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

Report commissioned by the Research Information Network (RIN)

Annex: detailed findings for the eight research areas

June 2008

www.rin.ac.uk

In association with:

CONTENTS

CLASSICS

Overview

Classics is an interesting subject to consider as part of this study because it is rather different from the other subject areas we have looked at. In the first place it has a long history of database publications and has been active in the electronic dissemination of data for nearly 30 years. In fact some classicists are already exploring more advanced data publishing related issues such as Virtual Research Environments, wikis and interoperability, often in conjunction with scholars based overseas. Second, although classicists on the whole do not produce a great deal of data, those who do reflect a culture of data sharing and of taking pride in publishing their datasets locally rather than simply sending them off to a central repository or data archive. The classicists who publish datasets include archaeologists, epigraphists, prosopographists, ceramists and art historians. They produce data such as edited catalogues, lexica, lists as well as numerical and statistical data.

Researchers in classics tend to be well aware of the issues to do with publishing datasets and, up to now, they have been well served by organisations such as the Arts and Humanities Data Service (AHDS) and the Centre for Computing in the Humanities at King's College London. Classicists are encouraged to produce data management plans and to archive their datasets by key funding bodies. There is, however, an atmosphere of uncertainty as the AHDS is to lose its funding in March 2008 and will cease to exist in its current form. The Archaeology Data Service will continue to function. This means that the official channel for giving advice to those who wish to publish data will no longer be available and that the official subject repository for published datasets will cease to exist. It is possible that the advisory role performed by the AHDS will be replaced by one or more technical reviewers.

This uncertainty has led to calls for subject centres of excellence to provide advice on data-related issues. As for archiving, in the absence of the AHDS many classicists feel that their own institution should host their datasets, but it is unclear how this should be funded or what might encourage their library or computer centre colleagues to take on this responsibility.

Data creation

Form and variety of data produced

Researchers working in the classics produce a wide variety of types of data. As well as catalogues, lists, lexica and annotated texts classicists also produce numerical and statistical data, tables and collections of digital images. In some cases researchers may create specific software for a project. In certain areas of classics data from external sources are being used to augment core data so, for example, some classical archaeologists are beginning to use spatially-oriented data to add value to the analysis of their own datasets.

Purposes of data generated as research output

Although few classicists set out to create electronic database resources for dissemination purposes, they do often produce datasets as part of their research projects. These datasets are then used as the basis for publishing books, monographs or journal articles. It is not uncommon for classicists to share digital images of artefacts. It would be very rare for raw data to be published; rather, the types of datasets that researchers in this field share have undergone significant analytical and editorial processes, adding considerable value to them.

The main purposes to which the data are put share common ground with the goals of researchers in other fields, namely to augment the authors' personal reputation and that of their employing institution. The RAE looms large over classics as it does over other subject areas, but we did discover an additional purpose that was rarely mentioned outside classics, which is the use of datasets as teaching tools for the academic and wider communities.

Metadata

For those classicists who produce datasets as part of their work, there is a general appreciation of the role of metadata – perhaps more so than in some other subject areas. This appreciation has come about relatively recently since older data that have not been migrated to more recent software platforms are less likely to have metadata associated with them. The bigger and better funded a project, the more likely it is that researchers will have sought advice at an early stage from a university computing centre or the AHDS which has specific policies relating to metadata and which has published guidelines on the subject. Most of these researchers understand the need for using standardised metadata but, where they consider the available standard metadata schemes to be lacking, they are likely to devise their own. They would be prepared to upgrade their metadata if this was thought to be important by their funders, but there is concern that this would carry costs in terms of time and money and that specific funds would need to be made available to do this type of work.

Adding value to data

As noted previously, classicists tend to add a great deal of value to the datasets they produce before they are made publicly available. Nearly all data published in the classics are subject to some form of editing and interpretative analysis, some of which may be substantial. Where raw datasets are numerical in nature, they are often analysed using statistical software to derive new variables and higher-order datasets.

Where major datasets are being produced, value is often added through the integration of links to third party online resources such as the *Thesaurus Linguae Graecae* and the Beazley Archive. There may also be links to online publications such as journal articles or books. In the case of some museum databases audio links have been incorporated. In sum, it is clearly the intention of the directors of the large database projects in classics that they integrate their database as closely as possibly with other relevant electronic information resources in the field, thereby adding a great deal of value to the data from a users' perspective.

Some datasets in classics are published electronically as supplements to printed materials. This route allows the publication of datasets where otherwise the cost of publishing the dataset would have been considered too expensive. These supplementary datasets permit

the development and updating of what would otherwise have been a static, print bound resource. This facility is seem by many researchers in classics to be a great advantage.

Long term viability of datasets

On the subject of the long term viability of datasets the prevailing attitudes of classicists are fundamentally different from those of researchers in many other subject areas. Whether classicists' datasets are designed to be static or continually updated, they are always intended to have viability as long as there is community interest in the subject. In that sense they are perceived to be no different from a book or other types of printed publication.

Where data are the intended end-product of a project, the funding bodies of the research all more or less insist on the datasets being made publicly available. The funding may come from a charitable trust (such as Leverhulme, which has no formal requirement but looks favourably on data sharing) or from the British Academy, JISC or AHRC. There are occasions where the publisher of a printed resource may try to prevent the publication of related datasets, or at least put in place an embargo. This approach invariably leads to disagreements with the content creators who usually want to make the data – print and electronic – as widely available as possible.

Projects in the classics obtain a substantial part of their funding from private and charitable sources. These sources usually want the datasets resulting from these projects to be made publicly available as a teaching tool but they do not necessarily provide funding for the maintenance of a dataset once it is completed. This can be a source of anxiety for classicists who are coming to realise the importance of highlighting in their funding applications the need for funding to future-proof project datasets. In contrast, large projects which require substantial funding have always included costs for long term sustainability in their applications for funds. In these cases typically the researchers will negotiate directly with the funders how best to make the dataset accessible in the long term.

Technical guidance for data management planning including curation is available but researchers need to proactively request such guidance from an advisory body such as the AHDS, the Oxford Text Archive or the Centre for Computing in the Humanities. While some researchers turn to their local university computing centres for help, these centres will not necessarily have the manpower or subject-specific expertise to be able to assist. This is perhaps why many researchers rely on their network of personal contacts to seek advice on how to ensure the long term viability of their datasets.

In order to maximise the potential usefulness of datasets to users other than their creators, it is important that appropriate contextual information is provided (normally in addition to the standard metadata). The availability and quality of contextual information is, however, highly variable and largely dependent on the quality of the guidance provided by advisory bodies or other expert sources. It seems only the projects that are particularly well organised with respect to data management planning have allowed for the production of contextual documentation in a systematic way.

Despite the fact that many classicists care about the long term viability of their datasets, they are also uncertain as to their future. The key problems are these:

- Because there is often no provision for the long term funding of datasets that have been made publicly available by classics researchers, software is not being upgraded and, in some cases, is becoming obsolete.
- Now that the AHDS will no longer be available to act as a repository, many dataset creators are left wondering how their datasets are to be hosted and maintained. At the time of our study, some dataset creators have not yet been told what is likely to happen to those datasets.

So what will these structural changes mean for the curation of datasets in the classics? At well-funded universities which regard their faculties of classics as fundamental to the reputation of the institution, Oxford University for example, researchers take it for granted that their datasets will be hosted by the university for as long as required and that the faculty will support the datasets and their creators for as long as is necessary. Some smaller universities may be willing to support the curation of classics datasets if they are judged to enhance the academic standing of the university. Other institutions, however, appear unwilling to look after datasets for the long term without specific funding for this purpose. Even King's College London is report to be prepared only to guarantee 10 to 15 years of support. Some institutions do not have digital repositories, at least not ones capable of hosting datasets and there is some doubt as to whether librarians or repository managers will be willing or able to be responsible for looking after classics datasets.

There is little doubt that dataset creators in the classics want the integrity of their datasets to be protected for the long term. In light of the decision to cease funding the AHDS from March 2008, researchers' options are becoming more limited. Some feel that curation anywhere would be better than not at all, even to the extent that a few researchers have considered approaching commercial publishers to take over the curation of their datasets on the basis that publishers sometimes wish to produce publications derived from those datasets. This option is not, however, generally thought to be a satisfactory solution since they want the datasets to remain in the public domain and they are not keen to see datasets locked up behind commercial toll barriers.

There is currently a division of opinion about how best to curate classics datasets. Some believe a central solution would be best, where a national data service for the classics would be set up to look after datasets for the long term. People believe such a service could be funded by the Government and offer subject-specific curation expertise. On the other hand, some favour a distributed approach where the institutions of dataset creators host and possibly fund the curation of datasets – with the national service perhaps acting as a backup. Most researchers agreed that the funding organisations could and should bear part or even all of the cost of curating the datasets resulting from the work that they choose to fund.

Working with the created data

Tools and technologies for analysis

Classics is a field that still relies on time-honoured and sometimes painstaking intellectual analysis; there are not many niches in the subject area that can benefit from specific

technological advances to expedite their work. Datasets are normally created using standard tools such as Excel spreadsheets and off-the-shelf relational database products.

Publishing datasets

Classics is one of those subject areas where, in general, researchers want their work to be seen and appreciated by as many of their peers as possible but unlike many other subject areas, researchers tend not to compete head-on; rather they develop particular expertise in particular niches. In this context, making their datasets publicly available is perceived to be a desirable end. They recognise the benefits to classical scholarship of being able to update datasets that are publicly available – to provide information about new excavations or new artefacts or texts that have been discovered, for example. Researchers tend also to only publish processed datasets which have had value added in various different ways. The exception is for classical archaeologists, who may present raw data in excavation reports, and museum databases where most of the data checking is a continuous process.

Although classicists are keen to disseminate their research findings, their views on how best to achieve this are mixed. Some decide at the very beginning of a project that their datasets will be published and that, where possible, any articles they produce from that dataset will be published on an open access basis. Others favour a mixed approach, making their datasets freely available to everyone but publishing their written analysis via traditional channels such as books and paid-for journals. Indeed there is a view held by a few that the traditional publishing channels garner more prestige but they don't discount the value to others of making the underlying datasets available even if only as a supplement to the printed product.

Many of the datasets produced by classics researchers are intended to be freely and widely available; the only proviso is that they are fairly referenced by users. Some datasets, however, are not published. Typically they refer to or describe objects that have not yet been academically attributed and cannot be made available to the public without the author's or project director's permission.

The notion of publishing in the traditional sense implies fixing a particular set of results or point of view at a certain point in the scholarly record. The publishing of datasets in classics is perceived to offer greater flexibility. For example, some key datasets are revised periodically, some as soon as new material is available and others on a scheduled basis irrespective of the amount of new material or revisions expected. Since the study of ancient history relies primarily on archaeological evidence, most of the databases are not expected to be static, while the classical philosophy databases will eventually be a completed entity once all the expected contributions have been received. Datasets that rely on analysis are usually published once all the data have been analysed; in a few cases a sample of data or discrete parts of datasets might be made available to other researchers at an early stage in the project.

Many classicists are prepared to countenance publishing their datasets themselves, but hold the view that universities might be minded to provide more help and support if this activity was explicitly recognised by the RAE to be valuable.

Ownership of data and constraints on publication and use

When it comes to the ownership of data, the relationship between the creators of datasets and their institutions and funders is not always clearly defined, reflecting the situation in many different subject areas. Where more than one researcher or institution is involved in data collection or processing the situation becomes even more obscure. One museum curator reported that the main dataset they look after belongs to the museum and university, but that some of the data are jointly owned with other museums. This can cause problems for managing access to the dataset for re-use.

As well as potential problems with intellectual property rights, dataset creators are aware of the pitfalls related to copyright law. Some of them actively avoid using material or data that may present problems – unpublished material for example. Others take a more relaxed view of copyright.

Response to requests for datasets

Academic specialties in classics are very narrow and the race to publish is not pronounced. The relatively small size of the discipline means that most of classics researchers know those working in their field and under most circumstances are content to share data. People typically wish to be acknowledged for sharing their data. Where there is a request for unpublished data there may be some negotiation, perhaps agreeing to co-author a publication, but it appears to be more common to wait until a dataset is made publicly available before requesting it or, for creators, before acceding to such a request. Once a dataset has been published normally people will be expected to access the datasets themselves.

Discovery, access and use of third party datasets

Discovering relevant datasets

By and large researchers in the classics can find most datasets they need in their field using web-based discovery tools, online bibliographic databases and their knowledge of the niche in which they specialise. The only grey area appears to be finding out about datasets being produced by new postgraduate students. That said, there are mixed views about the efficacy of the discovery tools available to them; while some are content with these others believe that a more standardised approach to metadata would improve the searching process. There are, inevitably in the view of a few, some datasets that are difficult to locate even if it is obvious from a published work that they exist. There is some expectation that Web 2.0-based tools will make the discovery process easier in due course. There is also a particular problem with some electronic data resources making them difficult to find: if the second edition of a hard copy volume is issued in electronic form, it cannot be catalogued. This means there is no trigger to alert libraries to the existence of these online resources since there is no separate ISBN. Although the British Library maintains a separate catalogue for electronic resources, if scholars are not aware of this catalogue they may not discover these particular resources.

Access to third party datasets

Researchers report few cost barriers other than the relatively low costs associated with downloading data. Occasionally access to classics datasets does have to be paid for but this

is thought to be unusual and is normally associated with the direct costs implicit in obtaining data objects that are owned by third parties, or where a private individual has made data available and wishes to recoup his direct costs.

Use of third party datasets

Researchers in the classics commonly want to use other people's data as building blocks for their own datasets. All researchers expect to acknowledge contributions from others and to be acknowledged themselves for data they have provided to others, but we received no reports of difficulties encountered in obtaining datasets from researchers in their field. Researchers would not publish unattributed material obtained from other researchers or museums without permission, though they do keep such data for private research.

Quality assurance

Quality assurance in the data creation process

Researchers tend to trust that the people working in their field produce datasets to a professional level of quality, though this varies a little according to the reputation of the creators of the datasets. The tradition of editing and proofreading data carefully before publishing them continues in the classics.

For those occasions where the provenance of a dataset is suspect, people might check the data for anomalies. There is an emphasis on trust primarily because it is not always made clear how datasets were generated. People would prefer to see quality assurance information such as analytical method, precision and sensitivity of the analysis published alongside the data.

Data management planning

Classicists who receive funding from the AHRC would normally follow the research council's guidance for data management planning and to offer their datasets to the AHDS for archiving[1]. Organisations like the AHDS have been highly valued by the classics community since they have been able to provide detailed advice about data management planning and archiving. Researchers applying for funding from the AHRC have become used to having to submit a Technical Appendix which has had the effect of focusing people's minds on issues central to effective data management planning. Of course not all classicists rely on funding from the research council and not all have sought advice from the AHDS. Some rely on their own knowledge or seek advice from their network of peers or their local university computing centre.

One of the key concerns that emerged through the course of this study is version management. Researchers want to know they are using the current version of a dataset but it is not always obvious that they are doing so.

Quality assessment of datasets

On the question of whether it is necessary or desirable for data to be subject to some form of external quality assessment process, classics researchers think this would increase trust but

[1] Detailed information on the work of the AHDS can be found here: http://www.ahds.ac.uk/about/index.htm

involve more work. Since they say that they already spend enough time reviewing the work of others they are unsure how it might work in practice. On balance people think that data quality was already adequately assessed through the process of applying for and obtaining a grant and the peer review of publications based that result from a research project. If formal external assessment of datasets were to become a requirement on the part of funders, though, most researchers would not object; in fact many think such a process might be helpful to them. Some, for example, had already received useful feedback from the AHRC. Informal feedback from users of people's datasets was thought to be potentially useful.

What motivates researchers to publish data?

On the whole classicists do not produce a great deal of data, certainly not by comparison with many other subject areas, nor do they publish very many datasets. This reflects the relatively small size of the discipline but it is also true that people often think that the datasets they do produce in support of their publications are too small to publish and that they are unlikely to be of use to other classics researchers.

In fact classics has an honourable tradition of database publication and has been practising electronic dissemination of data for nearly 30 years. Researchers tend to be very well-informed of the issues surrounding data publication. The classicists who do publish datasets do so willingly with the intention of sharing those data with others. There appears to be an open and collaborative culture of data sharing, and indeed an inclination to publish on local webservers – though many deposit backup versions of their datasets to the AHDS.

The nature of researchers in classics is different to researchers in some of the other subject areas we delved into as part of this study. There are fewer barriers to data sharing in terms of direct competition for funding or to be the first to publish on every small new advance in knowledge, and there is a prevailing culture of in favour of cooperation and data sharing. Researchers can study a subject in their particular niche for years and so really do have a love of the subject. In addition, the AHDS and other organisations have been on hand to guide and support people in issues central to data management and curation, and there has been a clear indication from several key funding bodies that they favour data sharing. The key factors which serve to motivate classics researchers to publish data are listed below:

- increasing scholarly reputation and reputation of institution
- establishing the identity of the institution
- to learn and teach others
- to reach a broad base of people
- love of the subject
- to enable dialogue between scholars
- job satisfaction and security, including positive feedback from peers
- possible career advancement

The classics researchers to whom we spoke always publicly credit the creators of the data they use in recognition of their value. As to whether there should be a standard method for citing datasets, most had never properly considered the subject though they appreciate that in due course citations to datasets may become a useful metric.

SOCIAL AND PUBLIC HEALTH SCIENCES

Overview

One of the aims in devising the list of eight subject areas to investigate was to include areas of study that produce and use data in particular ways within a particular data environment. A number of UK cohort studies and related longitudinal surveys conducted by academic bodies with government funding and by government departments exist in the public domain. These rich data resources support an active community of researchers that span the social and public health sciences. This community covers many disciplines in the social science arena including, for instance, epidemiologists, sociologists, psychologists, economists and geographers, so it is difficult to put a generic label on what they do. "Health & Development" is a label used by the UK Data Archive but is not widely recognised in the research community so, for the purpose of this study, we have called the community "Social and Public Health Sciences" but it is important to remember that the common feature of researchers in this community is that they contribute to and build upon cohort studies, examples of which are listed below:

- National Child Development Study
- 1970 British Cohort Study
- English Longitudinal Study of Ageing
- Longitudinal Study of Young People in England
- Millennium Cohort Study

The surveys conducted by government bodies are typically cross-sectional snapshots conducted at regular intervals using different samples of the population. Examples of these surveys are given in the list below:

- Health Survey for England
- National Diet and Nutrition Survey
- National Food Survey
- Expenditure and Food Survey
- General Household Survey

In addition to the cohort studies and national surveys, there is a network of nine regional NHS Public Health Observatories in England which monitor health patterns and trends.

The researchers covered by this section of the study fall into two groups. First, there are the researchers who create the data that are added to the cohort studies and national survey datasets on a periodic, regular basis. As well as creating the data, these researchers are also normally involved in the analysis and dissemination. Second, there is a larger group of researchers who are primarily consumers of the data contained in the cohort studies and national surveys datasets. These researchers extract value from the big, national datasets, often deriving new, added value datasets in the process.

There are a lot of researchers in the UK who use the cohort studies or national surveys. In 2005-6 the UK Data Archive – which hosts many of these datasets – delivered just under 50,000 datasets. UKDA saw 46,500 active users registered in the same year. The help desk

dealt with some 2300 queries in the same period, while over 56,000 students used various ESDS data in the classroom. For the purposes of this study we have spoken to a relatively small number of experts and data users in this field. Because the field is so broad, spanning many traditional disciplinary boundaries, we observe that the comments provided below are but general insights which contribute to our overall picture of researchers' attitudes to datasets.

Data creation

Form and variety of data produced

Given that this area is so broad, the form and variety of data produced are very varied, but many of the major cohort studies and national surveys are based on questionnaires. The national datasets can be highly complex and data verification, cleaning and documentation are important parts of what the creators of these datasets do.

Purposes of data generated as research output

For researchers whose work it is to create the data that populates the cohort studies and national surveys, the main focus is to produce high quality primary data. Normally, however, these same researchers also tend to do a lot of analysis on the longitudinal study data and, as with researchers working across most disciplines, producing high quality journal publications is important. Indeed, the quality and number of journal publications derived from particular cohort studies or national survey datasets are generally viewed as proxy indicators for the value of those data resources.

Researchers who rely on these national datasets as their primary source of data tend to focus on data analysis to answer a wide range of research questions. In the course of their analysis they might derive new variables and datasets, but these refined datasets are rarely destined for publications. Instead they form the basis from which publications are derived – normally in the form of scholarly papers for journals. Some researchers do produce their own primary data to supplement the data from national sources, but these new, often very tightly focused data, are rarely made available publicly.

Metadata

The main cohort studies and national surveys benefit from staff with data management expertise. Given the long term investment that has been made in long term data collection it is obviously important that the metadata are of high quality. This is assured not only by the expertise of the staff funded to produce and look after the datasets, but also by inspections, management reviews and funding reviews.

The datasets produced by researchers who use the national data resources are much less likely to have professional-level metadata associated with them. Awareness of the need for metadata is relatively low and, since in most cases the derived datasets are not destined for public availability, there is little incentive for researchers to invest time on metadata. It was observed that producing metadata is often one of the last things to be done in a project, and that there is no real legacy of good metadata. Of course an increasing number of funders are requiring award holders to produce a data management plan but, for small

scale, tightly-focused projects this does not yet appear to be yielding significant changes to researchers' behaviour with respect to metadata.

Adding value to data

The people charged with creating data for the cohort studies and national surveys are effectively paid to add value to data. It is collected then cleaned, verified, organised, documented and curated to a professional standard. These people tend to develop a close affinity with the datasets they help produce and take care of; they are protective of the value of the datasets.

Researchers who use these national datasets in theory add value as well since many of them derive new variables or analyse the datasets in ways that produce new datasets and new insights. Unfortunately it is rare for derived datasets, metadata or even the syntax people have developed to find their way back to the people looking after the original datasets. They are not even offered to the UK Data Archive or other data centres as a matter of course. The value added to the original data is manifested primarily in journal publications, but the derived datasets are not normally proactively shared with others.

Long term viability of datasets

The major longitudinal datasets described in the overview section are, of course, designed to exist and grow over a long period of time. These datasets are well managed and curated to a professional standard but in many cases long-term funding is not necessarily guaranteed. Many managers of these datasets need to re-apply for funding on a periodic basis and, to continue to receive funding, the research value of datasets will be taken into account.

Some research projects, particularly in the socio-medical field, do produce datasets that have been selected for curation by the UK Data Archive. They include, for example, *Urban Regeneration, Mental Health and Quality of Life in Wythenshawe, South Manchester, 1998-2001;* and *Childhood Vaccination: Science and Public Engagement in International Perspective, 2002-2004.* The Data Archive ensures that appropriate metadata and other necessary contextual information are collected and made available to researchers who wish to access these sorts of datasets in the future, thus helping to ensure their long term viability.

There are many datasets produced by individual researchers or small project teams that could have long term viability if they were offered to an appropriate data centre, but this tends not to be the natural course of things. The sharing of datasets from small scale research projects appears to be relatively uncommon at present.

Working with the created data

Tools and technologies for analysis

The tools and technologies used to analyse data are numerous and varied, reflecting the breadth of researchers' disciplinary backgrounds covered by the social and public health sciences umbrella. At a very general level, researchers in biomedical sciences tend to put lots of energy into collecting data well (which is one of the reasons they are often reluctant to share them) but the analysis is not always very sophisticated. Social science researchers,

on the other hand, are reported to place much more emphasis on the analysis of data. Typically researchers will spend months re-classifying data to derive new variables or they might add data from different sources in order to address new research questions. It would be unusual for the bespoke code written to enable this type of analysis to be made publicly available.

It is worth noting that although the national datasets are rich resources they are widely believed to be very difficult to use; this is why the dataset creators themselves tend to make the most intensive use of those data. Having an intimate knowledge of the characteristics of a dataset enables a tailored, sophisticated approach to data analysis.

"Publishing" datasets

As with all the other subject areas covered by this study, researchers' attitudes to making their datasets publicly available vary according to their personal perceptions, but these are guided by the prevailing culture within their subject area. On a macro scale, it is reported that researchers in the social sciences are generally more willing to publish and share their data than some of their counterparts in the biosciences. This situation is being addressed by the MRC which is continuing to develop their data management policy. Social science researchers have long been accustomed to data sharing mediated by the UK Data Archive.

Ownership of data and constraints on publication and use

In terms of the cohort studies the datasets normally belong to the funding organisation – the Economic and Social Research Council (ESRC) and the Medical Research Council fund a number of cohort studies, for example. Even though the main funders wish to encourage data sharing and would like researchers to be able to access national datasets, there are often many obstacles in the way, some of which are described in section c (ii) below. A recent study by Lowrance[2] indicated that there is a need to standardise current arrangements for facilitating access to national datasets and that there is scope to develop a model governance structure and guidelines on the policies and procedures adopted by the managers of important data collections. There are many good reasons for controlling access to national datasets – not least the need to protect their scientific value and reputation – but some researchers appear to believe that the desire to allow researchers reasonable access to important national data collections, and the desire of the dataset creators and owners to control access, are out of balance and that use of such resources in some instances is unreasonably constrained. This view does not apply to the UK Data Archive; researchers are satisfied with the access it offers to important national datasets.

Response to requests for datasets

Researchers may request access to data from cohort studies or other national sources and will sometimes find themselves in a negotiation. Their application for access is likely to be evaluated by the data owners or curators and, if access is granted, such access will typically be governed by written agreements or licences. Access to datasets via a data centre such as the UK Data Archive may be rather more straightforward, but users are required to adhere to conditions where these are required by the data creators or owners.

[2] Lowrance, W (2006) *Access to Collections of Data and Materials for Health Research.* A report to the Medical Research Council and the Wellcome Trust, March 2006

In terms of other datasets – those created from small scale projects – our interviewees believe that data sharing occurs only on a limited scale.

One of the reasons funders are keen to encourage data sharing is to reduce researchers' duplication of effort in terms of data collection. When it comes to analysis of that data, it has been observed that the UK Data Archive does not, as a matter of course, tell researchers that the dataset to which they have requested access has already been accessed by a different researcher (or group of researchers). However it does display which other users are using particular datasets and for what purpose, providing the user agrees to display this information about themselves and their project.

Discovery, access and use of third party datasets

Discovering relevant datasets

Since a good deal of the work in the area of social and public health sciences revolves around working with the data from the national dataset collections, relevant datasets are normally well known to researchers. Other means of finding relevant datasets include references from journal articles, journals or other publications that focus on issues to do with these big datasets, searches of data centres and general web-based searches.

Access to third party datasets

On the face of it researchers in social and public health sciences have a wealth of primary datasets available to them but, in practice, gaining access to such data is easier said than done. There are several obstacles to which our interviewees commonly referred:

- In this field much of the primary data are gathered from or about individuals and their right to confidentiality must be respected. In tandem, the terms of the Data Protection Act must be complied with. Data creators are often particularly concerned that data from or about individuals must not be able to be identified. Concern about the fallibility of some anonymisation techniques leads data managers to err on the side of caution, often leading to tight restrictions on access to datasets.
- Some datasets contain medical records which data creators are particularly keen to keep shielded for reasons of confidentiality.
- Where data creators deem datasets to be particularly sensitive, they may impose restrictions on access. For example, researchers may be allowed access to a dataset via one controlled access point, possibly under observation. Taking this one step further, in some cases researchers may not be allowed personal access to datasets, but they are permitted to send queries or syntax routines which the data owners will deal with.
- Cohort studies typically follow individual participants through part of or their entire life course. It is important from the point of view of completeness of the datasets that people continue to participate in the studies. This imperative naturally gives rise to a cautious approach to data sharing.
- As with most areas of research, there is competition between researchers to produce the best work in the best journals. This has a bearing not only on people's personal career path, but also on the reputation of the dataset. People involved in creating that major datasets naturally wish to be involved with extracting the best value from that data. To

this end there is often an embargo period where the data creators enjoy exclusive access to the data.

- It is also true that the people involved with creating such datasets are best positioned to analyse the data, chiefly because they are so intimately acquainted with them. This can lead to a feeling that letting third parties have unfettered access to the data risks those data being misrepresented, no matter how good the quality of the metadata supplied with it.
- It is often said that the data in the cohort studies is complex and can be very difficult to use on conceptual and practical levels. This is important when set in the context of the commonly held view that the UK has a deficit of quantitative analytical skills and training capacity to fill this gap is limited.
- Creating longitudinal datasets is an expensive business and therefore the people responsible for them tend to feel the need to protect them. This is manifested in reported anxiety about commercial organisations using data, deriving slightly or materially different datasets and claiming intellectual property rights over these new datasets. The fear is that such practice may adversely affect or somehow limit the scope for genuine scholarly research, and so managers tend to behave conservatively with respect to sharing data.
- Of course rights to data can be protected to a certain extent with licenses or other data sharing agreements. It is common practice for researchers to have to sign a licence or similar agreement before being allowed access to datasets. The terms of such documents may be deemed too restrictive or awkward.
- Major datasets may incorporate data elements which are themselves subject to licensing restrictions which may limit the extent to which they may be shared with third parties.
- Researchers might only be allowed access to a dataset on condition that one or more of the data creators is cited as a co-author on any publications that result. In many cases this may be a useful symbiotic arrangement but it is not clear that all researchers wish their independence to be restricted in this fashion.
- Some datasets reside behind toll barriers, barriers which may prove prohibitively expensive for some research projects. An example often given is the Avon Longitudinal Study of Parents and Children (ALSPAC) where external researchers are required to contribute towards the costs of administering the study.[3]

Use of third party datasets

Researchers working outside the institutions responsible for collecting data for the national datasets tend to make extensive use of third party datasets. Some researchers perceive a moral obligation to extract as much value as possible from these resources, which include the national cohort studies, national survey datasets, datasets produced by regional authorities (such as Primary Care Trusts) and those produced at a local level.

Quality assurance

Quality assurance in the data creation process

[3] Details of how ALSPAC may be accessed are provided in this management policy document: http://www.alspac.bris.ac.uk/collab/ALSPAC-policy.pdf

Those responsible for looking after national longitudinal datasets take quality assurance very seriously: the reputation and value of the datasets depend upon it. Funders expect that data creators follow relevant professional guidelines to produce high quality data that exhibit consistency and integrity. Researchers working on smaller scale projects which create data may not have data management expertise close at hand, but our interviews leave little doubt that they take professional pride in ensuring the data they create are as good as possible. Researchers take steps such as checking the distribution of variables and testing to ensure data are random. Experience is often cited an important part of the quality assurance process.

Data management planning

Data management planning is an integral part of what national dataset creators do, but for the majority of researchers it has hitherto not played a central role in their thinking. This attitude is slowly changing in response to the evolving requirements of research councils – though the pace of change is said to be glacial. Funders like the MRC now require award holders to consider data management planning as part of their overall project plans. The MRC has mechanisms in place to advise researchers about data management planning but sanctions are envisaged to improve the rate of compliance with award conditions with respect to data management planning. Compliance is easier to manage in the case of big projects, where adherence to data management plans including data sharing can be assessed during interim reviews, with continuing funding being contingent upon satisfactory compliance.

Quality assessment of datasets

Researchers tend to trust the quality of the third party datasets they use; to do otherwise would lead to time consuming checking procedures and could erode relationships with collaborators at an early stage in a project. Informal quality checks include things such as the track record of the dataset creator and the rigour of papers that have been based upon the dataset. The exception appears to be datasets created in countries with less than optimal track records for scholarly training and research. There are formal mechanisms in place to assess the quality of major national datasets, but for the plethora of smaller scale, more focused datasets there is no general agreement on whether they should be subject to some form of external assessment. It might be a "good thing" but finding reviewers with the time and particular subject expertise is widely thought to be too much to expect.

What motivates researchers to publish data?

Push factors: the effect of policy

In this sector of scholarly enquiry we found scant evidence of researchers wanting to publish datasets. Typically researchers will request data from one or more publicly-available datasets – a process that is often straightforward though sometimes difficult or impossible – and they will undertake data analysis. Often this process leads to the creation of new, derived datasets but these tend not to find their way to the public domain. The key funding organisations in this field, such as the ESRC and the MRC, desire data sharing and there are early signs that researchers are beginning to respond to the research councils' data management policies. Some of the researchers to whom we spoke are willing to share their datasets, but they are rarely asked for them and so they reside on a local server or disk

where they cannot easily be discovered. Indeed, the reuse of qualitative datasets is reported to be very rare. The goal is for funders' policies to change the overall culture in relation to publishing datasets, but the general view is that this may be a long process.

Pull factors: intrinsic rewards

Unlike in some of the other areas we have looked at there are no obvious rewards that accrue to researchers who decide to make their datasets publicly-available – though few deny that sharing datasets produced with public funds is a worthwhile principle. In fact it appears that one's dataset needs to be nationally recognised as being an important resource before it is submitted as evidence to the RAE. With the bar set this high, researchers producing small scale datasets see no reason to invest the time and effort required to make their datasets publicly available. This sector of research does have the significant benefit of world class data centres – notably the UK Data Archive – to which they may submit their datasets for archiving though, in common with researchers in other subject areas, clearly many believe their particular dataset will be of little use to others. Besides which, some want to control their data, limit the possibility of the data being misrepresented, and limit the scope for competition.

ASTRONOMY

Overview

UK astronomers are of two main types – observational astronomers or theoretical astronomers – and naturally only the former generate primary research data. Primary research data can be of three types, images, spectra from individual objects (brightness against individual wavelengths), or light curves (brightness against time). Theoretical astronomers utilise the data generated by observations and produce secondary data themselves.

Observations can take the form of sky surveys (whole sky or part of the sky) or single object studies. Projects can still be the domain of single astronomers, though this is less usual nowadays and most projects are worked in by groups of various sizes. Observations are made using either ground-based facilities (the big telescopes) or orbiting observatories such as the Hubble Space Telescope[4] and Spitzer[5]. Astronomers lease time on these instruments, collect all the data that they can in that time and then begin work on processing and analysing them. These large facilities – which may be outside the UK – archive the data and permit them to be worked upon by the astronomers who collected them for a period of twelve months. After this time they are opened up for access by all. This is a community-accepted convention. It allows the original data collectors to have time to carry out their own analyses and science and to publish their findings – or at least get their publications in preparation – before the data are made publicly-accessible. Thereafter, anyone can use the data for their own work.

Most astronomical research in the UK is now funded by the Science & Technology Facilities Council, formed by a merger of the CCLRC and PPARC (Particle Physics & Astronomy Research Council). The latter previously funded the work in this field. Policies on data have yet to be developed by the STFC but there is an expectation that its grantholders will make their data publicly-available at some point. The research council as yet offers no guidelines on how this should take place.

The sharing of data is, however, well-advanced and formalised in astronomy in general, whatever the explicit policies of funding bodies. The national facilities (telescopes) make data available, once the twelve-month embargo period has expired, to the UK data centres, which store and curate the data. These data centres, expert in manipulating and looking after data in the longer term, are at the big facilities (e.g. Jodrell Bank, belonging to Manchester University) or in specific university departments of astronomy – Edinburgh, Cambridge and Leicester. They seek to collect and curate as much of the data output of astronomers as is practically possible. There is data centre provision in other places, too, where astronomers may find and access data they need for their work, in Strasbourg, at the European Space Agency and the European Southern Observatory, located in Chile.

Data centres collect astronomical data from various sources, not just from the telescopes in their own country or region. For example, the University of Strasbourg's CDS (Centre de

[4] http://www.stsci.edu/hst/
[5] http://www.spitzer.caltech.edu/spitzer/index.shtml

Données astronomiques) database called SIMBAD[6] collects data on individual objects, harvesting articles on these objects from the world's astronomical literature as well as curating datasets relating to the objects. The same centre has other important databases: VizieR[7], which is a collection of catalogues on stars and Aladin[8], a sky atlas that permits users to interactively access data from multiple sources related to the focus of interest. NASA has databases comprising not only primary observational data but also full-text journal articles (the ADS – Astronomical Data System[9]), software and analytical tools (see, for example, HEASARC,the High Energy Astrophysics Science Archive Research Center[10]).

There is now a lot of effort going into moving to a higher order of data handling in astronomy with the development of the Virtual Observatory (VO). This aims to integrate astronomical data in one place and to provide a workspace for astronomers who wish to use the data. The International Virtual Observatory Alliance[11], the partnership of participating projects, now comprises 16 projects from countries worldwide. The IVOA has a number of working groups dedicated to the development of standards. The UK's VO project is called Astrogrid[12], which involves the collaboration of the main UK astronomy data centres and is led by Professor Andrew Lawrence in Edinburgh.

Data creation

Form and variety of data produced

Primary astronomical data are in the form of images, spectra and light curves. They are collected from optical, infrared, gamma-ray and X-ray telescopes. Astronomers using national facilities to collect data do not necessarily have to visit the facility itself. They book time – usually 'a few' kiloseconds (i.e. about twenty minutes) on the instrument and specify what they wish to observe and the data are collected by the instrument and archived at the facility. The astronomer then accesses his/her data and has exclusive use of them for twelve months before the facility opens them up for public use. Theoretical astronomers use supercomputers to work with the datasets of observational astronomers. Special software is required to produce visualisations of theoretical astronomy data.

Data from space-based and ground-based facilities vary: space missions are very organised with their data provided to the community in online databases, while astronomers involved in ground-based work use a telescope for a certain number of nights, write the data to CD and take them back to their laboratory for analysis.

Survey astronomy, a very fashionable area, involves the collection of data from a large area of sky and then people look for something interesting there – a data-mining exercise.

[6] http://simbad.u-strasbg.fr/simbad/
[7] http://vizier.u-strasbg.fr/
[8] http://aladin.u-strasbg.fr/aladin.gml
[9] http://www.adsabs.harvard.edu/
[10] http://heasarc.gsfc.nasa.gov/
[11] http://www.ivoa.net/
[12] http://www2.astrogrid.org/

Metadata and file formats

There are community-agreed standards for file formats and storing data. The most commonly (almost ubiquitously) used file format is FITS (Flexible Image Transport System). It is endorsed by NASA, which provides a large package of software tools for analysing FITS datasets. Astronomers either use the data in the FITS format or convert them to ASCII. The full metadata are included in the FITS file. There are a number of fixed metadata fields that must be completed, plus a set of free-text fields to give additional contextual information. The exact format of the FITS file metadata can vary from facility to facility, though. For example, the Edinburgh Data Centre loads data from the Cambridge archive, including the Cambridge-assigned metadata, but then enhances them by adding further metadata details.

The Virtual Observatory is attempting to develop standards in this area. This is considered to be a good thing, but there is some doubt as to whether it will prove possible or, if it does, how to achieve conformance or compliance. One new standard is the VO-Table, an XML-based standard for tabular datasets. It is still early days for the VO's work in this area so the possibility of any scalability issues for XML remain to be seen. Standard software tools are used to carry metadata over to derived datasets, and to add additional metadata, thus adding value to the data.

Adding value to data

The data that are collected from telescopes are in the rawest form and are processed into more usable data which are made freely available after twelve months. These datasets are then reduced further into products (spectra, images, etc) and the scientific analysis is performed on the data at this stage. Reduction is mainly carried out in the large data centres. For example, data from the European Space Agency (ESA) are shipped to the Leicester Data Centre for reduction and analysis. Nevertheless, there is the view that data could be supplied in more user-friendly formats than huge binary files.

Sky survey data are collected from optical and infrared instruments and shipped to the Cambridge data centre where the raw data are curated. Reduced data are sent to the Edinburgh data centre for further processing and preservation. Radio astronomy data are generally curated at Jodrell Bank.

Long term viability of datasets

Astronomers have always had a longer term perspective on data collection and handling than most disciplines, forced upon them by the fact that a telescope takes 15 years to build and is expected to provide data for decades afterwards. Despite this, the long term storage of datasets is not always optimal, even by data centres. Many raw data have been effectively lost as they are stored on reels of tape. Currently, data are stored on spinning disks and backed up at least once. Storage costs decrease year on year but considerable effort is needed for conversion of data from raw to reduced form and so there is a substantial cost to this work. Curation of data is partly addressed by the big facilities such as NASA and ESA who guarantee to look after data in the long term and are funded to do so. Nevertheless, there is a view in the astronomy community that when budget cuts are operating, it is more attractive to launch a new rocket than to put money into data handling.

The UK data centres maintain their own archives of data collected at other facilities, even though this is effectively simply mirroring something that is assumed to be securely curated

elsewhere. Partly there is a safety-first attitude in this activity but there are also advantages to the data centres in having the data directly under their care. The UK funding bodies, understandably, do not necessarily see great value in this though the STFC explicitly supports the curation of recent and new data. Nonetheless, such funding is usually for a single project at a time with tapering funding for data preservation after the project or mission finishes. Although datasets can become less important over time, it is never possible to predict with accuracy which, how or when and funders are loathe to pay over the very long term for data to be preserved. It is not easy to get funding for core R&D work, such as keeping up to date with data technologies, however. Every grant proposal needs to be linked to new work rather than this sort of activity. As regards legacy data, the data centres fund legacy data curation activities by shunting money between facilities and consumables budgets as appropriate to support all their activities in the best way possible. The marginal cost of storing older data on disk is not particularly high and so it is possible to cover these small costs from new project money. In the future, because of the e-science aspects of astronomical research, it is hoped the EPSRC will also be involved in funding the data centres.

Individual astronomers or groups curate derived data on their own computers locally. There is rarely any budget line in grant proposals for looking after such data in the longer term, though there is usually an element of every grant award that is allocated to computing in general. Data archiving activity is supported while a space mission is actually operating but it is difficult to look after the data once the mission is completed. There is a view that there should be a better balance between the funding of new projects and enabling the exploitation of older data. Some events happen only once and so images that are 20 years old remain valuable. Alternatively, some things happen over decades, so to carry out an analysis of such an event requires that older data are looked after.

While individual astronomers tend to keep their data on their own computer and look after them with a greater or lesser degree of expertise themselves, the data centres naturally bring professionalism to the task. A mix of skills is represented in the staff complement of the data centres though historically they have all been astronomers first – people who have built up an extra level of expertise in data management over time. Nowadays the centres are curating tens of terabytes of data and need to find specific expertise for this. Hiring from industry is expensive, but so is training on the job. Archive staff also have to keep up with what is happening in astronomy itself so they need to be given research time, too. Reviewers for grant panels do not particularly like this idea of people mixing data management work with basic research, but the view in the centres is that people looking after important data should understand how observers collected the data and what they may wish to do with them in future. This is especially important in the context of the periodic data recalibration exercises.

There is a further concern about data in astronomy: the long term *accessibility* of observational data. Accessibility depends upon having porting software that enables the moving of data onto modern computers on which scientists can access and analyse the datasets. The UK Starlink Project[13] provided this until recently when its funding from PPARC ceased. Starlink provided and coordinated interactive data reduction and analysis facilities, with Starlink's own tools being supplemented by such packages as IRAF and IDL.

[13] http://www.starlink.rl.ac.uk/index.htm

It also encouraged software sharing and standardisation to prevent unnecessary duplication of effort. There is no entity in the UK that is responsible for maintaining the Starlink software any longer. In Hawaii, the Joint Astronomical Centre (JAC) continues to support Starlink (funded in part by the STFC) at the moment because of its own need for the software, but it is expected to wind down this operation in future. The standard package for reducing and analysing optical astronomy data, IRAF (Image Reduction and Analysis Facility[14]), is supported and maintained in the US by an optical astronomy observatory run by a consortium of universities and funded by the National Science Foundation.

Working with the created data

Tools and technologies for analysis

FITS is a fairly elderly format, developed in the 1970s, but is being brought up to date at the moment. It is a flexible format so that there is sometimes a problem in accessing other researchers' datasets. Astronomers therefore write a lot of software themselves for this purpose and for data analysis. Software is usually written in FORTRAN, and all astronomers appear to view this as part and parcel of their everyday work. They generally write this software themselves because it produces a better result than trying to specify what is needed to a programming specialist. Sometimes, however, computer scientists may be employed on larger teams to assist with data manipulation and processing. Astronomers learn their programming skills as part of their training and so despite this seemingly *ad hoc* approach to accessing data, the system appears to work well in practice. Most astronomers work on data on their own local computer facilities.

Most astronomy research is standardised on UNIX-type computing systems (e.g. Linux). Windows is regarded as 'a step too far', though most astronomers are quite used to working on a number of operating systems and can cope with that without difficulty.

"Publishing" datasets

Astronomers do, in general, publish data wherever possible. Funders have an expectation that data collected in the course of publicly-funded work should be made available to the research community and the large facilities do this on behalf of astronomers (after the twelve month period has elapsed) with regard to raw or reduced data. Derived data, those produced by the astronomers themselves after processing and analysing raw or reduced data, are usually published as part of a journal article or supplied as supplementary data along with an article. Not all derived data thus end up in the public domain, therefore, but astronomers almost always supply data if requested by another investigator.

The data centres in the UK make as much data as possible openly available. As well as collecting and looking after data from UK observatories they will collect data from other sources too if possible. The Edinburgh data centre, for example, has made public the sky survey data from a closed-down Australian facility. In the context of the next section of this report, STFC considers that it owns these data, even though all the data conversion work on the old Australian datasets was carried out in Edinburgh by university-employed staff.

[14] http://iraf.noao.edu/

Ownership of data and constraints on publication and use

Data collected at the national, publicly-funded facilities are considered to be 'owned' by the observer (or their funder) for the twelve months after collection and then are publicly available. Some larger universities, particularly in the US, have their own observatories and may regard data collected as their own; in such cases those data do not become public unless an observer specifically makes them so. Some agencies – the Max Planck was specifically mentioned in this regard – view the raw data as their own, but make the reduced data publicly available. Small projects may keep their data private.

Response to requests for datasets

Requests for datasets made privately to the dataset creator are virtually always responded to positively. Frequently the reported findings from an observation are made public before the datasets are (i.e. before twelve months) and other investigators wish to access the datasets at that time. Researchers do give thought to how early they will share their data and how much they will give away in this early period, but essentially all publicly-funded data become freely available within short time after collection. Individual astronomers are often added to the list of authors of a paper if their data are used in the scientific process: sharing datasets can thus have a definite and positive outcome rather than just being an acceptable way of behaving.

Discovery, access and use of third party datasets

Discovering relevant datasets

Every observation is assigned an ID number and there are good methods for referencing an individual object in the sky, so given that metadata standards are quite high in this discipline, getting to a wanted dataset is quite simple. In journal articles datasets can be referenced by their accession number in the archive where they reside and if the data were collected on a specific mission then that piece of information is also declared. As a result of these conditions, *locating* datasets that support published work is not especially troublesome. What is more difficult is accessing some types of dataset (see below).

Access to third party datasets

Whilst observational astronomers' raw and curated datasets become openly available within a short time after collection, there is more of a problem with the accessibility of derived datasets. Theoretical astronomers work with these and process and manipulate derived data into new data products which they themselves publish along with their findings and conclusions. In general, there is a culture of sharing in the astronomy community and it is rare to find someone who keeps their data completely private (although as we have noted above *some* facilities or agencies can take this line). There is a difference, however, between actively disseminating data and providing it when asked. Some derived datasets – like those from small observational astronomy projects – are published in an *ad hoc* way on project websites (that is, unattached to a journal article), some are published with the article, but sometimes they may not be actively disseminated at all. In these cases, getting access to them is achieved by asking the data creator personally for the dataset. Whilst this is almost always provided it is often expected that the data creator's name will be put on the article that ensues. There is, therefore, a trading system going on with data being bartered for the

chance of formal journal publication. (This is not unique to astronomy: in the life sciences, reagents are often traded in this way).

Considerable effort therefore has to go into locating and obtaining datasets that are not formally disseminated. Although it is usually profitably spent, in that the dataset is eventually accessed, this may not always be the case. The astronomers we spoke to who use derived data for their work would prefer a system where data are always disseminated and they expressed a preference for using the arXiv as the locus. The message was that if all articles deposited on the arXiv were accompanied by their supporting datasets, including code, life for theoretical astronomers would be considerably easier.

Use of third party datasets

There are no significant issues around re-use of data except for the requirements around software to access others' datasets s discussed above. The culture of sharing data is strong in this community and with that comes the expectation that datasets will be re-used by others.

Quality assurance

Quality assurance in the data creation process

Most astronomy laboratories or groups carry out some sorts of high-level quality assurance on the data being processed by them, or before publishing. Quality control measures can be applied at two main stages in the data cycle in astronomy. At the data reduction stage, the data centres undertake data checking procedures. One example may serve to illustrate what is involved: at the Leicester data centre every image that is collected or harvested is checked for quality and consistency. Several images are checked per day. It is a difficult but rather tedious procedure and is shared out amongst a body of volunteers who undertake a training lasting a few days. When a new version of the software being used comes out, there is a need to go through all the stored images again. The new venture that uses the general public to differentiate between spiral and non-spiral galaxies – Galaxy Zoo – is an experimental way of sharing workload[15]. In this case, though, workers are not applying quality control procedures, but are helping to sort data. The Edinburgh Data Centre recalibrates its data from time to time as the understanding of how the telescope used to collect the data improves. Recalibration enables the team to spot lower level artefacts and apply corrections across the entire archive, changing all the data values in the process. This takes a lot of time and storage space.

Data management planning

The STFC does not appear to insist on a formal data plan in grant proposals but, in explaining the intent of a project, applicants write a section in their proposals on what types of data will be collected and what they plan to do with them. This does not usually include a description of plans for long term curation of data in the case of smaller projects. For bigger projects (e.g. 5-10 people employed, with a duration of more than 5 years) a data plan is required and such projects do make a play of returning data to the community because this is seen as likely to increase their chances of funding.

[15] http://galaxyzoo.org/

Quality assessment of datasets

The other time in the data cycle when quality control comes into play is during the publication process, when derived data are submitted along with an article to a journal. It is at this point that peer review is expected to play some role in controlling the integrity and accuracy of data. In practice, peer reviewers for journal articles cannot, and therefore do not, check datasets in the vast majority of cases. The datasets are too large and complex for this type of reviewing to handle. Reviewers read the article, assure themselves that the work was carried out properly and that the conclusions drawn are valid ones given the supporting data that the authors say they have, and pass the article for publication. If they spot something in the results section of the article that seems inconsistent or wrong then the reviewer may inspect the data or question the way the author(s) have reduced the data, but this is fairly rare. Under such a system it is possible for flawed datasets to be published, but the system is largely self-correcting in that errors are soon discovered by other scientists using the datasets and revealed to the author(s) and the community. The bulk of astronomical research is performed using very standardised procedures and in cases of non-standard procedures being employed, referees are likely to take a closer look at the methods used and the data produced. This may take place in fewer than 1 in 10 cases. In effect, then, quality assurance is effected by the community. This is aided by the fact that the standard software packages used provide the checks and balances needed to ensure that faulty datasets are discovered. As a result, astronomers generally trust one another's datasets and untrustworthy data, though arising occasionally, are not a big issue in this field.

What motivates researchers to publish data?

Push factors: the effect of policy

In the UK, the main funding agency for astronomical research, the Science & Technology Funding Council, has no explicit policy as yet covering the public availability of data. Its predecessor, PPARC, had a policy of encouragement without requirement. STFC does have a firm policy covering other research outputs (journal articles), so we can probably expect a policy on data to be developed at some point.

Pull factors: rewards

Journal articles are viewed as the primary output of an astronomer's efforts because it is through these that credit accrues. There is no overt recognition for 'publishing' (in the sense of placing in the public domain) datasets but astronomers acknowledge that there is an implicit reward in that people who make their data readily available enjoy a reputation within the community for being 'good guys'. There is also the possibility that someone using the dataset will wish to publish an article using it and frequently the form is that the dataset creator is included as an author. The fear of not being permitted time on a facility in future if they 'sit on' data that they have collected from that facility is another factor in encouraging the sharing of datasets. There is also the hope of driving users to the article that discusses the findings and conclusions drawn from using the data, and thus the hope of citations in the longer term. Impact is measured by citations and publishing in what are judged to be 'high-impact' journals. Much of this attitude is driven by the RAE exercise which explicitly rewards such things.

In the case of data centres there is another measure that comes into play: funding can in part be based upon usage of the data they curate.

A contrast was drawn by some astronomers we spoke to between the situation in the US, where money is given for collecting data, and that in the UK where money is given for doing the science using the data.

CHEMICAL CRYSTALLOGRAPHY

Overview

Crystallography is the most-used unambiguous technique for identifying the structures of chemicals. This structural characterisation is done either for basic research purposes or as a service for other chemists who do not have their own crystallographic facility.

The chemical crystallography community has long had a relatively organised approach to data. Crystallographic data are highly-structured. Any heterogeneity resides in the instrumentation and the software this uses rather than in the data outputs. Outputs are in 3 or 4 possible formats but there is a *de facto* standard adopted by the bulk of the community, the CIF (Crystallographic Information File), and a standard protocol for making data available to others. Probably 95+% of crystallography data is in the form of CIF files.

A CIF file represents derived data rather than raw data and chemists may derive further data from the CIF. Various players have made software available for generating and manipulating CIF files, including the Cambridge Crystallography Data Centre (CCDC) (see below for more on this organisation). Examples of such software are Mercury (from the CCDC)[16], WinGX from Glasgow University[17] and Olex, developed at Durham University[18]. All of them are free to the community. Additionally, researchers often write their own software to perform particular tasks. There have been people capable of writing software for crystallographic purposes in every major crystallography research group since the 1960s and frequently in smaller outfits everyone is a one-man-band, able to perform all roles in the data production process.

The major public funder of chemical crystallographic research in the UK is the Engineering and Physical Sciences Research Council (EPSRC), but a considerable amount of other crystallography work carried out in universities is paid for by pharmaceutical or chemical companies. This of course has some repercussions on what can be made publicly-available for sharing, but in general if a project is a purely academic one, even if funded by industry, then the funder places no barriers to public revelation of the data outputs. At the other end of the scale, some industry funded research is completely client-confidential, with the university crystallographic centres providing a bespoke service accordingly.

The EPSRC itself has no policy on data sharing. One interviewee expressed strong views about this, and argued for a planned approach to data sharing and preservation, citing an environment where employers (universities) are 'starting to raise eyebrows' about the EPSRC's hands-off approach to the importance of a planned approach to data sharing, curation and preservation. This is a view that is not out of place, perhaps, in an arena where Microsoft is investing in the development of ORE (Object Re-use and Exchange) in chemistry using OAI-compliant repositories and institutions themselves are coming under greater pressure to provide facilities for data storage and preservation. This is expensive, even in relation to the cost of the original research, but it is acknowledged by those who work in the field as a nettle that must be grasped. Some feel that the EPSRC's stance

[16] http://www.ccdc.cam.ac.uk/products/mercury/
[17] http://www.chem.gla.ac.uk/~louis/software/
[18] http://www.dimas.dur.ac.uk/olex

discourages researchers from including a proper data plan in grant applications for fear of making them look over-costly and thus damaging their chances of success.

Crystallographic data are often shared where there are no commercial interests involved, though estimates have been made that suggest that more than half the data generated on crystals are not openly published for various reasons[19]. There is an established, centralised system for publishing crystallographic data, and the main repository for chemical crystallographic data in the UK is the Cambridge Crystallographic Data Centre (CCDC). The CCDC was established in 1965, funded by the then Science Research Council, which subsequently metamorphosed into the Science and Engineering Research Council and latterly into the Engineering and Physical Sciences Research Council (EPSRC). The CCDC was funded by public money via the research councils until 1987 when it became a non-profit independent organisation. It was enabled to do this because its usefulness was 'discovered' by the chemical industry in the 1980s and that industry was willing to pay for the data it needed. Currently, there are around 130 industrial customers (subscribers) paying an average of 25,000 USD per year. Academic institutions pay a fraction of this – about 500-1500 GBP per year for unlimited access – and users from developing countries enjoy a reduced charge or a waiver. The CCDC's business is sound at present. It has running costs of over 2 million GBP per annum, employing over 50 fulltime staff to produce the database to the required professional standards, and these costs are covered by subscriptions and software sales.

Academic users of the CCDC database are largely structural chemists interested in the size and shape of molecules. Industrial users have traditionally been mainly drug design chemists but more recently there has been a lot of interest from drug formulation chemists working on molecules that have already passed through clinical trials.

The other main databanks where crystallographic data are deposited are in Germany (an inorganic molecules database), in Canada (a metals database) and the Protein Data Bank (PDB) in the US. The former two are subscription-based services; the PDB is funded entirely by US public money via the NSF and NIH and is completely committed to a free-access-for-all formula (even for industrial users).

In recent years there has been increased pressure from within the crystallography research community for more open access to crystallographic data. Two projects exemplify this development. The first is e-Crystals, an open access database of crystal structures from the Southampton Chemical Crystallography Group and the EPSRC-funded National Crystallographic Service (in Southampton)[20]. The second is CrystalEye, a service produced by the Unilever Centre for Molecular Informatics at the University of Cambridge[21], which harvests crystal data from the web and provides tools for browsing and manipulating them. Other examples of such services are ReciprocalNet[22] and the Crystallography Open Database[23]. Inevitably, there is concern that the growing success of these services will have

[19] Allen F (2004) High-throughput crystallography: the challenge of publishing, storing and using the results. *Crystallography Reviews,* **10 (1)** pp 3-15 DOI: 10.1080/08893110410001664864
[20] http://wiki.ecrystals.chem.soton.ac.uk/
[21] http://wwmm.ch.cam.ac.uk/crystaleye/
[22] http://www.reciprocalnet.org/
[23] http://sdpd.univ-lemans.fr/cod/

implications for the CCDC's business, since free access to CIFs may diminish drastically the subscription base, provided largely, of course, by the pharmaceutical industry. There is a move, therefore, to promote open data in crystallography as a complement to the CCDC's collection rather than as a competitive threat. The aim appears to be to organise alongside the CCDC and to collect as much open access data as possible to add to the CCDC database and, indeed, the CCDC harvests data from e-Crystals to further this aim. It is in the whole community's interest to get as much content into the CCDC database as possible because, after 30 years, it has now reached a size where scientific work can be done on the whole collection: new tools and technologies allow data to be interrogated, structural trends to be deduced and predictive methodologies employed.

Data creation

Form and variety of data produced

There is almost complete homogeneity of form for chemical crystallographic data in the public domain. They are produced in raw form from the analytical machines but converted by proprietary (machine) software to CIF files which are the ones that are generally made available to others. CIF files are thus derived data. There is some discussion around the desirability of being able to access raw data files for mining and quality control purposes, but such access is not a norm in this community. The raw data formats vary according to machine but all machines process these data into the standard CIF.

It is not currently possible to drill down through a CIF file to get to the data in their rawest format, yet this would be useful for processing and analytical purposes. The best *de facto* standard for the raw data at the moment is a plain ASCII file. Through CrystalGrid there have been moves towards adopting a new standard for raw data, the advantage of using these being that they can be shared without the need to be tied to proprietary software for processing them to a CIF. A new standard is being developed in the US, called IMG CIF, which will allow this raw data exchange. Once developed, there will be an approach to the instrument manufacturers to persuade them to conform to this.

Metadata

CIF files are self-describing, containing information about the instruments used to produce the data, the investigators who did the work and so forth. They are highly consistent. The eCrystals project has been working on the issue of more detailed metadata for crystallographic data files but this is in the development stage at present.

Adding value to data

The CCDC adds considerable value to CIF files in its care. It curates around 35,000 datasets a year, validates them, adds a search-enabling chemical diagram and additional metadata. The CCDC provides sophisticated search capabilities that enable users to carry out sub-structure searching by drawing a part of a molecule and then searching for that within the whole database.

Long term viability of datasets

The big databanks like the CCDC have professional standards in place to ensure proper curation and preservation of their digital files and are currently operating sound businesses, even though their business models may not be ideal. Other, smaller, databanks are vulnerable. One example is the Chemical Database Service at the Daresbury Laboratory, which holds some crystallography data amongst much other chemical information. Chemists like it because of its fast link. This is one of the EPSRC's National Services, but in October 2006 the EPSRC announced its decision to close the database, despite much protest and argument from the user community, including from crystallographers who use it because it was freely available online to UK academic users. The decision has been reversed, but only for crystallography content (all the other areas of chemistry have been lost) but the process has reinforced the impression in the chemistry community that the EPSRC has a low level of interest in data and their importance.

Larger crystallographic laboratories and services store all derived and raw data as a matter of good practice, but this may not always apply to small groups or single researchers. In smaller groups, it may sometimes be the case that CIFs are retained but raw data discarded. The problem with this approach is that the experiment cannot easily be replicated, since the original samples, which were synthesised in a highly specialised lab by a specifically skilled researcher, who due to the short term nature of academic research contracts may well have left the post, were not retained. Raw data, therefore, become extremely valuable in these circumstances.

Working with the created data

Tools and technologies for analysis

In many – perhaps most – cases, the text-based CIF file is adequate for the purpose intended, since it contains all the information needed to understand the basic structure of a chemical. Nonetheless, access to and use of raw data could be useful and demand for this is likely to become more common as tools are developed to look in new ways at digital data. Raw data also enable considerably more thorough validation (important in difficult cases) and provide a complete provenance chain. A fairly common practice is for researchers to produce a CIF and discard earlier data, partly because incentives to keep raw data are low at present and partly because there is little exchangeability in the images produced from the research machine. Some laboratories store and curate the raw data, as described above, but many do not (yet), though there is a low cost to doing this since data standards are well-established and not fast-changing in this community. Curation of diffraction patterns is, therefore, haphazard and incomplete. Yet access to these images would enable researchers to interpret findings better and establish a new quality control point. Developments in analytical tools for digital data will influence developments here.

"Publishing" datasets

The main route to the public domain for crystallography datasets is alongside a journal article. Partly, it is the insistence of the publishers of the major crystallography journals (including the Royal Society of Chemistry) that CIFs be submitted in this way that prompts this behaviour. For larger journals/publishers, authors submit a paper plus a CIF to the journal, and the CIF to the CCDC. Some journals publish the CIF on their websites to

ensure that anyone can access them (even those who do not subscribe to the CCDC or other subscription-based databank services), but where they do not do this, the CCDC is the only curator of the data. Smaller publishers require the CIF to be deposited with the CCDC before an article can be submitted, and the article must carry the CCDC item accession number. [But other publishers such as ACS prevent this? (if I have understood what is said at the foot of the next page)]

Crystallographers can also make datasets available in a way that peers can refer to by either submitting a CIF to the CCDC privately, or by publishing the CIF in a crystallographic journal. Some accept CIFs in this way (*Acta Crystallographica*, *Journal of Molecular Structures*, *Journal of Chemical Crystallography*).

Current best practice for publishing crystallographic data is by the series of journals owned and published by the International Union of Crystallography (IUCr). *Acta Crystallographica Section E* publishes short communications about crystal structures, with submissions being automatically processed because the 'papers' consist of little more than a CIF data file. Peer review is carried out on the file. The journal has been publishing around 5000 items a year. From January 2008, it will be an open access journal levying an article processing charge of about 150 USD, having already been through two rounds of funding by the JISC to examine and test such a business model. The IUCr is very active on the issue of publishing crystallography information and produced an extensive report[24] to the House of Commons Science & Technology Committee during its deliberations on access to the scientific literature in 2004 and a position paper on the matter[25].

As mentioned earlier, it is estimated that over half the datasets produced are not published, including a large proportion of the structures solved by equipment paid for by the UK research councils. The director of the CCDC estimates that the databank would have 'over a million' files in it by now, as opposed to the current c200,000, had all the crystallographic datasets produced been deposited there.

Many of the people we spoke to had datasets that they had not published yet. The datasets are accumulating in unpublished local databases and more than one of the crystallographers we spoke to joked that depositing them in the CCDC would be a task for their retirement. It is clear, then, that once a crystal has been solved and the findings used for chemistry research, publishing the structure is not always a priority. Some people said they always publish any structures they produce but these were not the majority.

Ownership of data and constraints on publication and use

Apart from the very clear case in crystallography where work has been done for a paying client, in which case ownership of the data lies with that client, investigators gave varied answers about the ownership of their data. Some decided that ownership of data produced using public money belonged with their university, some thought data belonged to the individual chemist who made the compound and some suggested ownership resided with the principal investigator of the team which did the work. Some differentiated between ownership of the data and ownership of 'the science' that came from an experiment (and

[24] http://www.iucr.org/iucr/stcttee04.html
[25] http://www.iucr.org/iucr/gicsi/positionpaper.html

concluded that the science belonged to their funder but the data belonged to the investigator).

Response to requests for datasets

Requests for data are not a normal occurrence in this community. It is assumed that datasets whose existence is in the public domain are available through the CCDC.

Discovery, access and use of third party datasets

Discovering relevant datasets

The CCDC databank is the main port of call for datasets that people wish to access. Datasets can be cited alone (i.e. without being part of a journal article) by their CCDC accession number or at other locations and this is the usual means of citing them when a researcher has deposited datasets without an accompanying article. The CDS 'Crystal Web' also provides a discovery service, allowing basic searching across a number of licensed databases. *Acta Crystallographica E* publishes items that are effectively datasets only, since there is little or no introductory or discussion material in an 'article' in this journal and datasets may be cited from this journal, though they also appear in the CCDC databank.

Access to third party datasets

Access is usually via the CCDC, though new open access collections are receiving considerable interest.

Use of third party datasets

Data re-use is growing in chemical crystallography with the development of open access collections of datasets and the machine tools to mine and re-engineer data. One elegant example of an open access collection available for re-use by the entire community is ReciprocalNet[26] developed by a consortium of crystallographic laboratories to share their results. Another example is the Crystallography Open Database[27], an open access repository for CIF files. The JISC-funded eCrystals project is also a ground-breaking initiative in this area, addressing issues of sustainability, linking/citation with the broader literature and integration with other information sources. The project is also assessing community acceptance of these new approaches[28].

The CrystalEye project, which has around 100,000 datasets, was also established to provide an open access database for crystallography data for re-use. It harvests its content (CIFs) from the websites of journals published by the Royal Society of Chemistry, the IUCr, ACS and the Chemical Society of Japan (only from *Chemistry Letters*) and Elsevier (only *Polyhedron*). Wiley, Springer and Blackwell do not expose CIFs so the data associated with articles published in their journals cannot be used. The American Chemical Society (ACS) claims copyright over the CIFS associated with articles published in its own journals, presenting a major barrier to re-use for this body of data. CIFs themselves cannot be copyrighted, but the publisher's format can, hence this stance by the ACS is legal.

[26] www.reciprocalnet.org
[27] www.crystallography.net
[28] www.ukoln.ac.uk/projects/ebank-uk and http://wiki.ecrystals.chem.soton.ac.uk

The re-use of CIFs from the CCDC is permitted for data mining purposes (though harvesting of CIFs from this source to create a new database is not). The re-use of CCDC content means that new kinds of scientific work can be done on this material using new computer tools. Evidence of such work can be seen in the areas of crystal engineering, supramolecular chemistry, crystal growth, polymorphism, crystal structure prediction and structure-property relationships.

Quality assurance

Quality assurance in the data creation process

Crystallographic data are created on dedicated machines that have their own quality checks embedded in the data creation process. Data integrity is also checked by hand in the laboratory before data are analysed or stored or published.

Data management planning

No-one we spoke to had ever been required to write a formal data plan when applying for funding, but most give thought to how they are going to store data. In practice, this generally means on DVD or hard disk. Some email data to colleagues as a backup. Several specifically mentioned the e-Crystals project as a hope for the future in that they will have a safe, open access database in which to deposit and find otherwise unpublished datasets.

Quality assessment of datasets

Datasets are peer-reviewed as part of the article if submitted to a journal. For journals that do not keep the CIF itself but instead require its deposition in the CCDC databank first, reviewers may request the CIF file from the CCDC for review purposes. The CCDC performs checks on the CIF file to ensure the integrity of the data at this stage. The CIF file contains derived data, however, and it is not possible for reviewers to drill down to the raw data to satisfy themselves that the correct interpretation and conclusions have been arrived at.

The IUCr provides a service called CHECKCIF which can be used to verify the integrity of a CIF to a certain extent and individual researchers can and do use this before working with a CIF generated by another worker. Workers need training in the operation of the software and interpretation of its outputs. Being able to judge a CIF and any accompanying information is a skill that comes with experience. Although the community is pretty well self-regulating in the area of quality control – it would be professionally damaging to be found publishing flawed datasets or erroneous conclusions – it remains to an extent a case of user beware.

What motivates researchers to publish data?

Push factors

In the absence of any policy from the largest funder in the UK, the EPSRC, the main impetus to publish datasets comes from the career reward of publishing journal articles.

Pull factors

Crystallographers publish datasets mainly as part of a journal article that is reporting new science because this is the primary route to recognition and career reward. Some deposit CIFs in the CCDC as a matter of routine or, often, when they think about it, when they have time or when they have accumulated a certain critical mass of CIFs.

For routine crystallography work – and it was pointed out by several people that only two decades ago it took months to solve a crystal structure and now several can be done in one day – CIFs are never going to become associated with a journal article simply because so many are generated these days. The impetus to publish these by depositing them in the CCDC databank is low because there is no explicit career reward for doing so. The CCDC recognises this and has developed a service highlighting the Top 200 authors, reflecting the number of datasets they have deposited or been cited on, though it is not prominent on the CCDC website[29]. The funding bodies currently have no mechanism for recognising such a contribution to science, though moves towards a metrics-based approach to research assessment provides an opportunity to introduce measures to assess contributions of this kind.

Despite the low overt recognition associated with depositing datasets, researchers are keen to be authors on articles because this does have explicit career rewards. There is rather a lot of discussion about this within the community at the moment. Many feel that the production of a crystal structure alone does not merit co-authorship, particularly if produced in the case of a crystallography service, and that there is a pressing need for a system that recognises and rewards the independent (i.e. non-journal) publication of crystallographic datasets. The format of 'articles' in *Acta Crystallographica E*, with little discursive matter around the dataset, is one variant of a moves towards recognising that datasets are a legitimate scientific output which deserve recognition.

[29] http://www.ccdc.cam.ac.uk/products/csd/statistics/

GENOMICS

Overview

Genomics is the study of the genome, the genetic complement of living organisms. Genomics work covers the sequencing, locating and physical mapping of genes on the chromosomal complement of an organism. Although in its widest use the term covers everything from gene (or RNA) sequences to chromosome structure and function, in practice its use is generally more restricted to the molecular end of the scale, with chromosome science more usually referred to as that, or subsumed under the header cytogenetics, cytology, cellular genetics, nucleus biology or similar.

The genomics research carried out in the UK is funded largely by the Biotechnology and Biological Sciences Research Council (BBSRC), the Medical Research Council (MRC), the Natural Environment Research Council (NERC) and the larger biomedical charities such as Cancer Research UK, most of them members of the Association of Medical Research Charities (AMRC has 114 member charities[30]). Of these charities, the Wellcome Trust is the biggest spender on genomics research. The Wellcome Trust led the way for mandating public access to research findings that it had funded, the BBSRC and MRC followed suit, and were joined by the other charities funding medical research. Over 90% of biomedical research in the UK is now carried out under a mandatory open access policy. This applies specifically to articles published in journals, but most of these funders also have a policy on sharing data.

The BBSRC produced guidelines on data sharing that were endorsed by its Council in mid-2006 and came into effect in April 2007. In addition to guidelines on making data available to others, one of the recommendations is that all raw data must be retained for 10 years. Applicants for BBSRC grants must include a data plan in their proposal and may also include a request for funds to support their data curation and sharing activities.

The MRC has a similar policy, implemented for grants starting from January 2006. As well as funding individual projects or programmes in universities the MRC has a number of large research centres, and is establishing data management facilities in these centres as part of a corporate plan for data.

The surge in interest in sharing data in recent years was partly a result of the community's open discussions about the desirability or not of patenting genome data, and more recently about the importance of data sharing in responding to developments such as SARS (Severe Acute Respiratory Syndrome) and avian influenza.

Genomics is a relatively mature science with respect to data archiving practice and there are norms that are mostly adhered to by the community for depositing data in the public databanks in the US (NCBI: National Center for Biotechnolgy Information, funded by the NIH[31]), the UK/Europe (the EMBL databank hosted by EBI: European Bioinformatics Institute[32]) and Japan (DDBJ: the DNA Databank of Japan[33]). Although strictly speaking

[30] http://www.amrc.org.uk/HomePage/Default.aspx?lang=en
[31] http://www.ncbi.nlm.nih.gov/
[32] http://www.ebi.ac.uk/embl/

only the US databank is named GenBank, in common parlance this is used as a general term to refer to all these three databanks, which share and mirror each others' content. For simplicity, we will also conform to this practice in this report. Researchers use whichever databank service is most convenient, which normally means that they use the one that has their preferred search interface or the one likely to be least busy at the time they wish to make a data query,. When uploading (submitting sequences) or downloading large datasets (or complete database copies), it is normal to use the one most geographically close, to save on transatlantic bandwidth: this is still a consideration for those on slower backbone connections or companies paying for connection time.

These three giants operate to high professional data archiving standards. They are all publicly-funded and their long-term resourcing appears assured at the moment. The entire international genomics community uses one or other of them. All content is freely available for access and re-use; indeed, that is their *raison d'être*, since genomics research is dependent upon being able to compare sequence findings and use other workers' sequences to produce new results.

There is also a large number of databases of derived data – that is, data derived from DNA and protein sequences – such as databases of protein functional domains, active motifs and so forth. The best-known metadatabase containing such data is Interpro – which contains 11 constituent databases[34]. Such databases provide a valuable resource for the functional annotation of genomic data. Following in their footsteps came the establishment of databanks for storing microarray data. The best known is probably ArrayExpress[35], curated by the EBI for EMBL. The raw data are in the form of .tiff files showing the image of the array plate. Image analysis software is then used to quantify the image, that is, to measure the intensity of the light from each spot. These intensities may then be plotted graphically.

Data creation

Form and variety of data produced

Genomics research produces several types of data. First, there are **genomic sequences** which may contain protein- or RNA-encoding genes and also intergenic regions and repeat sequences. Producing these is now a highly automated process carried out on sequencers capable of generating large numbers of sequences in a short space of time. Large laboratories with gene sequencing machines provide a service to smaller groups. The maximum length of a piece of nucleic acid that can be sequenced *in a single read* is up to 1000 base pairs of raw sequence and so whole genes or parts of genes must be sequenced as a series of short lengths and then the resulting sequence segments pieced together to arrive at the whole gene sequence.

Gene sequencers (machines) produce raw data in the form of a 4-colour trace (one colour for each base) with peaks corresponding to the presence of a specific base[36], and as a sequence

[33] http://www.ddbj.nig.ac.jp/
[34] http:www.ebi.ac.uk/interpro/
[35] http://www.ebi.ac.uk/microarray-as/aer/?#ae-main[0]
[36] http://www.ncbi.nlm.nih.gov/Traces/trace.cgi?&cmd=retrieve&val=448068227&dopt=trace&color=on&size=1&seeas=Show

of letters denoting the base sequences[37]. The machine's 'base-calling' software can misinterpret the peaks and assign the wrong base to a position on occasions (known as a reading error) and workers may manually check the letter sequence against the raw trace if there appears to be an anomaly. Detection of anomalies is possible because most experiments consist of running a number of repeat samples of a specimen through the sequencer – several hundred in some cases. The sequencer base-calling software also provides a statistically-derived quality score for each region of the sequence and this provides users with a certain level of confidence in the overall consensus assembly result. This becomes particularly important when looking for single nucleotide polymorphisms in the sequence.

Over the last year, the situation has started to change very rapidly, with the emergence of the new so-called 'high throughput' technologies for DNA sequencing. These new methods are expected to be particularly valuable in the re-sequencing market (where multiple individuals are sequenced in order to study variation – e.g. cancer mutations), also looking at inter-species variation in closely related species. The higher sequencing error in some individual high-throughput reads (approximately 3% compared with around 1% using current methods) and difficulties of assembling these far shorter read lengths are partially offset by the strategies of sequencing far more copies of each piece of DNA, and/or using a high-quality previously derived genome from the same or closely-related organism as a template on which to assemble the new reads. The software required to fully exploit the new data is at a nascent stage, with a number of open source early-release packages for assembling the new data types *de novo,* or as hybrid assemblies with other data. These have significant practical use issues until they become more mature. For example, they can be unstable, have bugs, and have very high CPU and/or memory requirements. Base-calling, assembly and viewing software used for conventional sequencing are not useable with the new data. The vast amounts of data produced in a single machine run mean that new paradigms for storage of unprocessed data (pre-base-called) need to be sought by the larger sequence producers since it will quickly become unfeasible to store all such data in the longer term.

Error detection is important because people use gene sequences to check for homologies between genes or between genomes of different organisms. The biggest laboratories and groups deposit *all* raw traces into GenBank but smaller groups generally deposit just one representative trace. This is because they do not have the time to deposit the several hundred they will have collected for each specimen. There is, therefore, a (relatively minor) issue of trading quality *versus* detail *versus* the number of sequences that can be worked through. There is no sequence without some errors and the community works with this knowledge.

The second main type of data produced is microarray data. Microarray technology is generally used by genomicists to look at gene expression, i.e. the mRNAs (messenger RNAs) produced by transcription (i.e. expression) of genes or in locating genes in linear space in the genome by hybridisation to known genes or oligonucleotides (gene fragments). Microarray technology is a methodology where thousands of genes or gene fragments are covalently bonded to a substrate and then hybridised to the mRNAs (usually in the form of cDNA) of interest. The substance of interest is commonly labelled with a fluorescent probe so that the

[37] http://www.ncbi.nlm.nih.gov/Traces/trace.cgi?&cmd=retrieve&val=448068227&dopt=fasta&color=on&size=1&seeas=Show

raw data produced from such assays are in the form of scanned images where each microarray appears as a grid of more or less luminous dots. The advantage of such technology is that thousands of genes or oligonucleotide sequences can be assayed at once, referred to as a high throughput process. Newer types of arrays, such as exon arrays, containing up to 5 million data points per chip (where multiple chips are used in a single experiment), require significant computational power for downstream processing, in particular when data normalisation is performed across multiple chips. Protein arrays, although less mature than conventional microarrays, are also becoming more mainstream.

Other data types in this field are amino acid sequences (the protein sequence that is derived from the translation of mRNA), and photographs or drawn images where gene sequences are physically mapped to a chromosome using FISH (fluorescence *in situ* hybridisation) technology. These may be published in journal articles but are not commonly deposited in any public databank, although there are examples of small databases for specific chromosomal/gene mapping data. Protein sequences, strictly belonging to the field of proteomics rather than genomics, are deposited in large public protein databanks of which Uniprot[38] is the primary service.

Metadata

The metadata requirements for GenBank are extremely demanding and reasonably strictly adhered to overall. Curators at the DNA sequence databases are highly-trained professionals who critically appraise each newly-deposited entry to ensure that the metadata are consistent and correct. If there are ambiguities or anomalies, then the curators will refer the issue to the depositor, entering into email correspondence about it until the matter is resolved. It is not, however, compulsory to complete all the fields.

The main problems are that there is no consistent vocabulary for data in many of the fields. Data may be missing, inconsistently annotated (that is, multiple phrases are usable to convey the same meaning), somewhat arbitrary (keywords are added or not by the author and may not reflect the actual content of the submission, just the author's particular interest in it), or inaccurate (coordinates are incorrect, incorrectly ascribed putative functions, circular annotation errors). There is also a huge legacy of poorly annotated entries from older, less strictly-policed times. Overall, one of the largest problems is that the evidence trail used to ascribe particular feature annotations is rarely available. For instance, if a putative function is ascribed to a feature (for example, it is deemed to be a gene or a regulatory region) this may have been done by noting the sequence's homology to a previously annotated feature which may in turn have been incorrectly annotated. If the function has been experimentally verified, both the method used and the level of certainty attached, although extremely important, will almost certainty not be noted. Efforts are being made to address this in other partially manually-curated DNA databases such as Refseq[39]. Also, Genbank-family databases have multiple sections containing preliminary data associated with high throughput sequencing projects where the data have minimal feature annotations and the sequences themselves are unstable and undergo multiple versions and changes as sequencing progresses.

[38] http://www.uniprot.org/
[39] http://www.ncbi.nlm.nih.gov/RefSeq/

There is also not a metadata element for the provenance of the specimen. In the case of some species or genera, where a central specimen service exists, this is not important since anyone wishing to repeat the science can order a new specimen from that service. This is relatively rare, though, and it is much more likely that the specimen used is in the collection of an individual scientist or herbarium, in which case information on its location would be very useful.

Microarray data constitute an even more complex situation in that the significance and meaning of expression data are only relevant in the context of the individual sample used. Because of this, each data item must be annotated with various pieces of information, including the gene name, details of the sample, the reagents used and so forth. Gene names are not yet fully standardised so there is ambiguity at this level; in addition, the situation is complicated by the possibility of many-to-many relationships between genes and the fact that there are no standard units for gene expression data. Standardisation of sample annotation by the development of ontologies is now the focus of an international effort[40]. There is also a community standard, MIAME (Minimal Information About a Microarray Experiment)[41] which the Microarray Gene Expression Database group (MGED) has defined for descriptive metadata.

Array databases to mirror ArrayExpress are being established at the NCBI and at the DDBJ in Japan and these players are collaborating in the MGED discussions, as are some commercial companies involved in microarray technology. Also, metadatabasing efforts to bring together related data from multiple constituent databases are now coming online. An example is the CELCIUS project[42].

Adding value to data

Genomics data are annotated by the experimenter to provide additional information about the sequence or DNA used in microarrays. Annotations may give information such as the gene name, function, position on the chromosome and so forth. The metadata required by the big databanks also provide considerable amounts of useful contextual information. There are also formal annotation projects that enable the adding of comments and notes to datasets so that sequences are more fully described in terms of location or function. An example is the GO annotation project[43] managed by the European Bioinformatics Institute. GO uses controlled vocabularies and formal ontologies for annotation terms. The main value of this is in defining how terms are related to one another and improving search capabilities by minimising use of acronyms, pseudonyms and so on. Another example of encouraging third-party annotations is the Third Party Annotation Project[44].

---

[40] http://mged.sourceforge.net/ontologies/index.php
[41] http://www.mged.org/Workgroups/MIAME/miame.html
[42] Day A, Carlson MR, Dong J, O'Connor BD and, Nelson SF (2007) Celsius: a community resource for Affymetrix microarray data. *Genome Biol*. 8(6):R112
[43] http://www.ebi.ac.uk/GOA/
[44] Cochrane G, Bates K, Apweiler R, Tateno Y, Mashima J, Kosuge T, Karsch Mizrachi I, Schafer S, Fetchko M (2006). Evidence Standards in Experimental and Inferential INSDC Third Party Annotation Data. OMICS: A Journal of Integrative Biology. June 1 10(2): 105-113. http://darwin.nerc-oxford.ac.uk/gc_wiki/images/0/0f/Evidence_stds.pdf

Long term viability of datasets

Data deposited in the big public databanks are guaranteed to be curated and preserved to professional standards, though most of the data are not manually curated. There are levels of curation: for example, although datasets are preserved for the longer term and metadata are checked, the quality of the actual sequence is not checked at deposit and the annotations are not checked for accuracy. Nonetheless, the level of professional care for sequence data is relatively high. Not all data generated are deposited, though. The large genomics laboratories, such as the Wellcome Trust-funded Sanger Institute in Hinxton, Cambridge, post gene sequence traces on their own websites as well as depositing in GenBank the traces from every repeat of an experimental run. Smaller laboratories or groups tend to deposit one representative dataset (of the many repeats) in GenBank and may or may not archive their raw data (traces). So although best practice is for all the raw traces from the sequencing machine to be kept this is far from universal. This may no longer be feasible, however, once high throughput technologies take off, although the NCBI trace repository has recently published standards and file formats for submission of high throughput trace data.

Microarray data deposited in the big databanks are also properly looked after, though again not all datasets are deposited and not all that are deposited are sufficiently well annotated to be re-used for comparative analyses between experiments. Image data from confocal microscopy tend to be massive and are generally kept by individual investigators on local computers or storage media.(local disk arrays, DVDs) and thus curation in the long term is an issue as media degrade, and storage technologies change.

There are many smaller specialised public databanks that also accept and curate datasets from genomics research. For example, in the area of gene sequences there are numerous databanks dedicated to a single species or genus, such as TAIR (The *Arabidopsis* Information Resource), which collects data and information on the plant *Arabidopsis thaliana*, a model organism. In molecular biology alone there are over 1000 of these smaller, specialised databases[45]. As metadata become more rich, and different types of data are integrated more routinely for a systems biology approach (see the Systems Biology section of this report) towards the single organism, the number of such sites is likely to increase. Many if not most of these are funded rather precariously, on rolling funding that is not guaranteed in the long term or on money that is sequestered from research grants – that is, money that was not originally specified for this purpose but which has been allocated to it from a project grant. Naturally, the future of these data stores remains uncertain although newer initiatives by the BBSRC have noted this and are trying to at least partially address this problem. There are many examples of such databases that have been set up and then run out of funding, or that have been set up as part of a project output and, once the project has ended and staff have left, have remained in place on the web but in static form with no future curation or maintenance activity planned.

[45] http://nar.oxfordjournals.org/content/vol35/suppl_1/index.dtl#EDITORIAL

Working with the created data

Tools and technologies for analysis

The software tools required for analysing genomics data are not standardised but all specialist laboratories and groups have a wide range of tools they need for this task. Individual non-specialist researchers who may need to work with publicly-available or their own datasets generally require specialist bioinformatics help as well as access to software and hardware above and beyond that available on individual desktops. Manipulating data does not present a technical problem in this community among specialists but may pose ongoing and significant problems for individual researchers who wish to work with even relatively small datasets and who do not have programming or computational expertise. Most genome biologists take local copies of raw sequence data and analyse these with their preferred software. Some write their own code for this, but others engage the help of bioinformaticians, computer scientists, mathematicians to carry out the task of software development. Searching for and building the required data subsets, reformatting in appropriate formats for the analysis software, and downstream processing of analysis output into human-understandable digests is often non-trivial and requires scripting and/or coding expertise to automate the process.

There is no overall community standard for raw microarray data as each scanner type generates data in a slightly different format to others, but bioinformaticians are able to write script that enables users to work across differing data formats. There do not appear to be any particular new problems associated with genomics data analysis other than complexities of size and incompleteness/errors.

"Publishing" datasets

Genomics datasets are mostly made public by depositing them in public databanks and in the main this does happen. The deposit step is taken when a journal article about the data and the science that has been done using the data is submitted for publication. The main journals in the field enable the sharing of data by having policies that require this[46]. They may police this requirement by insisting on a GenBank (or similar) accession number for the datasets associated with an article. There is, therefore, an explicit requirement for data sharing from both funders and publications. There is also a general ethos in the genomics community that sharing data is the norm to which everyone should adhere. Nonetheless, these measures and norms are not absolutely effective and a recent study reported that up to 20% of articles do not include GenBank accession numbers and up to 15% of studies never submitted sequences to GenBank at all[47].

It is generally agreed that while the rules are in place, the policing of compliance is low. Journal editors may take on this role, sometimes quite assertively if a researcher reports to them that the data behind an article the journal has published are not publicly available and the data creator is not responding to requests.

---

[46] See, for example, *Nature*'s stance: http://www.nature.com/authors/editorial_policies/availability.html

[47] Noor, MAF, Zimmerman KJ and Teeter KC (2007) dat sharing: how much doesn't get submitted to GenBank? *PLoS Biology*. 11 July 2006. http://biology.plosjournals.org/perlserv/?request=get-document&doi=10.1371/journal.pbio.0040228&ct=1

It is also possible to upload a dataset to GenBank without having an accompanying journal article. Users or referrers are able to cite the database accession number(s) of the dataset when referring to it.

Some researchers are now using Web 2.0 tools (blogs or wikis) to make data available the instant they are created. This is especially the case for DNA primer sequences which are spewed from the laboratory machine direct onto a website. It is possible to find these datasets using Google by searching using a very short base pair sequence.

Complementary to this is the increasing tendency to place methods and protocols information on websites to share with peers the full details of a methodology, something that is not usual in a journal article.

Ownership of data and constraints on publication and use

Most people consider that the university they work for or their funder is the rightful owner of the data they produce. Some think data are the property of the creator. Many were unsure. This is a hazy area for most genome biologists.

Response to requests for datasets

Whilst most people state that they would provide data in response to a request from another researcher, in practice this behaviour is patchy. Requesters report that their requests are never responded to or that the response is that the data cannot be found or have been discarded. The main sentiment appears to be 'willing but unable'. Of course, researchers who have complied with the community norms and deposited representative datasets in a public databank really consider themselves to have done all that is necessary, but there are occasions where other investigators need access to the *raw* data (which may not have been deposited), or perhaps to all the repeat runs of a sample sequence. In such circumstances, most people feel obliged to try to provide the data requested but that may not always be possible because they are no longer kept or cannot be easily located. There is no such compunction to provide access to unpublished data; that is, datasets that have not been referred to in a journal article or deposited in a public databank, even though their existence is known by other researchers.

Sometimes there may be some sort of trade-off is a data-provision arrangement between researchers, with datasets being provided on the understanding that the journal article about them is cited as a 'reward'.

Discovery, access and use of third party datasets

Discovering relevant datasets

Data discovery is routinely done by searching the large public databanks using keywords (such as gene names), author names or more complex terms. Because these databanks have such structured and detailed metadata using them should be very simple, though errors, missing data, pseudonyms and acronyms can still complicate detailed searching. Public web-based search programs are available specifically for these types of searches (e.g. NCBI's Entrez, EBI's SRS service). Researchers may also search for some data using Google, by typing a string of base-pair letters into the search box. Google can locate datasets containing

these strings if are exposed on websites. This type of search is frequently done when looking for DNA primer sequences, which are short lengths of DNA at the beginning of a coding sequence. This type of search can be useful but will be non-exhaustive since most public DNA/protein databases are not comprehensively indexed by Google.

Access to third party datasets

Access is by and large through the public databanks. There are no restrictions on access to these and anyone can take and use data from them. Some databases have a subscription system for access or a pay-per-access arrangement but the view is that these mostly come with such restrictions that they are not useful anyway. Access to privately held data is dependent upon the attitude of the data holder who may have legitimate reasons for withholding permission, perhaps because the data represent work-in-progress and further exploitation has to be done, or because of privacy issues to do with the provenance of the data.

Use of third party datasets

Re-use of publicly available data is a given in the genomics community. It is a cornerstone of research practice in this field. Re-use of privately-held data is a matter for negotiation between the prospective user and the data holder. Often this is seen as an opportunity for new collaboration and even procurement of joint funding, so approaches from one scientist to another for data can be regarded as a positive thing with a potentially fruitful outcome.

Quality assurance

Quality assurance in the data creation process

Standard practice is for gene sequence data to be checked manually after each series of runs on a sample. These checks are to identify and correct reading errors as the machine assigns a nucleotide letter to each peak on the raw trace. Checks can also identify other issues, such as where different allelic sequences are revealed. Once the data have been manually checked in this way the datasets are ready for deposit in the public domain as soon as the data creator deems fit. With most high throughput sequence work there is not the time or manpower to operate in this way. Assemblies may be checked manually but basecalls in individual traces are not corrected manually in even modest size projects. One point of note is that such manual checking could, in fact, increase overall error rates in assemblies (due to the subjectivity of the checker) and basecall statistical quality values are generally used to calculate the overall consensus quality. Most reads from small projects are not deposited until assembled, though the large sequencing 'factories' follow the essence of the BERMUDA agreement and deposit raw sequence nightly. EST (expressed sequence tags) sequences are a good case in point: with these, individual single pass reads straight off the sequencer are deposited to their own section in Genbank with no manual sequence checking at all, In these cases, unresolved basecalls remain in the sequences as strings of 1 or more N's and have an overall error rate in the sequence of around 1%.

Data management planning

Project proposals being made to the BBSRC, MRC, NERC, the European Union and Wellcome Trust must have a formal data plan in them. The plan should detail what data will be generated, how they will be used and how they will be shared. In practice, people do

write a plan into their grant applications but it is not always detailed and formalised along the lines above. A number said that the data plan is the last thing they write and it is not a part of the application that they spend much time on; rather there are a set of words that they know are needed for this part of the proposal and they may be included without being given too much thought. If the experimenter needs new computer equipment or staff time for manipulating or storing data as part of the proposed project then this need will be spelled out in more detail.

Quality assessment of datasets

At deposit, the database curators also check the datasets. There are stringent checks on the metadata and contextual information and there may be correspondence at this stage between the database curator and the depositor to ensure the integrity and validity of the dataset.

After deposit, quality assurance is largely carried out through data use by the community and the reporting of significant errors and flaws.

Journals also have a role to play in quality assessment, since editors/reviewers check the sequence data for general quality and for the methodology and so forth. A journal can accept a consensus sequence rather than insisting that all repeat runs are published. Journals are also meant to check that data actually have been deposited if they are mentioned in an article though this does not necessarily happen: editors or reviewers may assume a dataset has been deposited in the public domain but do not actually check. Submitters can also request that data are not made available until after the journal article's publication date. Occasionally, the data are not made public in the database at this point by error (on the part of the journal), until requests are made to release them.

What motivates researchers to publish data?

Push factors: the effect of policy

With all the main funders of genomics research having a policy on data sharing the push factor to make data publicly available is very strong in this field. Not everyone yet complies with these policies, but the great majority do.

Pull factors: intrinsic rewards

There is no explicit reward for making data publicly-available but the agreement to do so within the community is real and largely borne out in practice. Researchers reported that they share data anyway, even if their funder may not require it, because it is expected behaviour with a 'warm feelings' factor for them. There is a strong esteem factor involved here and researchers who do not share data in genomics are very much frowned upon by the rest of the community. Most people would include on their CV the fact that they had deposited a number of datasets into the public domain since this would be regarded as 'playing the game' and makes the researcher a valued member of the community.

SYSTEMS BIOLOGY

Overview

Systems biology shares many of the approaches and tools of genomics but it has a broader scope and reflects a new way of thinking about biological questions. The systems approach is integrative, using mathematical and statistical techniques to manipulate biological data with the aim of developing predictive models of biological behaviour, identifying emergent models that could not be predicted by studying individual parts of a biological system in isolation. Experimental biological data are processed to enable an understanding of the relationships in a biological system that can ultimately be described by differential equations or other mathematical or computational methods (e.g. Markov chains, process calculus, inductive logic programming). Accordingly, systems biology teams include not only biologists but scientists from other disciplines too.

In 2005 and 2006, the Biological and Biotechnology Research Council (BBSRC) and the Engineering & Physical Sciences Research Councils (EPSRC) funded the establishment of six systems biology centres in the UK at a total cost of £40 million in the following universities: Newcastle, Imperial College, Manchester, Edinburgh, Nottingham and Oxford. The BBSRC has a policy on data sharing and use; the EPSRC does not. There is therefore some confusion in the community on this topic. Other major funders of systems biology work, or work that contributes to the systems approach, are the Medical Research Council (MRC), the Wellcome Trust and the large medical charities.

There are two main ways to create data in systems biology. First, there are experimental procedures. These may be high-throughput processes that provide ways to attempt maximum and systematic coverage of a biological problem. Examples of this are the microarray experiments that were described earlier in the Genomics section, or experiments to determine characteristics of protein-protein interactions. Or they may use, for example, sophisticated microscopy techniques that combine light microscopy (usually confocal work) with electron microscopy studies to produce structure/function interpretations of a biological process.

The second main method is to work with curated datasets, piecing together data already in the published literature (or those plus experimental data) to assemble evidence of a new interaction, metabolic network or whatever biological phenomenon is being studied.

In both kinds work, there is a high degree of reliance on informatics and computer science technologies to enable the sophisticated manipulation and interpretation of the data produced at the bench. Mathematicians, computer scientists, engineers, physicists and chemists may play a part in generating and processing data in systems biology. One characteristic of systems biology is the vast amount of data being produced. This will only increase, and in some areas biologists are looking at the grid as the solution to the problems they face in processing their experimental (or derived) data.

Systems biologists use the public databanks described in the Genomics section of this report for depositing relevant datasets and for obtaining data for re-use. There are also other public databanks curating data that pertain particularly to systems biology. The EBI at

Hinxton curates biological modelling data, for example[48, 49]. Other examples of interest include databases of metabolic pathways (e.g. KEGG -Kyoto Encyclopaedia of Genes & Genome[50]) and protein interactions (e.g. the Database of Interacting Proteins DIP[51]).

Data creation

Form and variety of data produced

Gene sequence and expression data, as described for genomics, form part of the panoply of outputs from systems biology work. So-called 'tiling arrays' are common, where an entire genome of an organism is present on the assay plate. Other types of experimental data are biochemical data, such as protein characterisations from mass spectrometers (the field of proteomics), metabolic pathway information (metabonomics) from nuclear magnetic resonance studies, and images (such as confocal micrographs and electron micrographs). In addition, there is another level of data produced in systems biology – data derived *from* experimental data: these may be models of various types – mathematical, computer models, simulations, 3-D models and so forth. The field is broad and encompasses many different and disparate systems, though there is currently a strong focus on the molecular biology of yeast (which is a model system in molecular/cellular biology) and on particular metabolic processes in higher organisms.

The size of datasets can be enormous. For microarray data a text file is produced that may have up to 5 million rows, each representing a gene or gene fragment, and 80 columns of information. An individual experiment may involve from half a dozen to several thousand arrays and arrays may themselves be one-channel (one colour) or two channel (two colour). Experiments that produce images and use image analysis software to process the information contained in them can produce 200 gigabytes datasets. They will approach petabyte size in future, demanding grid computing facilities for effective manipulation.

Standards for data are not fully mature overall though there is considerable progress in this area. There are well-established standards for gene sequence and microarray data, as discussed in the Genomics section of this report. The proteomics community is now fairly well advanced in developing standards[52] (there is, for example, along the lines of microarray data, the MIAPE – Minimal Information about Proteomics Experiments) and standards are currently being developed for metabonomics. For model data there are now standards developed by the EBI (European Bioinformatics Institute), standardised on XML which is

---

[48] Le Novère N. (2006) Model storage, exchange and integration. *BMC Neuroscience*, 7(Suppl 1):S11. http://www.biomedcentral.com/1471-2202/7/S1/S11

[49] Le Novère N., Bornstein B., Broicher A., Courtot M., Donizelli M., Dharuri H., Li L., Sauro H., Schilstra M., Shapiro B., Snoep J.L., Hucka M. (2006) BioModels Database: A Free, Centralized Database of Curated, Published, Quantitative Kinetic Models of Biochemical and Cellular Systems. *Nucleic Acids Res.*, 34: D689-D691. http://nar.oxfordjournals.org/cgi/content/full/34/suppl_1/D689?maxtoshow=&HITS=10&hits=10&RESULTFORMAT=1&author2=le+novere&andorexacttitle=and&andorexacttitleabs=and&andorexactfulltext=and&searchid=1&FIRSTINDEX=0&sortspec=relevance&resourcetype=HWCIT

[50] http://www.genome.ad.jp/kegg/

[51] http://dip.doe-mbi.ucla.edu/

[52] See the Proteomics Standards Initiative: http://www.psidev.info/

much richer and structured than spreadsheets or text. Existing markup languages, for example SBML (Systems Biology Markup Language), may not be sufficiently rich to capture all required information for particular scenarios, however, and the standards developed for differential-equation derived models may be inappropriate for models generated by inductive logic programming. Some of the larger and as yet unresolved problems of making models available in a meaningful manner are that multiple underlying methods are used to generate models, there are few current standards to publishing models – that is, multiple file formats can be generated by differing programs. Perhaps more fundamentally, the model itself is usually only a small part of the package required to evaluate, and ultimately re-use, the model. In addition, researchers may well require the algorithms used to manipulate the model, that is, to fit the model to data, solve the model mathematically, simulate the model or compare models. 'Algorithm' here includes the actual implementation of the mathematical model in a computer language (e.g. Python, a specialised mathematics package). In the ideal scenario, researchers would also have access to the actual datasets used for building and testing the model.

Metadata

Metadata are made available via journal websites, using appropriate legends or descriptors, and databases have certain requirements for metadata but outside the province of the professional databanks metadata standards can be very variable.

Adding value to data

Many systems biology datasets are annotated to give extra context, meaning and purpose. A large component of this is to enable provenance tracking of data in the biological and experimental context. In this area more than perhaps any other, it is vitally important to capture rich information about the biological context, samples and experimental methodologies used. Without this, it is impossible to relate data meaningfully across multiple experiments (possibly carried out in different laboratories or when data are derived from public repositories, or from different types of experiment (e.g. microarray, proteomics, confocal microscopy) performed on the same biological system. Furthermore, data quality information becomes particularly relevant (highlighting the importance of errors and missing data) since this can be critical in the modelling process. The effects of errors are quickly compounded when looking for informational signals across multiple datasets and data types. It should be noted here that inadequacies in existing repositories are widely acknowledged. For example, even annotations that are MIAME-compliant for microarrays may be insufficient in this context, and some databases such as GEO[53] do not insist on annotations to even this basic standard.

Long term viability of datasets

The BBSRC requires raw data to be kept for ten years and the large systems biology groups it funds comply with that requirement. This is a major undertaking by itself, requiring considerable storage capabilities and, ideally, great curation expertise. Storing for the long term the outputs from confocal microscopy used in the systems approach, especially where derived data are produced in the reconstruction of images, is a problem. That is probably too large for smaller groups who lack the funding and facilities necessary. High-throughput

[53] http://www.ncbi.nlm.nih.gov/projects/geo/

molecular work (producing sequencing or expression data, for example) can also produce large datasets: one machine can generate a dataset of 4 terabytes per run. These outputs are on a comparable scale to those of astronomy, where centralised data warehousing solutions have been found. The same solutions are used by the big systems biology groups for all wetwork raw data, but smaller teams generally store data on DVDs or a server on which they can obtain space.

In systems biology it is not just the data that need to be stored and looked after in the longer term; the software tools for handling and making the data available are also needed. Researchers do not often itemise this kind of expenditure in grant applications for fear of raising the overall cost too much and jeopardising their chances of being awarded a grant. This situation is easing a little as more 'small tools and resources' grants become available as funders begin to take the issues of data management and sharing very seriously.

Working with the created data

Tools and technologies for analysis

Systems biology data by definition eventually require manipulation by computer scientists and mathematicians for model-building processes. Computer scientists tend to be involved in developing machine learning aspects of the work. Mathematicians use differential equations and other methods such as pi calculus, Markov chains, Petri nets, and process algebra to model discrete biological processes and then use these in a predictive way to discover new biological information (see, for example, the work of Jaroslav Stark's group[54]).

Each systems biology group may also have one or more data managers or programmers who are experts in handling data of various types, constructing databases that the group (and others) can interrogate and manipulate, and producing software to enable this work. The BBSRC/EPSRC-funded systems biology centres generally have two such persons. In some cases these people are on an academic job grade but in others they may be in posts akin to technical grades and thus deemed – officially – to be of lower status than the researchers, though all the researchers we spoke to regret this state of affairs and value extremely highly the expertise and participation of such group members. Others are on 'academic-related' professional services grades which are even more of a grey area, sitting between purely academic and purely technical grades.

Larger groups may buy commercial solutions to enable viewing and working with data, exporting and checking data integrity. There is a quality assurance element here for those who can afford these tools. Examples are Rosetta Resolver or Expressionist for analysing and managing microarray data. Commercial analysis and management solutions – including Resolver – are often based on ORACLE databases, which require additional cost-prohibitive licenses for the underlying database technology. Both open source and commercial LIMS (laboratory information management systems) are available for various data types (e.g. proteomics 2D gels and mass spectrometry data, DNA sequencing, microarrays) but the end cost of these is outside the remit of smaller labs. Whatever the source, these systems need considerable expertise to set up and configure for the specific

[54] http://genomebiology.com/2006/7/3/R25

application and do not work 'out of the box'. They may form an important part of the quality control process, though, as they ensure that appropriate data about the experiments, materials and methods are captured at source, together with the raw data.

Some scientists are uncomfortable with open source solutions, though, but commercial companies usually write software for Windows, an operating system that does not cope well with moving around very large volumes of data. Linux or other UNIX-based solutions are considered preferable in this respect.

A number of issues prevent the optimal generation and handling of data, but the most important is that the field is evolving fast, with wetwork technologies evolving alongside data formats and standards. Only the best-resourced groups employing real data experts can hope to keep up.

"Publishing" datasets

Gene sequence data and microarray data are published as described in the Genomics section of this report. This is the simple end of the scale of systems biology data. At the other end of the scale are datasets that are never made public, though the prevailing ethic is to share, and most researchers try to make data available where they can. Some are using Web 2.0 tools to manage projects and work collaboratively within their group(s) but as yet there are few examples of such tools used to make data public. There are a number of web-based systems for data management and analysis under active development in the microarray arena.

Journals in the area of genomics mostly have a policy of requiring datasets to be made publicly available and this is becoming commonplace in proteomics. Metabonomics data policies are also springing up and that community currently has some standards development going on that will formalise the situation somewhat. To some extent, this is being driven forwards by the journals themselves (e.g. *Nature Biotechnology*), which are promoting the various standards issues.

Larger laboratories upload sequence data nightly to the public databanks (i.e. GenBank) but microarray data tend to be published at the end of the research cycle, when a journal article is written. Larger groups also publish their raw data on project websites, including more contextual information than is usually required by the public databanks.

Other types of systems biology data, such as the outputs from microscopy, cannot all be published with the journal article and there are no public databank set-ups to parallel the gene sequence ones for this type of dataset. Normal practice is for scientists to publish some micrographs as exemplars but to hold the bulk of the outputs locally on their own systems. In any case, it is not possible to represent fully the data that this area of systems biology is generating, since they are complex three-dimensional computer-generated models annotated quite extensively by the researchers and requiring a *lot* of computer space for their manifestation. It is this area of systems biology that is looking at the grid as a solution to the problems of data manipulation, visualisation and sharing.

Some data are published on project or research group websites. This is better than nothing, but this should imply the freezing and building of datasets. Some groups upload data and then replace those data at a later date with newer data. This is not best practice, since other

workers cannot then access the original datasets. Best practice is to 'freeze-and-build' to that all 'versions' of a particular body of data remain available for analysis by other workers. Even with freeze-and-build procedures it is not always clear whether multiple validations have taken place on the data or whether earlier data are simply incorporated without further validation (i.e. assumed to be correct and the experiment not repeated) into the later dataset that is published. Version–control procedures are obviously paramount to this issue.

Freeze-and-build should ideally always be the practice in formal databases that are maintained on the web but often it is not insisted upon, though with the right kind of database structure the practice would be enforced automatically.

Ownership of data and constraints on publication and use

Some researchers may be funded by, or work in collaboration with, commercial organisations. In such cases there may be acknowledged ownership of data by the commercial entity and constraints on how open the data may be made. These situations are relatively rare and in the majority of cases there are no such commercial constraints. As with genomics, however, most researchers are unclear as to who owns their data and usually guess that it is their university or funder.

Response to requests for datasets

Most researchers say that they respond as helpfully as they can to requests for data from other workers. When asking for data from other people they report that in some cases they are supplied willingly and rapidly but that their request may also be met with a refusal or simply ignored.

It is also possible to prevent easy re-use of data whilst appearing to share. Some datasets are made public as 'flat' pdf files to accompany a journal article (journals usually insist on this because it is easy for them to handle) and as text files for a group's website. Such formats are difficult to use, especially pdf and the practice is known as 'protecting by pdf" within the community.

Discovery, access and use of third party datasets

Discovering relevant datasets

Discovering datasets in the large public databanks is a relatively simple process and most people working in the field are also familiar with smaller, specialised databases that are relevant to their work. Some workers, however, reported that they routinely 'trawl' around the web looking for relevant datasets to use. At other times they may need to access data that they know are 'out there' but which are difficult to locate. Even when they know they are on a certain group's website, actually finding the datasets can still be difficult. Every website has a different structure and routes to datasets can be tortuous and non-intuitive. An individual who had consulted all the colleagues in his group before he was interviewed told us that "When we thought about how much time we spend searching other people's suboptimal websites for data it was a shock!"

Access to third party datasets

The main problems in accessing systems biology data are:

i) locating the data
ii) navigability (is it easy to get to the data in a website?)
iii) data formats
iv) consistency of formats between different websites
v) freezing of data

Access to systems biology data is via four main routes:

i) From the websites of journals publishing the articles to which the data relate. This is a fallback position and far from ideal since most journals insist on publishing data in pdf format. Bioinformatics journals are the exception in that many of them will accept data in other formats more suitable for re-use. Even where the data are linear, text-based data, pdf is hard 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 at worst simply impossible

ii) From websites of individual researchers or research groups/projects. This can be a very good option because data deposited here can be very fresh and good quality. The possible downside is that finding the right datasets on websites can be difficult, as discussed above, and the freeze-and-build issue may apply here, too. In addition, web-sites maintained by individuals or groups can be labile where '24/7' access for external researchers is not a primary concern for the group, and the hardware infrastructure resources and personnel to provide such a service may not be available.

iii) From web-based databases maintained by individuals or research groups[55]. These are less *ad hoc* than (ii) but they can be subject to the vulnerability described in the Genomics section; that is, many of them are funded on soft money or on funds sequestered from research grants and so their long term existence is uncertain, and they are not so expertly curated as the large, professionalized databanks.

iv) From large public databanks, which may also provide a centralised service where more than one database can be searched at once, thus permitting related or contextual information to be determined at the same time.

There are no cash costs for users involved in using options (ii) to (iv) above (usually) but there are considerable costs (not usually borne by the users themselves) in subscribing to the journal literature in order to access data. The other cost of access comes from the need for software to decode other workers' datasets, and time/expertise to use it and or build it. One specific example may help to explain this. Mass spectrometer data are almost all produced using MASCOT software, a proprietary tool which is ubiquitous in proteomics laboratories. Groups that do not specialise in proteomics but which wish to use proteomics data are obliged to pay for a MASCOT licence that costs several thousands of pounds per year, per processor. This may not be the only software a group requires, of course, so the cost of tools to enable access to data can be quite considerable. Software for microscopes presents much the same problem. Scientists 'slice and dice' software themselves to create a solution they can work with but it may not be a particularly satisfactory one. There is now

[55] See, for example, BioGrid: www.thebiogrid.org

some discussion in the community of involving computer scientists to address the data incompatibility issue and find a suitable solution.

Accessing data is one thing but accessing the *full* dataset is another. Some datasets in the public domain are incomplete. Expert data managers can piece together incomplete datasets from disparate sources but this requires a level of expertise in data handling that 'average' researchers (biologists) do not possess. Thus, some data are not being made available for use optimally.

Data may not always be of good quality, either. The big public databases have very good internal consistency, as discussed in the genomics section of this report, but the standards upheld do vary between databases, with virtually only the biggest players maintaining the highest possible standards.

Freezing and building of datasets is not a ubiquitous practice and, as mentioned earlier, this is proving to be a big problem in systems biology. Overwritten datasets are better than no datasets at all, but several people made the point that science should at its very essence be reproducible and overwritten data prevent this practice.

Researchers who have work-in-progress are loathe to release data for others to mine and even when work has effectively been completed on a certain body of data some scientists prefer to keep the data under wraps for a year or so until they are sure that there is no more exploitation they wish to do themselves.

Scientists at the EBI were the first to develop XML standards, which are much richer and more structured than text or spreadsheets, for systems biology models data. The EBI also provides a large toolkit of standards, formats and ontologies to describe, annotate and exchange biological models. These, however, are not for machine learning-based models, and this highlights the fact that the science is in its infancy and much further work on standardisation needs to go on. To publish these models, which are small files in mathematical notation or machine language, software needs to accompany the files in order for other workers to access and use the data. This almost inevitably means that some training is needed to work with such data effectively.

Use of third party datasets

If datasets are in the public domain then there is rarely any problem in re-use. Data sharing, despite some anomalies, is the prevailing ethic in this community. Problems of re-use centre around the technical issues discussed earlier – the variety of formats, the non-standardisation of formats, the need for proprietary software and so forth.

Quality assurance

Quality assurance in the data creation process

Quality assurance of systems biology data (outside genomics, which has been discussed in a previous section) resides mainly in the technologies used to create the data in the first place. Most machines have an inbuilt data check system that ensures that data are of good quality and consistency. Manual checks of data outputs are also made by researchers before making

their data public. Note, though, that some groups are 'publishing' data straight onto websites in real time, and others are using these data with confidence, signifying the reliability and accuracy of the machines used to generate these types of data. It may also, though, signify a lack of understanding of the errors inherent in such data, or an acceptance that they are there and need to be accommodated.

Data management planning

Large systems biology groups develop thorough data plans under the auspices of their expert data managers. Smaller groups work largely along the lines of the smaller genomics groups, producing plans of a greater or lesser degree of sophistication according to time available when writing a project proposal. Data management is an issue that everyone in this community acknowledges as very important, but many admit that it can take a back seat under the day-to-day pressure to get benchwork done and finding the resources to pay for it.

Quality assessment of datasets

Referees of journal articles check (or should do) whether datasets have been deposited in the relevant public databanks, thus assuring that the author has conformed to a journal's own policy. They can and do check the integrity of the dataset in some cases but not all. More likely, they can support minimal quality assurance, at least, by enforcing data deposition with suitable standard information. For complex image data and papers on high volume high throughput data this would be a task beyond an article reviewer and such data outputs have to be taken largely on trust as to their accuracy and integrity.

Researchers by and large do trust one another's data, working on two assumptions: first, that no dataset is flawless and second, that seriously flawed data will be revealed by community re-use. That said, there is a properly scientific degree of caution in place when a dataset produced by another worker or group is obtained for re-use. Even within a dataset regarded as good quality there are still internal variations.

What motivates researchers to publish data?

Push factors: the effect of policy

Funder policies are influential in helping to ensure that data are openly available. The situation that pertains in the genomics area also pertains here with respect to those sorts of funded work. Where funder policies do not reach, there is a mix of results. Some researchers make great efforts to share data while others may retain their findings or publish in a form that means that although data are available, they are not readily accessible.

Pull factors: intrinsic rewards

There is no formal recognition for publishing data. Researchers do so because they say they wish to see their data's value increased by citation (of the accompanying article or of the dataset itself by referring to its accession number in a public databank) and re-use. They also cite the desire to play a proper role in their community, which means sharing their data and enjoying the 'nice, warm feeling' that comes from doing so.

RURAL ECONOMY AND LAND USE PROGRAMME

Overview

The Rural Economy and Land Use programme (RELU) is a £24 million research investment which brings together natural and social scientists from over 40 different disciplines and 50 institutions to address four key research themes: the integration of land and water use; the environmental basis of rural development; sustainable food chains; and economic and social interactions with the rural environment. The programme is jointly funded by the Economic and Social Research Council, the Biotechnology and Biological Sciences Research Council and the Natural Environment Research Council with additional support from the Department for Environment, Food and Rural Affairs and the Scottish Government. RELU is an interesting programme to investigate not just because it is explicitly interdisciplinary in nature, but also because RELU has a clearly defined data management policy supported by a dedicated Data Support Service.

Data management has been given a starring role within the RELU programme and, as such, it may provide a model for furthering the data management goals of the Research Councils. RELU's Data Management Policy is based on current best practice within the Research Council community and is based upon five principles:[56]

- Publicly-funded research data are a valuable, long-term resource
- Data must be well managed
- Data must be made available by researchers for archiving
- RELU funds will support data management through the life of the project
- Post-programme data management will be the responsibility of the Research Councils

The pragmatic drivers for putting data management at the centre of the RELU programme are envisaged in the terms listed below. It is worth noting that this programme is one of the first to focus on knowledge transfer between researchers in different disciplines.

- Researchers will be better able to apply learning from one field to another
- Effective data management will enable the combination of different methodological approaches and sources of information
- Data management will facilitate the cross fertilisation of ideas and concepts
- The effective management of project data will help the research community to understand scientific, technological and environmental problems in their social and economic contexts.

The RELU Data Support Service (DSS) was originally conceived to provide a single point of data-related expertise for RELU award holders. The Service is fully embedded within the UK Data Archive which provides the data acquisition and dissemination service for ESRC. RELU DSS exists to provide advice and guidance on matters ranging from the development of data management plans at the beginning of a research project to the process of depositing datasets with one of the established data centres such as the Economic and Social Data Service (ESDS) at the end of a project. DSS staff undertake proactive liaison with

[56] These are listed on the RELU website [www.relu.ac.uk/about/data.htm] and expounded in RELU's Data Management Policy: www.relu.ac.uk/about/Data%20Management%20Plan.pdf

researchers through evaluation of data management plans, training workshops, site visits, evaluation of data at the end of awards as well as providing a data processing and support role. Award holders are not compelled to use the services offered by the DSS, but they are expected to complete a Data Management Plan at the outset of a project and to sign up to RELU's Data Management Policy. Notwithstanding these requirements the overall tone of the programme is to *encourage* researchers to consider the benefits of effective data management data sharing such that, in time, there will be a cultural shift – a widespread recognition that effective data management leading to long term data curation is a valuable activity. They are also expected to submit data or samples of data at the end of their project via the DSS

For this part of the project we spoke at length with ten people, all principal investigators (PIs) or others closely involved with RELU-funded projects. While clearly this is not a comprehensive survey it is sufficient to provide sound insights into award-holders' attitudes to data management and to make a summary judgement about the overall effectiveness of this type of approach to data management.

Data creation

Form and variety of data produced

The interdisciplinary nature of projects funded by the RELU programme means that a wide variety of data is produced spanning the biological, environmental and social sciences. The range includes field observations, monitoring and laboratory experiments through to qualitative interviews. The actual type of data produced includes: numeric and tabular data; GIS and CAD data; survey and qualitative data; audio and image data. Although much of the data that is core to award holders' research are primary, many researchers also rely on input data from third parties, notably the spatially-based large scale databases held by organisations such as DEFRA, the Environment Agency and the Ordnance Survey.

Purposes of data generated as research output

One of the key purposes of the data generated as research output from RELU-funded projects should be the availability of those data to other researchers. Data management and ultimately the curation and re-use of data created using public funds should be fundamental to all RELU projects by design. In reality attitudes to data management vary markedly even between researchers in the same research group. In fact the attitudes of researchers working on RELU-funded projects are basically the same as researchers whose projects are funded by other means: the datasets collected and processed in the course of a research project are but stepping stones to producing journal publications. Journal publications remain central to career progression and reputation; publishing datasets, by contrast, appears to offer few if any tangible benefits. The RELU programme does, though, describe a clear pathway that datasets should follow in terms of them being offered for curation. By and large researchers comply with the requirement to offer data, though in most cases this seems to be a passive process, not an opportunity grasped with much enthusiasm – particularly among researchers who doubt their datasets will be of use to, or be used by, other researchers.

Metadata

Researchers who receive funding from the RELU programme are expected to provide metadata that describes their datasets. It would be fair to say that awareness of the role of metadata is greater among RELU award-holders than among some other areas of the research community, largely due to the requirement for them to read and sign up to the Data Management Policy. However, the extent to which good quality metadata are actually produced to an appropriate standard (that is, in a form acceptable to data centres) varies widely in accordance with the importance project teams attach to the process of data management. Attitudes to data description typically range from informal ("we do it on the hoof") to formal, where the project has a dedicated data manager.

Adding value to data

Because the RELU programme covers many different disciplines, even within project teams people find it difficult to generalise about how value is added to raw data. In the main, though, processes such as data cleaning, data coding and deriving higher order data are perceived to add value to datasets. Such activities are almost always specific to the goals of each research project; it is rare to encounter researchers who are adding value with the aim of making data more accessible to others, or usable by others, beyond mainstream functions such as providing standard metadata.

Long term viability of datasets

Most researchers working on projects funded by RELU know that one of the conditions of their award is that the datasets that result from their work should be offered within three months of project completion to either the ESDC or one of NERC's designated data centres, where it will be considered for curation. If data creation methods or data management have been poor then the resulting datasets may not be suitable for curation: data formats may be inappropriate or crucial metadata may be absent, for example. Data centres are not obliged to accept for curation datasets that suffer such deficiencies, nor are any sanctions imposed on project teams that deliver datasets unsuitable for curation or subsequent re-use. The result appears to be that, except for those researchers who recognise and have internalised the value of making their data available to others in the research community, researchers who do not present their datasets to an acceptable standard for whatever reason are not subsequently held to account. While ESRC researchers who are unable or unwilling to comply with data sharing (through reasons of not seeking consent and so on) are written to by ESDS,a formal mechanism for linking non-compliance to a penalty of any kind is not evident. This could be retention of final part of an award grant or jeopardising successful future funding are not yet in place, which both happen if an End of Award Report is not submitted or fails.

Working with the created data

Tools and technologies for analysis

Because the disciplines covered by the RELU programme are so numerous it is not surprising that many different tools and technologies are employed for data analysis. Tools that are common to many projects include statistical processing software such as SPSS and

software in the realm of Geographic Information Systems since many projects have spatial elements to them.

"Publishing" datasets

The general impression derived from the interviews is that researchers tend to believe their datasets will not have a life after the end of the project, and that they are very unlikely to be used by or be useful to other researchers. Many researchers admit to paying little attention to the longevity of datasets they produce; once a project is finished their attention shifts to the next project. In this context, where datasets are accorded little importance, it is not surprising that researchers tend not to dwell upon issues to do with "publishing" their datasets. In fact RELU award-holders are relieved of having to worry about making their datasets publicly available since, once they have submitted their final datasets to a relevant data centre in compliance with the terms of the Data Management Policy, they need worry no more about those datasets unless they choose to have a greater level of involvement.

Ownership of data and constraints on publication and use

In the majority of cases researchers believe that their employing institution owns the data they produce. These researchers tend not to be troubled by issues to do with intellectual property rights since rarely is there any monetary or other commercial gain associated with the outputs of their research. It was mentioned that, for the tiny proportion of projects where there is potential for commercial exploitation, universities are quick to claim ownership of any applicable rights. Other than in these rare cases, data ownership is not, *per se*, a constraint on making that data available to others with some exceptions:

a. Where researchers have used data owned by third parties there may be constraints on its re-use. Typically researchers must negotiate or purchase access to datasets owned by, for example, the Environment Agency or Ordnance Survey. The licences associated with these transactions will describe the limits on the subsequent use or publication of these data or conceivably any derived datasets.
b. Where research is largely based on interviews with individuals, unless their consent to publish their responses has been sought and given. Project teams are sometimes reluctant to seek such consent at the outset for fear of interviewees refusing to participate in the study – but then researchers are also reluctant to seek consent at a later stage because of the cost of doing so and, again, the fear that they may withdraw from the study.

Response to requests for datasets

As with some other disciplines covered by this study, researchers within the RELU programme are generally reluctant to grant access to their data at least until the end of the project but normally until a later stage. The Data Management Policy allows for a period of exclusive access by the data creators of up to one year so they have the opportunity to benefit from the data, normally in the sense of deriving publications. This reluctance to share is based on several factors:

- Researchers compete for funding and intangible benefits such as prestige. Having gone through the process of winning funding and collecting data, the natural course is for those researchers to be able to analyse the data and publish papers before releasing the data to others.
- The nature of researchers' responses to requests for access to their data is conditioned by their social networks. People are likely to treat requests from people

they know differently from those they don't, particularly where the latter are perceived to be competitors in some sense.

- Researchers are also somewhat fearful that their data may be misrepresented. The best way to minimise the chance of this happening is to provide good quality metadata and other contextual information that would help a third party use the data in an appropriate way. Of course not all research teams have this type of information to hand and most do not have the time to spend describing the characteristics of the data to third parties, so responses to requests for datasets are not always forthcoming or are deferred until the datasets are managed by a data centre.
- There may be licensing restrictions where research groups do not own all the data represented in a particular dataset. Few such groups have the expertise or time to devise and negotiate appropriate licenses which would enable them to release datasets to third parties.
- Many people to whom we spoke did not believe people would want access to their datasets. They think that typical small scale, highly-specific projects yield data that will not be of interest to anyone else. It is also thought that, in many cases, people do not want access to raw data; instead they prefer datasets where some degree of value has been added.
- And yet there are some project teams which plan to make their data available to third parties not just via a data centre but via their own websites. These tend to be projects where the dataset is the main output rather than publications based on it.

Discovery, access and use of third party datasets

Discovering relevant datasets

In the subject themes covered by the RELU programme the relevant big datasets are well known to researchers. The RELU DSS also provides a list of information resources relevant to the programme's themes. Researchers seeking more specific datasets use their personal networks together with attendance at conferences and other relevant meetings. They are also likely to search the web. No-one to whom we spoke was dissatisfied with the discovery tools at their disposal, but in most cases researchers were producing primary data and are not especially reliant upon finding datasets that are not already known to them or their network of peers.

Access to third party datasets

Although the majority of the projects we looked at were producing primary data of different types, many needed access to one or more external datasets with which, as mentioned above, they are normally familiar. These tend to be the large, comprehensive datasets typically produced by organisations such as the Ordnance Survey, the Meteorological Office, the Environment Agency and DEFRA. Most of these operate along commercial lines, charging for access to data. The charges can be significant. Access to the Point X dataset produced by the Ordnance Survey was cited as an example of an expensive dataset where even with higher education discounts the charges can run to tens of thousands of pounds. In fact the cost of accessing important nationwide datasets was highlighted as a particular problem for award-holders. Within its remit the RELU DSS has investigated group licensing for some of the more popular but expensive datasets, but the negotiation of any

discounts for award holders needing to access such data will take time, and may not be realised within the life of the current Programme. Researchers clearly had high hopes for this idea and some even thought the DSS would be in a position to pay for such access. However, as yet, this ongoing work has not yet achieved any specific licensing for RELU researchers. All researchers were thus advised to budget for full purchase costs of data in their research submissions.

Use of third party datasets

As well as the major datasets produced by public sector bodies many RELU award holders have a need for third party data as a key building block to their own work. These may be economic data, census data, medical records or other public health longitudinal data produced by the main cohort studies. Sometimes researchers purchase very specific data produced by commercial organisations, such as those who regularly collect information from farmers. Normally people want to use processed data, typically from big datasets; it is rare for researchers in this programme to want to access small scale raw datasets produced by other researchers.

Quality assurance

Quality assurance in the data creation process

In general researchers believe their peers conduct their research and data collection to a professional standard. There is an interesting additional element to this programme because each funded project is interdisciplinary in nature. It is clear that researchers from different disciplinary backgrounds are left to do their own part of a project; people tend not to involve themselves in data collection being carried out by colleagues from a different discipline. The different disciplinary strands of a project are brought together at project meetings and workshops as well as through more informal means of communication. This serves to highlight that research professionals trust that their colleagues from different disciplines are applying appropriate levels of quality assurance. It would be unusual, for example, to hear of a social scientist checking over the laboratory notebooks of a natural scientist.

Data management planning

The quality of data management planning is highly variable. Relatively few researchers have taken advantage of the expertise offered by staff at the RELU DSS though many were mandated to attended DSS training workshops at the outset of their awards. Quality appears to be a function of the expertise within a project team but also the inclination of the researchers on the team: there are reportedly many researchers who are not convinced of the benefits of data management and fail to engage with the data management planning process in any meaningful way. RELU award holders are expected to produce high quality data management plans as a requirement of their contract. These are reviewed and evaluated by DSS, and only signed off if they are lacking in detail or quality or clearly do not address data sharing adequately. Substandard plans must be revised and submitted after a discussion between the PI and DSS. Across subsequent Calls 1, 2 and 3 of the RELU DSS programme, it was noted that later data plans were of higher quality, perhaps reflecting the insistence of the Programme Manager to take the matter seriously. It might be a useful exercise to compare the quality of a project's initial data management plan

(before revision), its data management processes during the project, and the quality of the final dataset offered to the data centre. Does poor quality data management planning inevitably lead to poor quality datasets? If a causal link can be established there may be a case for raising the bar in terms of funders' requirements for high quality data management plans, perhaps as a condition of releasing project funding. Interviewees have indicated that the current requirements are fairly basic.

Quality assessment of datasets

There is little appetite among our interviewees for external quality assessment of datasets in terms of their scholarly content (rather than format, metadata and so forth). Of course people expect the quality will be of a professional standard and that people's publications will reflect that quality of the underlying data, but more than then there is a concern that any external assessment process will divert money from funding research and will entail researchers surrendering even more of their time to review datasets which, in the view of many, are not very likely to be re-used by others in the field.

What motivates researchers to publish data?

Push factors: the effect of policy

This glimpse into the realm of RELU-funded projects has been interesting not least because award holders were aware from the very beginning that they were required not only to produce a data management plan that was acceptable to the RELU DSS and the programme managers, but also they were required to sign up to the RELU Data Management Policy. In addition, award holders have access to a dedicated Data Support Service whose staff are available to provide advice and guidance on all aspects of data management. Has this particular approach and investment in infrastructure yielded substantive results? Well, bearing in mind the limited extent of our investigation, the answer is somewhat equivocal: there have been some positive results but overall the effectiveness of this approach could be described as not having achieved its full potential. The cultural divide between funders [who want datasets to be made available for other researchers to use] and researchers [many of whom might agree with the sentiment of data sharing but are not convinced that they personally should be the ones doing the sharing for a wide range of reasons] clearly exists. The approach of encouragement and education taken by the RELU programme has proved useful in terms of beginning the process of bridging the divide but in the absence of effective sanctions for those researchers who choose not to engage with the spirit of RELU's Data Management Policy its effectiveness is going to continue to be limited.

Beginning with the positive aspects, a number of people said that being required to consider and produce a data management plan focused their thoughts on the data-related aspects of their project in a formal way, a process to which few were accustomed. People report that many of their colleagues have been receptive to the idea of data management planning. Indeed we received the impression that the effect of having to produce a data management plan will have a long term effect on the approach to data management planning taken by a number of principal investigators for future projects. This accords well with the programme's approach of encouraging researchers to appreciate the benefits of effective data management throughout the lifecycle of research projects. Some project teams do

actively engage with data management experts at the RELU DSS; personal visits by DSS staff to offer guidance and advice to project teams have been appreciated by those teams which have chosen to take advantage of this resource.

Turning to the less positive perspective, it is possible to argue that the effect of RELU's policy efforts have been muted on the whole simply because researchers are not personally persuaded of the benefits of investing time and resources in data management. They might agree with the argument that publicly-funded research should be made available to others and that the steps taken to oils the wheels of knowledge transfer between disciplines are worthwhile, but some are adept at coming up with reasons why their particular case is worthy of exemption. The key difficulties are listed below:

c. Some researchers hold the general view that the prevailing culture in the part of the research community covered by the RELU programme does not value data management. Some researchers are reported to hold the view that the requirement to produce a data management plan is but a bureaucratic hurdle that needs to be dealt with as expeditiously as possible. Many project teams did not avail themselves of the opportunity to access the data management expertise offered by the staff at the RELU DSS.
d. Some researchers – even within the same project team – are ambivalent or uninterested in matters to do with data management. Others don't see the point in producing a data management plan at the beginning of a project because they believe the research focus invariably shifts as a project progresses, to the extent that data management planning done in the early stages is bound to be redundant later.
e. In parts of the social sciences issues of ethics, consent, confidentiality and data protection are often invoked as reasons why data cannot be shared. These issues could be adequately addressed with better data management planning, appropriate anonymisation techniques and licensing controls – but funders need to be very clear at the outset what their data sharing requirements are so project teams can plan their work in a way that enables them to comply with these requirements.
f. There appears to be a general view that the data outputs from "small science" projects – that is, very specific, small scale studies - are somehow not worth keeping and that the data are not likely to be used by others. In fact it has been said that researchers see data management issues as being separate from the science (and therefore not really something they should need to be worrying about).
g. The traditional view that research projects are set up to produce publications still prevails; researchers seem not to be interested in the long-term fate of their datasets. The fact that datasets are required to be offered for curation under the terms of the RELU Data Management Policy doesn't' appear to influence people's attitudes to data management. As with the requirement to produce a data management plan, for some it is simply a bureaucratic hoop to jump through. In fact with compliance running at 80-90% most people do jump through the hoop, but since the requirement is simply to offer datasets there is no incentive for people to offer them in a state whereby they are suitable for curation.

To summarise this section, the RELU Data Management Policy has had a positive effect for many researchers involved with the programme. There are a number of researchers for whom the discipline imposed by the requirement to produce a data management plan has been beneficial. But counterbalancing the group of researchers who embrace the spirit and goals of RELU's Data Management Policy, there are others for whom it is an inconvenient obstacle on the funding trail. The approach of encouraging researchers to consider data management seriously appears to work but only for individuals prepared to change their traditional ways of working. Changing the overall culture in respect of data management and sharing may require a harder edge to funders' policies.

Pull factors: intrinsic rewards

It would be a fair reflection of our insights into the RELU programme to conclude that publishing or sharing datasets does not feature near the top of people's list of priorities. When pressed, people see the recognise the principle behind the notion of publishing datasets, but there is a disjunction between this general view and their own particular experience: few people see the benefit to themselves of sharing data, cultural precedents for data sharing are weak, and in many cases researchers don't think their datasets will be of interest or use to other people (except when they are producing large scale datasets). In a sense, the requirement to offer datasets to a data centre at the end of a project may in fact reduce the need to think about why data sharing might be important for knowledge transfer and the good of science; many simply do it because it's a box that needs to be ticked to fulfil the funder's requirements. There are some people who are motivated to share their data for primarily altruistic reasons, but for most researchers datasets exist as a basis for publications from which the true career rewards is gained.

CLIMATE SCIENCE

Overview

Climate scientists create and use large volumes of data. The Natural Environment Research Council (NERC) gives a clear lead on how data should be treated, expounded in the NERC Data Policy Handbook[57], and funds a network of data centres[58] to look after important datasets for the long term. The goal of NERC with respect to data is very clear, as demonstrated by this extract from NERC's rationale for its data centre network: *It is essential that data generated through NERC supported activities are properly managed to ensure their long-term availability. Our network of data centres provide support and guidance in data management to those funded by NERC, are responsible for the long-term curation of data and provide access to NERC's data holdings[59].*

While NERC has evidently taken important strides in setting up a clear policy framework with respect to data management, curation and the necessary technical infrastructure, a proportion of datasets do not become publicly available. Although climate scientists who are funded by NERC are normally required under the terms of their grant conditions to offer their data to the relevant data centre for curation, NERC's data centres do not have the capacity to take everything and are not obliged to do so. Judgements need to be made by each data centre about what should be curated, primarily in terms of the usability and usefulness of datasets. Datasets that do not fulfil the criteria for inclusion in one of the data centres tend not to be made publicly available via any other route.

It is one thing for datasets to be – in theory at least – available for the academic community to use. It is often something quite different for climate scientists and others to be able to find and then access datasets that may potentially be useful to their work. This study has shown that people's approaches to information discovery is somewhat unsophisticated and tends to be based mainly on their personal network of peers, sources they know through experience, or references to datasets from published papers. NERC is addressing the issue of data discovery through its investment in the NERC DataGrid[60]. This discovery gateway will provide access to thousands of datasets in addition to tools designed to help with the processing of these datasets. Integral to the development of the DataGrid is the development of the Climate Science Modelling Language (CSML) information model, toolbox and data services. These initiatives should help set the conditions necessary to facilitate greater use and sharing of datasets in climate science.

[57] NERC Data Policy Handbook, http://www.nerc.ac.uk/research/sites/data/documents/datahandbook.pdf

[58] Links to NERC's network of data centres can be found here: http://www.nerc.ac.uk/research/sites/data/

[59] http://www.nerc.ac.uk/research/sites/data/

[60] http://ndg.nerc.ac.uk/index.htm

Data Creation

Form and variety of data produced

It is possible to characterise the types of data produced by climate scientists in four broad categories:

- First, large volumes of data are produced by climate models. The model output data (the base data) are often processed and analysed.
- Second, data from long-term monitoring and observation work. Such data are often used for comparison with climate model data. These data may also be the result of *in situ* or *in vitro* studies which tend to require more complex metadata than the metadata typically associated with model run data.
- Third, there is a group of data which is variously termed historical; recreated; derived or proxy data.
- Finally, graphical representations or visualisations of data may be construed to be a fourth type of data. Typically these data are an amalgamation of the three types of data listed above.

The climate science community is well used to creating and working with data although, as with researchers in all subject areas, there is room for improvement in terms of producing good quality, consistent metadata. While NERC does not explicitly require the researchers it funds to produce a data management plan, it does require them to consider the long term viability of the data they produce. Other funders, notably those at the European level, do require researchers to focus on data management plans, but conversely they have in the past put less emphasis on the curation of datasets. In theory data centres stand ready to give advice on data management issues to researchers but, in reality, requests for this type of advice are the exception rather than the norm.

Climate data formats are well defined and there are tools available to convert the data between different formats. Formats that are widely used by climate scientists include Hierarchical Data Format (HDF)[61], the Network Common Data Format (NetCDF)[62] and the NASA Ames format[63]. The Meteorological Office tends to use binary data formats.

Data produced by climate scientists may be subject to many different stages of transformation from being raw data through to final datasets which, in theory, are in a form suitable for public dissemination. It is thought unlikely that data in these intermediate stages will be available to third parties.

Metadata

Many climate science researchers spend little or no time thinking about metadata. In an environment where time is the main limiting factor, metadata are clearly not a priority. Although there exist sound standards for metadata – and CSML looks set to become a robust and useful standard - their adoption is limited. This may be because there is a lack of tools to create or exploit metadata, or it may be because metadata are mainly useful for sharing datasets with others, not something essential for researchers who are analysing their own datasets. Given that the sharing of datasets between researchers or research

[61] http://www.sesp.cse.clrc.ac.uk/Publications/data-management/report/node13.html
[62] http://www.unidata.ucar.edu/software/netcdf/docs/faq.html#whatisit
[63] http://badc.nerc.ac.uk/help/formats/NASA-Ames/

groups is limited in scope and scale, there may simply be little or no incentive for the creators of datasets to worry themselves about metadata.

Insofar as metadata are used, it is possible to characterise this use in three key ways. First, raw data may have basic or standard metadata, often associated with the use of one of the common data formats. Second, to facilitate quality assurance and data sharing it is important to produce preliminary provenance metadata. Standards are emerging which will make this task more straightforward. Finally, towards the final stages of the lifecycle of a dataset – after transformations, analysis or visualisations – the main form of metadata is text.

Adding value to data

Climate science data may be classified according to its stage in the research process in the following manner: raw data; analysed data; collated data; multi-source data; presentation data and, finally, published data. The characteristics of data vary according to these stages. Normally data near the end of the research process are perceived to be the most valuable, having been processed to reach that stage. Model run data in the rawest form are unlikely to be especially useful to anyone other than the creator; normally such data need some sort of processing to make them useful to others. Observational data, on the other hand, can be very useful to people in the original raw state – assuming the metadata are sufficient to enable users of the data to understand how they were collected. In the field of palaeo data, for example, raw data are unique and, since they are potentially valuable for a long period of time and need to be properly curated.

Long term viability of datasets

It is generally agreed that important datasets should be looked after for the long term, but climate scientists are realistic about the need to trade off the cost of long term curation with the value and ageing profile of those datasets. Model run data, for example, are widely assumed to have a maximum useful life of five years. The exception to this is where large scale or high resolution models are run; these are expensive, can take a long time and consequently remain viable for longer. Raw data born of observational or remote sensing techniques are unique in that they represent properties with fixed temporal and spatial dimensions, and they can be very expensive to collect. It is widely believed that these types of data should be curated for the long term.

Working with the created data

Tools and technologies for analysis

Climate scientists use a wide variety of tools and technologies for data analysis. In the modelling community the emphasis is on building and refining computer models and doing many model runs with slightly different permutations each time. For the biggest models these can require significant computing power, which is why they can only be run by a limited number of organisations which have the appropriate resources - the European Centre for Medium-Range Weather Forecasts for instance. There is also a requirement to use analytical tools to enable the visualisation of model run data.

"Publishing" datasets

On the whole, climate science researchers give little thought to making their datasets available to a wider audience. There is little in the way of incentives to do so and, for many, their training and experience does not equip them to "publish" their data. There are other forces at play which will be discussed later but for the present, with a few exceptions, for the majority of climate science researchers disseminating their datasets tends not to feature near the top of their "to do" lists. What most will do, however, is comply with grant conditions which required them to offer their datasets to the appropriate data centre for curation.

Ownership of data and constraints on publication and use

The NERC Data Policy is clear on the issue of ownership, making it plain that ownership of datasets often resides with the organisations that have funded the project (such as NERC) or the creators' employers (HEIs, for instance). The policy underlines the obligations of data creators to ensure the rights of the owners of the data are not compromised. This may involve the use of formal licence agreements. There is a clear desire on the part of research funders not to give away rights that may subsequently be exploited by third party organisations.

Although NERC requires grant-holders to offer a copy of their final datasets together with the appropriate metadata and other relevant documentation for archiving in one of the NERC data centres, it does not necessarily require the transfer of the Intellectual Property Rights in the datasets from the owners of those rights. The data centres will ensure the rights of the owner of datasets are respected through formal means such as a data protocol, the terms of which academic users of the datasets are required to observe. It is worth noting that datasets and their owners are subject to the requirements of legislation such as the Data Protection Act 1998 and the EC Directive on the Freedom of Access to Environmental Information.

Response to requests for datasets

Researchers rarely request model run datasets from other researchers or research groups, though there is demand for processed, higher-order datasets. Some – though by no means all researchers – will try to oblige by sharing their datasets where they are not otherwise available from one of the data centres. The quality of the response is reported to vary. For instance, while people might be willing to locate and send a data file, without the appropriate metadata and in some cases specific analytical or visualisation software, the value of the raw data is limited. In the cases of sample materials and observational data, there is a general sense that these should be more readily available to researchers since, at present, only a proportion of such datasets are held by national data centres. NERC is currently consulting with the palaeo-data community to determine how best to manage and share the important datasets they produce.

Discovery, access and use of third party datasets

Discovering relevant datasets

Finding relevant datasets is no simple matter and researchers' discovery behaviour is not necessarily systematic or optimal. We have adduced five general categories of behaviour with respect to finding datasets:

- First, in some areas of climate science research there are few sources of relevant datasets and they are well known to researchers.
- Second, researchers turn to their network of peers to get help in finding or identifying datasets that could be useful to their work.
- Third, published papers are widely regarded as signposts to datasets. Many papers describe the datasets upon which they are based, and people contact authors to see whether they could gain access to the dataset. Whether or not they get a response, let alone access to the dataset, is a hit and miss affair.
- Fourth, people may use the search facilities offered by the NERC-funded data centres or their equivalents abroad, together with some specialised climate science metadata searching websites such as ACCENT.
- Finally, apparently when all else fails, people will turn to a generic search engine such as Google. This is regarded by many as a sub-optimal method because no contextual information is provided about the hits to help researchers decide whether or not to pursue a particular line of enquiry.

These various approaches are the main tools used by climate science researchers, but they tend not to be used in a systematic or hierarchical way. The NERC DataGrid, currently under development, seeks to address the issue of dataset discovery by offering a relatively sophisticated, wide-ranging mechanism for searching out relevant datasets. The ability for researchers across the climate science community to discover datasets more effectively may well encourage people to share data more readily.

Access to third party datasets

Discovering a dataset that might be useful is just the start. Even when researchers identify a dataset they think might be useful to their work there is no guarantee they will be entitled to, or be able to negotiate, access to it. The creators of datasets are not always willing to share access to them. In cases where the dataset has been produced by a collaboration of different people from different organisations, there is likely to be a requirement to negotiate a licensing process to ensure none of the rights of the owners of the dataset are compromised. Researchers who want to use data from one of the big providers, such as the European Weather Centre, must have a licence to do so.

The extent to which datasets are accessible to the research community depends on the nature of the infrastructure that exists to facilitate access. Our investigations indicate that there are five broad categories.

- First, NERC has invested in a formal network of data centres and associated specialist centres and these offer the possibility to access a large number of climate science datasets.
- Second, there exist a number of national or international organisations or specialist centres from which it is possible for researchers to access datasets. Examples include National Oceanography Centre, Southampton (NOCS), European Centre for Medium-

Range Weather Forecasts (ECMWF), EUMETSAT, Program for Climate Model Diagnosis and Intercomparison (PCMDI) and the World Climate Data Program (WCDP). These are well-managed, long-term funding assures their stability, and they provide metadata and tools for data discover, data re-formatting and data delivery.

- Third, some (very few) climate science departments in the UK have websites from which users can access datasets. These may not be searchable and there are no guarantees about their availability over the long term.
- Fourth, many UK climate research projects look after their own data archives. These are primarily for internal use, though sometimes external access is available (on a request basis for instance).
- Fifth, many researchers resort to storing their datasets themselves either on a personal disk or an institutional server. It is not normally possible to access these without making a personal approach to the creators of the datasets.

Use of third party datasets

In general the demand among climate science researchers for datasets from other researchers or research groups is, at present, very limited. Where researchers make their model simulation data publicly available downloads of datasets or requests for particular sections of datasets, tools, algorithms or help are infrequent. Climate science modellers, it appears, rarely look at other modellers' raw data. Few people want raw data; if they request data at all they normally ask for processed or derived datasets. There is a higher level of demand for datasets that contain monitoring, remote sensing or observational data, not least since these are often used to check the veracity of model simulations. But perhaps the greatest demand for third party data is for large datasets that are beyond the capacity or capabilities of researchers to produce themselves. These are mainly large national or European datasets including, for example, high resolution model simulations that may take a lot of money and time to run.

Quality assurance

Quality assurance in the data creation process

Climate science researchers in the UK clearly believe that there are few if any problems with quality assurance in the data creation process.

Data management planning

Although the quality and veracity of the content of datasets produced by climate scientists in the UK is widely regarded as being more than sufficient, the quality of metadata and documentation is patchy. Some researchers are careful to produce useful metadata and methodological explanations; others pay little or no attention to such matters. Researchers are currently free to request advice from NERC data centres about data management, and staff at those centres are willing to provide help and guidance, but up to now climate science researchers have not been required by NERC to develop and submit a data management plan as a condition of being awarded a grant. This position has been under review and is likely to change in the near future. Effective data management plans can form the foundation for the use of appropriate formats but, importantly, adherence to accepted data management standards and the creation of metadata means that more datasets become more accessible to a greater number of people.

Quality assessment of datasets

Most researchers in the climate science community appear to believe that producing datasets of an acceptable quality is the province of dataset creators themselves. As trained scientists with experience in their specialised subject area they are best placed to ensure that the data they produce conform to accepted standards. All the climate scientists to whom we spoke trust that datasets they obtain from peers for use in their own work are of an acceptable quality. The exceptions are when datasets are being sought from research groups not known to the requesting researcher, and when dataset creators are based in countries that are not known for scientific training and rigour.

When asked whether datasets should undergo formal quality assessment, most researchers believe that would be a step too far. They are already over-burdened with requests to review papers submitted to scholarly journals, without having to review datasets. Instead, climate scientists are content with the current system whereby peer review of journal articles is, by proxy, a review of the methods and therefore reliability of the science behind the creation of the datasets that underpin those articles.

There is also an important connection between quality assurance and the longevity of data. A dataset and its associated metadata may be of the highest order when it is first made publicly available, but there is no question that data have a definite shelf-life. Model run data have a life of up to five years; observational datasets are longer lived because they have value because they provide a unique record relating to a specific time and place. The ageing of model-run datasets comes about for reasons such as refinements to climate models, the re-calibration of satellite data, the development of new algorithms to transform proxy data, and the periodic re-analysis of important longitudinal weather datasets. As model-run datasets become outdated, so their inherent value diminishes.

What motivates researchers to "publish" and share data?

At a conceptual level the idea of data sharing for the greater good of climate science and, by extension, the UK's knowledge economy, is accepted as being worthy by many. There are also some national and international research projects which are having the effect of encouraging the sharing of datasets – projects such as the creation of consistent re-analysis climate datasets (ERA or NCEP) or major climate model inter-comparison projects (IPCC). For the most part, however, the publication of datasets for the purpose of sharing is very patchy. The sharing of datasets is more prevalent among climate science researchers than within the weather data community where commercial imperatives pertain, but even among climate scientists there is no dominant culture of data sharing across the board. Instead, researchers in different parts of climate science exhibit different attitudes to sharing: climate modellers tend not to share (often because they have no particular need of other modellers' datasets) whereas sharing is more common in ocean modelling. Researchers dealing with observational data do tend to share data, and there are often data sharing collaborations between the modelling and observation data communities.

At present, data sharing is based largely on expediency mainly because, for climate scientists working on regular projects, there are few if any explicit rewards for publishing datasets. In fact the impediments to doing so appear to outweigh the potential benefits that

may accrue to the creators of datasets by making their data available. There are costs in finance and time to producing datasets that are capable of being shared; that is, datasets that are properly formatted with good quality metadata and supplied with all the necessary contextual information including computer syntax or other tools required for other people to re-use those datasets accurately and effectively

Push factors: the effect of policy

In the absence of any tradition for data sharing across the breadth of climate science, the actions taken by NERC, the main funding organisation for the field in the UK, are of key importance. The network of designated data centres facilitates curation of datasets and access to them, while NERC's data policy requires the datasets resulting from projects it funds to be offered to the NERC data centres. In the near future, award holders will also be required to develop a data management plan as part of their funded project. These policies have the effect of influencing researchers' behaviour: if climate scientists apply for NERC grants they know the data-related conditions they are signing up to. There is no guarantee that the data centres will choose to archive all the resulting datasets, but at least they have the opportunity to accept those they perceive to have long term value. It remains to be seen whether these policies will have the effect of changing researchers' deep-seated reluctance to proactively share data. Many may still need to be persuaded of the benefits of doing so. There are benefits to be had, as we explain below.

Pull factors: rewards

There are few climate scientists who have taken to sharing their project-level datasets by developing their own websites for the purpose. It is, we are told, a complex and time-consuming task, one that until now has been largely unfunded.

There is work underway now, though, that is designed to explore the idea of developing two-way citation links between traditional publications and published datasets. The CLADDIER[64] project is currently experimenting with linking publications in institutional repositories with datasets held at the British Atmospheric Data Centre (BADC). In due course this initiative may benefit climate scientists who publish their data in terms of increased exposure and citations.

[64] http://CLADDIER.badc.ac.uk/