to ask advice from NERC's data centre staff about data management planning, but submitting a data management plan has not been a condition of awarding funding This position has been under review and may change in the near future, reflecting the patchy nature of the quality of data management processes employed by some researchers	We found no enthusiasm for formally assessing the quality of datasets produced by climate scientists They are content with the current system whereby peer review of journal articles is, by proxy, a review of the reliability of the science and the datasets underpinning those articles [though this process does not properly address the usability of those datasets]

SUMMARY OF THE POSITION IN EACH OF THE EIGHT AREAS COVERED IN THE CURRENT STUDY				
	Culture of sharing data	**Infrastructure-related barriers to publishing data**	**Effect of policy initiatives to encourage data publishing**	**Overall propensity to publish datasets (with appropriate metadata and contextual documentation)**
Astronomy	High	Low	Medium	High
Chemical crystallography	Medium	Low	Low	High
Genomics	High	Low	High	High
Systems biology	Medium	Medium	High	Medium
Classics	High	High[19]	Medium	Medium
Social and Public Health Sciences	Low	Low	Low	Low [20]
RELU	Medium	Low	Medium	Medium
Climate science	Low	Low[21]	Medium	Low to Medium

This table provides a summary of the position in each of the eight areas covered by the current study. Detailed reports on each area are available in the Annex to this report, available on the RIN's website at www.rin.ac.uk/data-publication . Note that because 'high' may be positive or negative depending on the factor being graded, we have denoted the overall grade by means of colours for cell shading. These denote where a factor merits a positive grading (green), medium grading (amber) or negative grading (red).

[19] AHDS ceased to exist on 31 March 2008

[20] (for researchers not directly connected with a national data collection)

[21] NERC provides data centres

6. Methodology

6.1 The broad aims of the project

The broad aims of this project were:

1. to investigate, with reference to selected areas of the scholarly spectrum, the nature and range of arrangements for making research data as widely available as possible (referred to as "data publishing"), and the role that data outputs currently play alongside or as an alternative to conventional publications in the research communication process; and
2. to investigate current practice for ensuring the quality of such data.

It was clear from our initial investigations that researchers in different parts of the scholarly spectrum would report different experiences with respect to data publishing, not least because different subject disciplines are served by different infrastructures in relation to data management and curation. The challenge, therefore, was to select subject areas that not only covered the four broad disciplinary areas (arts and humanities; social sciences; life sciences; physical sciences) but could also provide insights into particular issues. These might include, for example: disciplines where researchers have a tradition of data sharing and others where they don't; disciplines where researchers are well-served by a network of data centres and so on. After much discussion and consultation with a panel of experts, the following six subject areas were selected:

- classics, where the Arts and Humanities Data Service has been key to increasing the community's level of knowledge about data management and has provided a clear route for the curation of datasets;
- astronomy, where there is a tradition of data sharing and where the quantities of data produced are enormous;
- chemical crystallography, where innovations in infrastructure provision has paved the way for more efficient data publishing;
- genomics, which in many respects leads the way in terms of science data publishing;
- social and public health sciences, a field where a few important longitudinal datasets provide source data for many researchers in the social and medical sciences; and
- climate science, where researchers are served by a network of NERC-funded data centres.

In addition to these subject areas, a further two *inter-disciplinary* areas were selected. These were:

- *systems biology*, where policy initiatives are having a positive effect on data publishing; and
- *the Rural Economy and Land Use programme*, which brings together scientists and social scientists to work together to tackle some of the fundamental issues affecting the UK's rural economies.

6.2 The approach to the project

The issues central to data publishing are complex and each different subject area has its own set of discipline-specific issues and terminology. For these reasons it was decided to adopt a qualitative approach to the project. For each of the six subject areas fifteen interviews in-depth telephone interviews were conducted with a mix of experts (researchers with significant experience of publishing data) and researchers. For the two interdisciplinary areas, ten in-depth interviews were conducted with a mix of experts and other researchers. Although we encountered the usual problem of long lead times to book interview slots with researchers (up to several weeks in advance in some cases), the majority of participants engaged fully with the interviews, many of which lasted more than 60 minutes. Clearly the sizes of the

judgement samples in each subject area are limited but they were, in our view, sufficient to provide reasonable insights into researchers' attitudes within different subject areas and across them.

In preparation for the interviews, comprehensive lists of subject-specific data-related issues were drafted with the help of subject experts. These long lists were then considered and distilled to a list of core issues which then formed the basis for interview guides. These guides are essentially lists of questions and issues to be covered during the course of an interview (though not necessarily in a particular order; the conversational approach is more palatable to interviewees and the guides help the interviewers ensure all the bases are covered). The questions and issues were broadly common across the interviewees across all the subject areas, in order to enable comparisons to be made. Indeed the results not only signpost important issues within each of the subject areas studied, but they highlight attitudes to data publishing and the quality assurance of datasets that pertain across the breadth of scholarship.