Data Dimensions: Disciplinary Differences in Research Data Sharing, Reuse and Long term Viability

A comparative review based on sixteen case studies

Key Perspectives Ltd

With a foreword and recommendations by Chris Rusbridge and Liz Lyon

DCC SCARP SYNTHESIS REPORT

Commissioned by the Digital Curation Centre.

Deliverable B4.11

Version No. 1.0

Status FINAL

Date 18 January 2010

Copyright

Catalogue Entry

Title Data dimensions: disciplinary differences in research data sharing, reuse and long term viability
Creator Key Perspectives Ltd (authors)
Subject Data curation; formats, processes and issues; system development; standards; legal factors; methodology, and problems overcome; human factors
Description This synthesis report has been produced for the Digital Curation Centre (DCC) SCARP project, funded by the Joint Information Systems Committee (JISC) to investigate disciplinary attitudes and approaches to data deposit.

Date January 2010 (creation)
Type Text
Format Adobe Portable Document Format v.1.3
Resource Identifier ISSN 1759-586X
Language English

Citation Guidelines

Key Perspectives. (2010), "Data dimensions: disciplinary differences in research data sharing, reuse and long term viability. SCARP Synthesis Study", Digital Curation Centre, Retrieved <date>, from http://www.dcc.ac.uk/scarp

Foreword

This synthesis study, commissioned by the Digital Curation Centre from Key Perspectives Ltd, forms a major output from the DCC SCARP Project, which investigated attitudes and approaches to data deposit, sharing and reuse, curation and preservation, over a range of research fields in differing disciplines. The aim was to investigate research practitioners' perspectives and practices in caring for their research data, and the methods and tools they use to that end. Objectives included identification and promotion of 'good practice' in the selected research domains, as expressed in DCC tools and resources. The approach combined case study methods with a survey of the literature relevant to digital curation in the selected fields.

The resulting case studies, each with recommendations and findings for the research groups and for the range of stakeholders in digital curation, demonstrated that curation tools, such as the Digital Curation Lifecycle Model, were useful provided effort was applied to adapt them to the context of 'live' data creation and curation. Participating research groups generally did not have formalised curation or preservation processes, so the case studies aimed to understand expectations, risks and constraints, and find appropriate ways to build on current capabilities.

This synthesis report (which drew on the SCARP case studies plus a number of others, identified in the Appendix), identifies factors that help understand how curation practices in research groups differ in disciplinary terms. This provides a backdrop to different digital curation approaches. However the case studies illustrate that "the discipline" is too broad a level to understand data curation practices or requirements. The diversity of data types, working methods, curation practices and content skills found even within specialised domains means that requirements should be defined at this or even a finer-grained level, such as the research group.

The main body of this report (excepting this foreword, the overview sections and conclusion) is faithfully focused on the content of the case studies. To maintain the integrity of the process of synthesising the case study reports, the authors have refrained from making comments, judgements, implications, references to more recent developments, or changed subject-specific terminology in the main body of the work. This is also why there are (unusually) no references: the main body of the work is derived solely from the case studies.

Building on this work, as Directors of the Digital Curation Centre and the DCC SCARP Project, we believe the project and this report suggest some important Recommendations, which we include here for consultation.

Many people worked on the SCARP project, and it is impossible to credit them all. However, we would like to thank Sheridan Brown and Alma Swan, who have produced this Synthesis, and Esther Conway, Elizabeth Fairley, Colin Neilson, Jenny Ure and Angus Whyte as lead authors of the SCARP studies and reports.

Draft Recommendations

JISC

1) The JISC Managing Research Data Programme is an important contributor to developing practice in data curation throughout the sector. We strongly recommend that JISC regard this change management initiative as a priority to be developed further as soon as funding permits, and take specific account of discipline issues.
2) JISC should commission a study to map a wider range of discipline (or sub-discipline) curation requirements in depth, to inform this further work.

Higher Education and Research funders

3) Mandates for data sharing and curation are important and valuable in the cultural change process: such policies should be developed by research funders in consultation with researchers in their various disciplines.
4) Formal recognition of the value of curated data as research outputs is necessary and would be an important signal to institutional leaders and researchers. We recommend that all those involved in the assessment of research should be explicit about the position and value of curated data in such assessments.
5) We strongly recommend that national-scale research data support services be defined and supported: for example, a development from the UKRDS Feasibility Study, or a service similar in scope to the Australian National Data Service. Such developments need to take specific account of discipline-related issues.

Publishers and Learned Societies

6) Publishers and Learned Societies should be more specific about their requirements for lead authors/researchers to be responsible for ensuring the retention of their data, and for making it available for re-use under appropriate conditions.
7) Publishers should ensure that Supplementary Data are openly available in forms and under conditions appropriate to the disciplines, and suitable for re-use.
8) Publishers should promote direct citation of datasets used in research underlying the articles they publish.
9) Learned Societies should use their key position in relation to researchers, as agents for change, and support their disciplinary requirements.

HEIs and Research Institutions

10) Institutions have a vital role to play in supporting curation of data produced by their researchers, and in managing the medium to long term retention of some of those data. It is critical that institutions take a lead in this area: all institutions with a research focus have a responsibility to address these issues.
11) Institutions should identify how data curation will be supported for their researchers. This could be through local services (such as the Library and IT services), through regional partnerships, or through external services such as data centres. In taking these decisions, institutions should take account of discipline needs and resources.
12) Institutions should explicitly identify funding to support data curation development, infrastructure, staffing and training and make full use of the resources provided by the DCC.

Researchers and Scholars

13) Researchers should work with colleagues and their discipline community to develop selection/appraisal criteria to identify priority data.
14) Researchers have a responsibility to curate their data; where data supports another research output such as a journal article, persistence may be assured by transferring data to an appropriate data store or archive .
15) Researchers should seek to reach consensus on appropriate standards and formats for their discipline or sub-discipline in partnership with relevant professional bodies, in order to enhance data discoverability and re-use. .

Chris Rusbridge, DCC
Liz Lyon, DCC and UKOLN

Contents

INTRODUCTION

The SCARP project, directed by the Digital Curation Centre, was conceived to investigate disciplinary attitudes to data deposit, sharing and reuse, curation and preservation. A series of immersive case studies in selected disciplinary areas have been conducted. The purpose of this study was to review the SCARP studies, together with relevant case studies conducted or commissioned elsewhere, and to bring together the various insights from the case studies into a single coherent report. Sixteen different case studies have been explicitly reviewed, ranging from audits of single departments or projects in individual disciplines to overviews of researchers' attitudes and infrastructure arrangements in discrete disciplines and cross-disciplinary fields. Synthesising disparate case studies with their different styles and structures such as these is not straightforward, but there is value to be found in presenting the key findings under the following common headings: data sharing (including insights into the culture and practices that prevail in different disciplines); data discovery, access and reuse; and long term care and preservation of data. Here "data sharing" is used to describe the act by a data creator of making their data available for others to reuse. This is the natural pre-condition for data reuse and has to do with the physical aspects of deposit but also with peoples' attitudes and behaviour with respect to the value of their data to others through time. The two concepts - sharing and re-use - are closer together in some disciplines than others but it is useful here to differentiate them. If researchers choose not to share their data, others cannot reuse them.

It is clear that researchers' attitudes and practice with regard to the creation, sharing, reuse and long term care of data are closely linked to the discipline in which they work. The fact that these disciplinary differences exist, and that they exist on a variety of technical and behavioural levels, presents a significant challenge for those who may become responsible for curating the breadth of research data produced by a single institution. In some disciplines the capacity of national and smaller data centres is insufficient to accommodate the volume and variety of data being produced by researchers. A growing number of people hold the view that individual institutions should be responsible for curating the data produced by their own research community where those data have no natural home. If institutional repository administrators are to be given the responsibility for curating their institution's research data outputs they will need to develop domain-specific strategies since, as this project demonstrates, a generic approach to data curation will not be sufficient to cope with the different data-related needs and expectations of researchers working in different disciplines other than at a superficial level.

Overview

Data Sharing Overview

Among the research community at large there is growing recognition of the merits of data sharing in principle and in practice. Many research funders too, particularly those which disburse public funds, are pushing the data sharing agenda in recognition of the potential value to be gained from encouraging researchers to make their research data available for others to find, review and reuse,

possibly in ways that had not been imagined when the data were originally collected. New software tools and techniques that facilitate data mining and re-engineering of data are already yielding new scientific insights in some disciplines.

It may seem intuitively clear, but the evidence of the case studies reviewed for this project demonstrates that there exist strong disciplinary differences in terms of researchers' attitudes to sharing the data they produce. Their attitudes are influenced not only by structural limitations on sharing data (such as the lack or inconsistent application of metadata standards or unequal access to data storage facilities) but also by the cultural heritage of individual disciplines. In this section we draw on the available evidence in the case studies to highlight disciplinary differences with respect to data sharing.

Data Discovery, Access and Reuse Overview

There are stark disciplinary differences in the existence and application of technical standards for describing and storing data. Some disciplines benefit from well-established, discipline-wide standards. For many disciplines, on the other hand, there exist no commonly agreed standards. This matters because data that is well structured and described with sufficiently rich and appropriate metadata is easier to discover. Such data may not necessarily be easy to access or to reuse since sometimes datasets are secured behind access barriers. Barriers may be overcome by monetary payment or by agreeing to mechanisms that mitigate issues to do with ethics or confidentiality, but sometimes they prove insurmountable. In some disciplines, however, researchers are untroubled by such matters and benefit from being able to discover, access and reuse well-curated data from publicly-funded, open access databanks or data centres: astronomy, crystallography and genomics are good examples.

It is hardly surprising that different technical standards have evolved to reflect the types of data peculiar to different disciplines. In climate science, for example, data formats are well defined and there are tools available to convert the data between different formats. The Climate Science Modelling Language is a standards-based data model and Geography Mark-up Language application schema for atmospheric and oceanographic data and is already proving its worth in climate science, but it is obviously domain-specific. This example highlights the problem that will confront people charged with curating data in disciplines with which they are unfamiliar. Will institutional repository administrators in a university setting be willing or able to comprehend the details of data formats and metadata schemas across a whole range of disciplines? And for disciplines where such standards have yet to evolve – in the classics for example – the absence of standards may simply add to the overall challenge of curating data across a spectrum of disciplines.

If the true potential of reusing data is to be realised, researchers need to be able to discover relevant datasets. In some disciplines the main sources of data are few and well-known to the researcher community. From the cohort of case studies we have looked at, astronomy, genomics and some disciplines in the social sciences certainly fit this category. In other disciplines researchers may turn to their peers for help in identifying or finding datasets that could contribute to their work. Published papers are sometimes useful signposts to datasets, and journals in fields such as systems biology regularly flag up links to underlying datasets. People may also use the search facilities provided by their disciplines' main data centres, so social scientists might

conduct specific searches on the UK Data Archive's own website and, before it was closed in 2008, researchers in relevant disciplines could have contacted the Arts and Humanities Data Service for guidance. Finally, researchers may resort to the use of generic search engines such as Google or Yahoo to see what they can find. This can be a sub-optimal and laborious method mainly because no contextual information is provided about the search results to help researchers decide whether or not to pursue a line of enquiry.

Data Preservation Overview

Digital data preservation may be defined as the management of digital data over time in order to maintain the intellectual record. Most researchers would probably not dispute the theoretical value implicit in preserving data to validate work and for reuse but, given that it needs considerable resources to adequately preserve data, the question arises as to which data are worth preserving and for how long. The answer will be different in each discipline but, in general, peer review of datasets produced by small to medium-sized research projects is uncommon. In astronomy nearly all observational data regardless of when they were collected are thought to be valuable; astronomers need the ability to track back through the record. The MST dataset, for example, is an irreplaceable earth observation record; once lost it cannot be replaced by the repetition of an experiment or model simulation. In systems biology researchers bemoan the fact that sometimes datasets are overwritten with new or updated data, depriving researchers of the opportunity to reproduce earlier findings. In some disciplines, however, the requirement to preserve data for ever is less clear cut. In climate science while observational data is deemed valuable and unique enough to be worth keeping, climate modelling data is thought to have a useful life of around five years and, even so, such data are rarely in demand from researchers beyond the team which created the data in the first place.

It is plain that the drivers and potential for data preservation will vary to a significant extent between disciplines and those responsible for mapping preservation strategies will need to take account of the volume, structure, quality, uniqueness and current and potential scholarly value of data within each discipline. Whether or not such data can, in fact, be preserved will depend upon the existence or development of appropriate data formats, metadata schema, legal and ethical impediments and storage infrastructure together with the expertise and funding required to manage the preservation process. As this report shows, some disciplines are already well positioned in this respect; others have a mountain to climb.

The cost of inaction is the loss of datasets which may or may not have had a role to play in the advancement of scholarship. It is, in economic parlance, an opportunity cost - though it is difficult or impossible to quantify that cost. One case study, Presto Space, focused on the difficulties of preserving audio visual materials – films and television programmes – due mainly to the monetary costs of preservation and the enormous task of migrating data to modern storage media. The study estimates that, at current rates of data migration, at worst 70% of tape-based content will be lost by 2025; the best case scenario is for 40% of tape-based content to be lost by 2045. The Presto Space project offers guidance on how to calculate the value of items using the concept of public value, especially in relation to heritage and broadcast collections. The goal is to increase recognition of the notion that the public value of audio visual materials far exceeds the commonly used but narrow concept of commercial value.

Arts and Humanities

Arts and Humanities Data Sharing

Compared with some other disciplines researchers in the arts and humanities do not publish a great deal of research data. Although many researchers in these disciplines still prefer to publish considered works in monographs, there are nevertheless a growing number who produce and share their datasets. Researchers in the fields of archaeology, epigraphy and the history of art produce, for example, lexica, edited catalogues and statistical data. Historically data sharing has been largely influenced by the work of the Arts and Humanities Data Service, an organisation that provided a range of data-related services to scholars. The withdrawal of funding by the Arts and Humanities Research Council led to the demise of this service in March 2008 and the disaggregation of subject-specific data repositories, but its legacy is a cadre of researchers in the arts and humanities that know something of data management planning and the mechanisms and benefits of data curation.

One of the sixteen case studies covered by this project is in the arts and humanities: classics. Classicists have a long tradition of sharing the data they produce and many take pride in publishing their datasets themselves in preference to putting them in an institutional or subject-specific data archive. They also tend to add value to data before making it widely available through, for example, editing, annotating or interpreting the data. Researchers may seek technical guidance from bodies such as the Oxford Text Archive or the Centre for Computing in the Humanities in order to enhance the long term viability of their datasets, though data management planning is not required by all the relevant public and private sources of research funding.

Scholars in the classics typically develop their own particular fields of expertise, which means there is usually less direct competition between them for recognition or resources than is the case in other disciplines. This situation helps foster a supportive attitude with regard to sharing data and disseminating their work to as many peers as possible. Many of the datasets produced by classicists are widely and freely available and are often the subject of addition or revision. The fact that community-based collaborative development of key datasets goes on is indicative of the extent to which classicists are prepared to share their data for the advancement of scholarship in their field of study.

Arts and Humanities Data Discovery, Access and Re-use

For eleven years to March 2008 the arts and humanities research community was well served by the Arts and Humanities Data Service (AHDS). This service was appreciated by researchers in the field and offered domain-specific advice on many aspects of digital data curation. Among other things, the AHDS offered a facility to search across its data collections which included historical databases, literary texts, linguistic corpora, performing arts multi-media collections and so forth. While this search facility still exists it is not clear how long this will continue to be the

case and its demise would undoubtedly make the task of finding datasets in the arts and humanities more difficult and time-consuming.

Arts and Humanities Data Preservation

In the classics, researchers tend to take it on trust that the datasets produced by their peers are done so to a professional level of quality. This is important because people working in the classics commonly want to use other researchers' datasets as the basis for their own work. The emphasis often needs to be on trust because quality assurance information such as details about the analytical method used is not always published alongside the data. It is important to note that the prevailing desire in the field is for datasets to be viable as long as there is interest in the subject and therefore they are viewed no differently from printed publications. Where data are the end product of a project, it is usual for the funders to require that the datasets are made publicly available. In situations where funding comes from private and charitable sources (as is common in the classics), the awards do not necessarily include funding for the maintenance of datasets once the project is complete. In contrast, large projects which require substantial funding normally do include provision for long term preservation of datasets. Technical guidance on the preservation of datasets can be sought from the Oxford Text Archive, the Centre for Computing in the Humanities, subject-specific data centres, local university computing centres or sometimes from researchers' networks of peers. The extent to which contextual information is provided with datasets varies greatly. As to where datasets can be preserved, researchers' options have become more limited with the demise of the Arts and Humanities Data Service, and many hope that their own universities will step in to support the preservation of these resources.

Social sciences

Social Sciences Data Sharing

Perhaps more than in any other research category the preparedness of researchers in the social sciences to share data depends very much on the traditional characteristics of individual disciplines. This study draws on four case studies in the social sciences which cover rural economy and land use, social and public health sciences, social studies of interaction and architecture. What many social science disciplines have in common is their focus on collecting and using data that is to some extent bounded by rules or agreements relating to confidentiality together with legal and ethical considerations. These factors can be significant barriers to the sharing and reuse of research data. Even though it is generally believed possible to render data anonymous, some data managers are concerned about the fallibility of these techniques and opt for a cautious approach, often limiting access to the team responsible for collecting the data.

The existence of such barriers is exemplified in researchers' use of video data in social studies of interaction. Researchers working in a variety of social science disciplines use video in studies designed to improve understanding of human interaction. This has application in, for example, technology design. Video data is generally associated with other data – audio, transcripts and annotations for instance – and may be analysed in a variety of different ways. The data is frequently shared on an informal basis with known peers, often as part of the analysis process,

but more general sharing of the data is constrained due to ethical and legal considerations. That said, because re-analysis and re-annotation is seen as a desirable goal, some video archives are developing models to support this type of activity online.

In the field of social and public health sciences, despite the desire of the Medical Research Council and the Economic and Social Research Council to encourage the sharing of research data, in practice researchers are not often inclined to do so. Typically researchers will request data from one of the guardians of national longitudinal surveys, such as the General Household Survey, or from a centrally-funded cohort study such as the 1970 British Cohort Study, and perform some re-analysis of parts of the data to answer specific research questions. Although such work typically leads to publications, the new derived datasets are not systematically shared with others. There are signs that this culture is beginning to change as funders promote the merits of effective data management planning and curation. In addition the social sciences benefits from the world class service offered by the Economic and Social Data Service (ESDS) which professionally curates worthy data sets. The capacity of the ESDS is, naturally, finite so decisions have to be made about which original or derived datasets will be looked after for the long term. Researchers in the socio-medical field report few personal incentives to share datasets; it is worth noting in this context that only datasets that are nationally recognised as being important would normally be included in an institution's national research assessment submission.

The Rural Economy and Land Use programme (RELU) was conceived to bring together natural and social scientists from many different disciplines to address key concerns to do with environmental, social and economic aspects of rural development. The programme benefits from a dedicated Data Support Service which is embedded within the UK Data Archive and which provides help with data management planning to facilitate the sharing of data. This explicit focus on data management and the aspiration of senior partners in the programme to share data has had a useful, positive impact in helping researchers appreciate the long term usefulness of the data they produce. There are, however, many reasons why data sharing has been less successful than some had hoped. Some researchers have not internalised the value of data management and sharing, perceiving it to be a bureaucratic hurdle, while others think that their "small science" data is unlikely to be reused by others. As with other social science disciplines, issues of ethics, consent, confidentiality and data protection are often invoked as reasons why data cannot be shared.

In applied disciplines such as architecture, it is thought that the true value of curating and sharing data lies not simply in the data itself, but the combination of that data with an understanding of the process of how something was designed. Thus for effective reuse, architects need to be able to discover enough about a project to understand its context and the evolution of the design – something that is difficult to do at present.

Social Sciences Data Discovery, Access and Re-use

In many social science disciplines researchers are accustomed to working on data sourced from national, often publicly funded datasets, performing analyses on those data and deriving new datasets. On the whole, the process of discovering relevant datasets is not a major obstacle: researchers tend to know where to find the datasets that are central to their work, whether in

economics or social and public health sciences. Social scientists are fortunate in having access to the services of the Economic and Social Data Service (ESDS), widely regarded as offering a gold standard data curation service for the social sciences. The ESDS professionally curates a wide range of datasets that have been selected for their potential usefulness to the research community.

Despite the efforts of the UKDA, access to social science datasets is incomplete for a variety of reasons. This is a field in which a lot of primary data includes information about individuals whose privacy and rights to confidentiality must be respected, particularly when medical records are involved. In such cases access to the data beyond the boundaries of the data collection team or organisation is often limited to some degree. An example from the social and public health sciences case study tells of a situation where researchers are required either to send their data requests or query code to the data creators for attention, or researchers might be allowed local access to a restricted dataset via a monitored workstation. There are a host of other reasons given by data creators to justify their decisions to limit access to those data: cohort studies rely on the willingness of individuals to participate over a long period of time, so researchers are reluctant to do anything that might damage their relationship with these individuals (such as their identity inadvertently coming to light); cohort or other large scale datasets are sometimes said to be so complex that they can be difficult for people not closely associated with the data to use on a conceptual and practical level; some data creators fear that unfettered access may result in their data being misrepresented; and finally, the teams responsible for building and maintaining important national datasets need to justify their value to their funders and, since this value is often measured in terms of published papers, data creators have an interest in extracting as much value as they can from the data they collect and look after.

Some social science datasets are just plain expensive to access: a few important longitudinal datasets are available to researchers so long as they pay a fee, contributing to the costs of administering the study. Major datasets produced by public sector organisations such as the Ordnance Survey, the Meteorological Office, the Environment Agency and DEFRA often operate on a commercial basis, charging for access to data. The PointX (Points of Interest) database produced by the Ordnance Survey has been cited as being particularly expensive; even with higher education discounts the cost of access and reuse can run into many thousands of pounds, normally beyond the budgets of research groups.

Architectural practice and research spans many disciplinary boundaries. The cross-cutting nature of the field militates against domain-specific data centres and means researchers and practitioners may need to search an array of sources to find data that might be useful to their projects or research. Although architectural departments in the UK do have repositories of research and learning materials, common methods for discovering what lies within them are lacking. With the help of JISC funding, the Lincoln School of Architecture attempted to address the discoverability problem through the development of the Lincoln Repository of Learning Materials. The ambition was for the repository to support objects like digital animations of 3-D models, architectural documentation such as technical briefings and materials, together with supporting text based materials. Although the technology to support this ambition was insufficiently developed at the time of the project, the attempt to build the repository highlights the need to bring together different types of digital objects together with their metadata so they can be discovered and reused by the architectural community.

To complicate matters further, it has been reported that conceptual design information – the nuanced, creative application of human knowledge – is by no means simple to capture in a form that can be adequately curated. Efforts are being made to tackle this issue: the Building Stories programme at the University of California, Berkeley, attempts to capture the story behind a number of building projects in an online repository to facilitate the sharing of design knowledge. In Europe, the MACE project (Metadata for Architectural Contents in Europe) aims to improve architectural education by integrating and connecting content from diverse repositories containing information and data relevant to architectural design communities. The project is creating new tools to help people find and reuse data that was hitherto accessible to only a few.

The case study that investigated the roles and reusability of video data in social studies of interaction highlights the difficulties researchers experience in finding, let alone reusing such data. On the positive side, the development of novel analytical tools offer search and concordance capabilities across video collections and corpora. In addition, metadata schema and archival expertise can be transferred from linguistic domains, psycholinguistics and language documentation for example. While the potential for discovery reuse exists, the reality is that local and national repositories are only now beginning to make available information on the availability of video data. Metadata and annotation schema for video data are not widely implemented and vary to a significant extent between research communities. Finally, the search tools that do exist are generally unable to search across video content except where this has been described with meaningful metadata, or where the content has been transcribed into a textual format. In short, the potential for finding video data for the research purposes described in the case study is currently very limited.

Social Sciences Data Preservation

It was noted previously that the portfolio of research projects funded under the umbrella of the Rural Economy and Land Use Programme had access to data management expertise from a dedicated Data Support Service (DSS). The DSS is available to offer advice at all stages of the life cycle of the programme's research projects, from data management planning to helping with depositing datasets in an appropriate data centre such as the Economic and Social Data Service. Not only were project leaders required to sign up to the programme's Data Management Policy, but a condition of the award was that datasets were to be offered to the ESDS or one of the NERC's data centres within three months of the end of the project, at which point the receiving data centre would consider the suitability of the datasets for preservation. The sanctions for not doing so, or for offering up datasets that are unsuitable for curation, perhaps due to the use of inappropriate metadata, are limited at present. In reality, some projects took data management seriously, perhaps appointing a dedicated data manager; some did not. It is interesting to note also that within multi-disciplinary teams, researchers from different disciplines are by and large left to get on with their own specialist part of the project. Colleagues trust that research group members will apply appropriate levels of quality assurance to their work and the production of data. The key point to observe is that despite the support structure in place for this programme, adherence to the Data Management Policy and the suitability of datasets for curation was variable, reflecting individuals' personal attitudes towards the value of datasets with regard to their long term viability and reuse. The traditional view that the primary outputs of a project are journal articles prevails, with the fate of the underlying data coming further down the list of priorities.

Similar attitudes to the long term viability of datasets can be found among the researchers who use video data in analyses of human interaction. The projects that were covered by the case study had been the subject of a data management planning process to comply with funding body requirements. It was found, however that these plans and reality diverged at an early stage in the projects, due in part to the speed of change of video capture technology, format changes and the sheer volume of the video data being produced. For long term viability, researchers needed access to a data centre but access to such facilities can depend on which research council was the main funder of what are normally multi-disciplinary projects, and some funders do not support data centres. It is worth noting that researchers expected their own institutions to be able to provide affordable managed storage, technical support and a preservation facility – but few institutions appear to be able to offer such services at this point.

The situation in the field of social and public health sciences is somewhat different mainly because many of the producers of large datasets also look after those datasets for the long term. The units funded to collect data for the major cohort studies and national surveys also add value to the data through data cleaning, verification, organisation and documentation. The researchers working in these units can develop a close affinity with what can be complex datasets and are protective of the value of the datasets. Even though the professionally curated datasets are nationally recognised for their quality and contribution to scholarship, long term funding for them is by no means guaranteed. Managers need to re-apply for funding periodically at which point the funders will typically review the value of the datasets. Often their value is measured using proxy indicators such as the number of times the data is reused or, perhaps more commonly, the number of published journal articles that have been largely derived from the datasets. Even though it is possible to cite datasets directly (and indeed the Economic and Social Data Service spells out a standard method for doing so), at present researchers' perceptions are that citations to published papers carry more weight in the view of research funding bodies.

While the longevity of important national datasets in this field is relatively secure, if not necessarily guaranteed, many smaller datasets often fall by the wayside. The long term viability of datasets created by individual researchers or research groups seems not to be important to many of them since there are currently no explicit rewards for investing the time and money to look after data once a project has finished. The process of curating data is normally secondary to the cycle of publishing papers and applying for research funding. Even in situations where data has been requested from a national dataset, re-processed and analysed to create new value, these new derived datasets are, it seems, not often offered back to the data centre from which the raw data originally came even though the managers of those data centres would often welcome the opportunity to review those new datasets for possible long term preservation.

Life sciences

Life Sciences Data Sharing

The quantity of data produced in the life sciences is growing at an enormous rate, largely outstripping attempts to curate it. The size of individual datasets can be huge: in systems biology, for example, microarray data text files are produced which may have millions of rows and eighty columns of information. A single experiment may involve up to several thousand

arrays. Handling data on this scale requires large amounts of storage and very powerful computing facilities to permit effective manipulation. It is commonly perceived that life sciences researchers in general are willing to share data although, anecdotally, the sharing ethos can be somewhat selectively applied. Even where researchers publish in journals which specifically require authors to share their data, a recent small scale study of authors has found that many fail to comply.[1] As with some social science disciplines there are, of course, a number of life science disciplines where patient confidentiality precludes the sharing of some data.

There are fields of study in the life sciences where the idea of data sharing has yet to be internalised. Researchers working on the CARMEN project – designed to engineer, utilise and refine scalable e-Science architecture to serve research into the nervous system – concluded that they required complete control over the sharing of their data. Although there is an accepted expectation that data would be released to interested parties following publication of the project's outcomes, the definition of "publication" is not entirely clear. If it is taken to apply to the publication of papers derived from the data, the delay may be rather long. There is no consensus on the sharing of derived data and the issue of licensing data to, for example, commercial users or medical charities remains unresolved.

In the field of neuropsychology (and based on a case study of the work of the Neuroimaging Group in the University of Edinburgh's Division of Psychiatry), obstacles to sharing neuroimaging data include concerns about disclosure of confidential data and misinterpretation of data without sufficient contextual and technical information about the original experimental context. The Group is sharing data with other research groups through the Medical Research Council funded e-Science projects Neurogrid and NeuroPsyGrid. In the work done by the Group, the Principal Investigator is the gatekeeper of any identifying demographic details (for images that have had identifying data removed), while senior imaging researchers ensure that scans are also de-identified. This facilitates sharing, but multi-centre projects raise complex issues about balancing individual privacy by pseudonymising shared data, with the recruiting centres' clinical research need to be able to re-identify images, for example to conduct repeat scans for longitudinal studies. Overall, the probability of identity disclosure depends on the measures taken to anonymised the data linked to scans (patients' medical histories and demographic details for instance). While scans contain personally unique features, they are les sensitive due to the high skill levels and computing resources needed to match de-identified scans with identified ones, though this risk may increase with improvements in automation and analysis tools. There is limited appetite among Group members for making the data available more widely except where there are clear research benefits: concern to protect the research value of the data is as much a factor as privacy. The data sharing model the Group would prefer is "give to get" rather than "give away". They are nonetheless committed to following the researcher funder's policy which places an onus on custodians to make data available for new research purposes in a timely manner, while balancing the interests of data creators, custodians, users and data subjects.

The muted level of enthusiasm for data sharing indicated in the two cases outlined above contrasts with the experience of the Edinburgh Mouse Atlas Project (EMAP). EMAP is a biocuration project designed to develop an expression summary for each gene in the mouse

[1] http://www.plosone.org/article/info:doi/10.1371/journal.pone.0007078

embryo. EMAGE (the Edinburgh Mouse Atlas Gene-Expression Database) is one of the first applications of the EMAP framework and provides a spatially mapped gene-expression database with associated tools for data mapping, submission, and query. The EMAP service is a public database and researchers in the molecular and developmental biology community, known for their willingness to share, are keen to deposit datasets.

This project also reviewed a case study focusing on integrative biology, which examined the aims, funding, technological infrastructure and operation of a project designed to develop a theory of biology and biological function. Much of the work involves voluntary sharing of data and software and discussions on open source licensing are actively encouraged and supported through the JISC-funded Open Source Software Watch group. The case study revealed that providing access to raw data is often less of a problem than providing access to lab note books and the series of actions or methods which may reveal a researcher's scientific agenda.

As the section overview indicated, there are life sciences disciplines in which data sharing tends to be the norm rather than being conditional and partial. Research in genomics, for instance, is largely funded by agencies which have policies designed to encourage positive attitudes to data curation and sharing. The Biotechnology and Biological Sciences Research Council (BBSRC), for example, recommends that all raw data be retained for ten years, and applicants for BBSRC grants must include a data plan in their proposal which may also include requests for funds to support data curation and sharing. Genomics has a long tradition with respect to data deposition in one of three large publicly funded databanks. Typically researchers deposit their data – though not necessarily all the raw data - when an article based on that data is submitted for publication in a journal (and in fact some journals require evidence that deposition has taken place). The community push factor to make data publicly available is very strong in this discipline; not everyone does so but the majority do, not least since there is a strong esteem-based advantage in so doing. Some researchers even share their data via blogs or wikis as soon as they are created.

Systems biology has a broader scope than genomics and reflects a new, integrative approach to biological questions drawing on mathematical and statistical techniques to manipulate biological data. Researchers working in the six systems biology centres in the UK produce a vast amount of data through experimental procedures and by working with curated datasets to assemble evidence of new biological phenomena. One machine run for high throughput molecular work can generate a dataset of four terabytes. They use the same three public databanks as the genomics research community for depositing datasets and obtaining data for reuse in addition to other more specialist databanks such as the Database of Interacting Proteins (DIP). Some large laboratories upload gene sequence data to the public databanks every night, while microarray data are more usually published at the end of a research cycle, when a journal article is produced. For some types of data – such as the outputs from microscopy – there is no publicly-funded databank provision so, while some micrographs may be published, most of this type of data is held on local systems. The prevailing ethic is to share data and most try to make data available, often on a project website and sometimes using Web 2.0 tools, even though there is no formal recognition for sharing data in this way – just the intangible reward of playing a full role in their researcher community.

Life Sciences Data Discovery, Access and Re-use

Although not all the life sciences case studies we covered had specific information about data discovery, access or reuse, those that did offer useful insights. In neurophysiology, data discovery will be relatively simple once the CARMEN project has established all the necessary structures and procedures, since neuroscientists will deposit their data in the requisite store and others will be able to go to that store via a CARMEN service to access and reuse the data. Initially, there will be two repository sites, one at the University of Newcastle and the other at York. Although they will have the same infrastructure, data will not be mirrored across the two sites. This environment of federated data will be accessed by consortium members via a single interface (that is, through a process of dynamic deployment, users will be routed to data irrespective of its location). Together, these repositories have been named the CARMEN CAIRN (CARMEN Active Information Repository Node), which is to be considered as the original nexus of a distributed whole. In the future, assuming a larger proportion of the global neuroscience community eventually subscribes to CARMEN, it is envisaged that geographically dispersed groups would create their own CAIRNs, and that the CARMEN integration function would be scaled up to support the presentation of data and services across this expanded family of CAIRNs, as if they were a single entity.

There is a pronounced sense of data ownership within this research community and the habit of making data available for others to reuse is not well developed. It is even felt that the CARMEN project may face a potential threat because of this. A view held amongst the CARMEN project membership is that data security is critically important to experimenters. It is also firmly believed that they do not wish to be merely anonymous contributors of data but expect to be directly involved in any further analysis of their data sets, and that this expectation will be supported. When users login they will have immediate access to their own data and those datasets that others have specified they may access. If they wish to see other datasets, an email is generated to the data owner asking for permission to allow the user access. It is hoped in time that the nuisance of this procedure may encourage greater willingness to data sharing amongst data owners.

In the case study of the integrative biology project, the quest continues to enable the reuse of datasets for the benefit of the user community. The development team will review future use of a Storage Resource Broker (SRB) across continents, looking into ways of federating catalogues thereby reducing the need for mass implementation of global metadata. For the present, raw model simulation datasets appear to be reusable by project partners. Current integrative biology users are predominantly data creators, running numerical codes to model physiology. As the integrative biology infrastructure develops, it is predicted that new numerical physiologists will become reusers, accessing the stored results of previous calculations for further analysis.

As is the case with many disciplines, researchers in neuropsychology seek ways to integrate data in order to enhance opportunities for reuse, but there are obstacles to be overcome. In this particular case study, for instance, researchers point to the fact that scanners vary in many ways, including magnetic field and image intensity; research centres may recruit from markedly different populations and there is wide variation in image analysis tools. Projects are, therefore, trying to focus on developing and using standardised tools, harmonised methods, the normalisation of scanner output, and the coordination of quality assurance. The potential rewards

from improved data description are high: reusing data with novel multivariate analytical techniques would enable studies that would otherwise be uneconomic. These techniques could be provided as web services to allow remote analysis by external researchers (and steps toward this have been taken in the NeuroGrid e-Science project). It is worth noting that, in this field of research, while people would like access to data there is less interest in reusing derived data to replicate previous analyses than might be the case in other fields of study.

Turning to the EMAP project, data discovery is provided via the EMAGE data service, a public database. Initially, users were mainly researchers in the USA and Europe. These were tracked regularly by obtaining information of requests for mouse atlas data (on a CD) and online EMAGE software, publication references and website usage. Since 2007, exact use of the EMAGE interface has not been logged, but there are approximately 2 million requests per year for static EMAGE web pages and there has been a significant increase (from 105,000 to 170,000 requests) in the Repository Browse / Quick Search function since 2007. This is an interesting view of researchers' discovery behaviour from the data provider's perspective. It seems that if researchers are offered effective means for discovering data, they will use it in increasing numbers.

The ease of discovery and access associated with the EMAP project is also evident in the disciplines of genomics and systems biology where access to data is straightforward. There are generally no restrictions on access to data from publicly funded databanks. Some databases do have a subscription or pay-per-access arrangement, but these tend not to offer databases central to many researchers' needs. Access to privately-held data depends on the goodwill of the data creators who may have legitimate reasons for restricting access, perhaps because the data are work-in-progress or due to reasons of confidentiality. The reuse of publicly available data in the genomics is a cornerstone of research practice.

Discovering datasets in the large public genomics databanks is a relatively simple process and most people working in these fields are also familiar with smaller, specialised databases that are relevant to their work. Because these large databanks benefit from well structured and detailed metadata fields, searching by keywords (such as gene names), author names or more complex terms is straightforward. Web-based search programs exist to facilitate these types of searches, NCBI's Entrez and EBI's SRS service to name but two. Google is also routinely used to find DNA primer sequences (short lengths of DNA at the beginning of a coding sequence), but such searches will not be comprehensive because most public DNA/protein databases are not completely indexed by Google. Despite the existence of big publicly-accessible databanks serving the genomics and related research communities, some researchers still report having to trawl the web looking for datasets that they know exist but cannot find, or for datasets they are not familiar with. This can be a time-consuming task since each research group's website has different structures and routes to find datasets, not all of which are necessarily obvious or intuitive. The problems of discoverability of smaller web-based datasets reported by researchers in systems biology – which no doubt apply to researchers across the life sciences to a greater or lesser degree – can be summarised as follows: it can be difficult to locate datasets in the first place; websites can be difficult to navigate, complicating the task of finding datasets; data formats can vary in their application and consistency; even when datasets are discovered they may be inaccessible.

Access to data in systems biology is via four main routes. First, data can be accessed via the websites of journals in which articles built on the data are published. Most of this data is represented in pdf format which is far from ideal for reuse. Second, researchers can access data from the websites of individual researchers or research groups. This is perceived to be a good option since data can be recent and of good quality, though the websites can be difficult to find and navigate. Third, data may be obtained from web-based databases maintained by individuals or research groups – though because they are normally funded on a short term basis, the long term viability of the data is often far from certain. Finally, the main public databanks offer excellent access and some also provide a centralised search facility enabling cross-database searching and discovery. There are usually no direct costs to researchers associated with these access routes, though their institution may need to subscribe to journals to enable access to a web-based article and associated data resources. There are sometimes costs associated with the need to buy software or licences to decode or otherwise work with other peoples' data. For example, mass spectrometer data are mostly produced using a propriety software tool known as MASCOT. Researchers wishing to use proteomics data need to buy a MASCOT licence that costs several thousands of pounds per processor.

The systems biology case study also highlights further issues to do with access. Some of the datasets available to all comers are incomplete. Data managers in research groups with sufficient expertise can sometimes piece together incomplete datasets from different sources, but the expertise required is in relatively short supply. Datasets vary in their consistency and quality, which can cause problems for reuse, and the practice of over writing data with newer data is a problem in systems biology. It is argued that science should be reproducible and over written data prevent this.

Life Sciences Data Preservation

The CARMEN project provides a very interesting study into how a planned system of data capture, description and storage can yield dividends in terms of the long term viability of data. In the field of neuroscience, there are two main methods of collecting experimental data: by direct recording of neuronal activity by means of electrodes that pick up electrical impulses; and by light microscopic visualisation of neuronal activity using fluorescence-labelled ion-channel markers. Under these two main headings fall many differently-detailed techniques. The goal of the CARMEN project is to provide a Storage Research Broker (SRB) file store facility for both raw and derived data. The objective of retaining raw data is to ensure that they are frozen for science and permit "freeze-and-build" work to be carried out in future. Freeze-and-build simply means compiling a dataset and putting it somewhere, and then maybe doing some more work on that experiment, or repeating it and getting slightly different or slightly extra results, and adding them to the original dataset in such a way that the original dataset still exists. In this way the original dataset is not changed or replaced. Where datasets have not been used for two or three years they will be moved to a different storage facility (probably a tape store), simply for reasons of cost: they will still be accessible to anyone who wants them, albeit after a slight delay whilst they are recovered.

Data uploaded by experimental neuroscientists will include both raw voltage signal data that has been recorded in the laboratory and images captured from neuron and neuronal network activity.

From prior experience it is known that these will represent a very high volume of data, with an initial storage requirement predicted at 50 terabytes.

Terminologies and nomenclatures are not consistent across the different neuroscience disciplines, nor are they used consistently. Sometimes, individual neuroscience laboratories produce data in their own preferred format, with locally-developed metadata schemes. The potential for data sharing has been low as a result. The CARMEN Project is to provide a Grid-enabled set of tools and standards for archiving and describing the data so that authorised analysts and other experimental researchers can access and re-use data easily. Most importantly, neuroscience data are not suitable for reuse without clear descriptions of the experimental conditions for their collection.

Users of CARMEN will interact with the CAIRN (CARMEN Active Information Repository Node), via a Web portal, which will serve as a conduit for uploading their data, for annotating it with metadata, and as a facility for locating and browsing data within the CAIRN. The portal will also provide the means to create, run and monitor workflows. The CAIRN is seen as a repository for the long- term storage and curation of both data and analysis services. The original aim was to store raw data in a separate location (a Storage Resource Broker filestore) to preserve its integrity, while derived data were to be held in a database system fully accessible for re-use at all times, but for cost and technological reasons the decision was made to store raw and derived data together in the SRB. As yet no decisions have been made about sustainability. Given the effort that has gone into developing the system and its potential value, ideally it will continue indefinitely. This will, however, require considerable resourcing and it is not clear where that money might come from.

Further insights into the complexity of finding funding for curating datasets are provided by the EMAP case study. Currently, core funding for the work completed by the EMAP is from the Medical Research Council. There are a number of additional external funding bodies including BBSRC, the NIH and the European Commission. In the past funding has been obtained from the pharmaceutical company, GlaxoSmithKline, for the reconstruction of a new model (there were no intellectual property restrictions, which enables EMAP to incorporate the new model into the atlas). Obtaining funding for the development of database and ontology resources is possible; however obtaining funding for the maintenance of ongoing databases is currently more difficult. There are a number of key organisations that EMAP personnel are working with, establishing and strengthening relations, in the hope that these organisations will be the gateway to longer term funding opportunities.

Funding for the professional curation of datasets in genomics is less of a problem than in other fields, though the task of securing funding is a necessary and time-consuming task for dataset managers. Funding is typically derived from multiple sources which ameliorate the risks synonymous with relying on single sources. Datasets that are deposited into the big public databanks are looked after to a good standard, though most of the data are not manually curated. It is worth noting that the meaning of curation tends to be slightly different in this discipline with a greater emphasis on annotation and linkage to publication evidence. Although the level of care for sequence data is relatively high, the quality of the actual sequence is not checked at deposit and the annotations are not checked for accuracy. Not all data are deposited exclusively in the public databanks. The large genomics laboratories such as the Sanger Institute in Hinxton,

funded by the Wellcome Trust, post gene sequence traces on their own websites as well as depositing the traces from every repeat of an experimental run in GenBank. Smaller laboratories or groups will normally deposit one representative dataset in GenBank but may choose not to deposit the raw data, deviating from best practice. Microarray data is also professionally curated by the big public databanks though not all of them are annotated sufficiently to permit reuse for comparative analysis of experiments. Image data from confocal microscopy tend to be massive and are usually kept by researchers on local storage media; the long term viability of these datasets is questionable.

In addition to the large databanks, researchers in this field are served by many smaller specialised public databanks that curate genomics datasets. In molecular biology alone there are over 1000 such databanks and in the area of gene sequences there are databanks dedicated to a single species or genus. These types of resource are typically funded using project money and their long term existence is uncertain. When the funding runs out or a project team disbands, the data may remain available via the Web but usually only in static form with no prospect of active curation.

The problem of small databanks surviving on precarious funding sources is common across the disciplinary spectrum, though in many disciplines there are at least a number of data centres which curate important datasets for the long term. There is no such provision to support the archiving of neuroimaging data. Although the involvement of the researchers in the neuroimaging case study in multi-centre Grid projects is designed to facilitate data integration and sharing, a database infrastructure to maintain and add value to *locally held* data is still in the planning stage. This is a situation that is quite typical of centres in the UK and is attributed to a historic lack of infrastructure funding for the neuroimaging field.

The integrative biology (IB) case study describes a rather different approach to data curation. The project is designed to exploit e-Science and Grid computing technology, harnessing the power of distributed computer and data resources to run ever more complex *in silico* simulations: complex multi-scale models from cell level to whole organs, including the assimilation of *in vitro* and clinical data to fine-tune the models. Computational biologists conduct collaborative simulated experiments within a distributed virtual laboratory or 'collaboratory'. The location and content of data is self-documented and self-managed.

The project architecture comprises five main components, which are essentially technologies for analysis: first, infrastructure components such as managed data access and control of data visualisation; second, a simulation subsystem providing 'solvers' for user-supplied model codes; third, a data management subsystem stores users' data files and association metadata and provides facilities for data retrieval and curation/annotation; fourth, a visualisation subsystem offers a range of techniques for examining simulation and experimental results (which can be run in real time with remote colleagues); and finally, users can interface with the service via the web or a stand-alone IB interface. Longer term storage of data is handled by the ATLAS Data Storage facility at the Rutherford Appleton Laboratory, a discipline-neutral, sustainable large scale facility operated by STFC. A strategy is being developed to determine when particular files should be archived as part of an automated data management infrastructure – one in which IB tools disassociate the end user from details regarding the physical location of individual datasets.

In recent years the publicly funded research councils have been reviewing and strengthening their policies with respect to research data and this is evident in the field of systems biology[2]. The Biotechnology and Biological Sciences Research Council (BBSRC) has a clear and comprehensive policy whereby it expects research data from the projects it funds to be not only shared in a timely fashion (and no later than the point at which the main findings based on the data are published), but also that existing standards for data collection and management should be used and that data should be made available through existing databases known to the community. The BBSRC also explicitly recognises that different fields of study will necessarily adopt different approaches to data management, echoing the realities of disciplinary differences in relation to the data curation lifecycle that have been made plain in this review of case studies.

In terms of the long term viability of datasets, the BBSRC requires that data are retained for a period of ten years. What happens to the data after this period remains to be seen, but this policy points to the difficulty of deciding how long to preserve data and at what point their value begins to diminish. Looking after datasets even for ten years, particularly for smaller research groups, can be challenging mainly because data outputs can be massive and the facilities required to store those data can be expensive to maintain. The outputs from confocal microscopy used in the systems approach can be voluminous and the curation problem is made more onerous when derived data are produced in the reconstruction of images. The data outputs from high-throughput molecular work are on a scale comparable to astronomy – but systems biologists are not as well served as astronomers in terms of dedicated data centres. The storage of data in systems biology is but one part of the challenge; the other is the curation of the software tools required to handle the data. Obsolescence and the cost implications for maintaining necessary software licenses can make such curation expensive, though increasingly "small tools and resources" grants are becoming available as funders come to recognise the full range of costs associated with effective data curation.

Physical Sciences

Physical Sciences Data Sharing

As with the other research categories, researchers' attitudes and practices to sharing data in the physical sciences vary by individual discipline. At one end of the spectrum, the sharing of data is well-advanced and formalised in astronomy, while at the other end our design and manufacturing case study indicates the paucity of data sharing activity in this discipline.
Observational astronomers produce primary research data of three types: images; spectra from individual objects and light curves. These data are used by theoretical astronomers who may produce derived datasets as part of their work. Typically observational astronomers lease time on ground based telescopes or orbiting observatories, collect all the data they can in their allocated time, and then they have twelve months to process and analyse the data themselves. After this, the data are made available to everyone; this is a community-accepted convention. Data are curated in distributed UK data centres: Jodrell Bank at Manchester University and departments of

[2] Information on these policies is collated and presented by the Digital Curation Centre and can be viewed via this link: http://www.dcc.ac.uk/resource/curation-policies/

astronomy at Edinburgh, Cambridge and Leicester universities. These data centres try to make available as much data as possible. Astronomers can also access data from large data centres around the world, not least those run by the North American and European Space Agencies. Derived datasets produced by astronomers tend to be stored on local computing systems and although they are not in the public domain, it is the norm for researchers to respond positively to requests for derived data from other researchers. While there is no specific reward for sharing data in this way, people think this behaviour is not only good for the community but that it potentially enhances their own reputation.

Continuing the astronomical theme, this project covers a case study looking at the work of the Wide Field Astronomy Unit (WFAU) at the University of Edinburgh. The goal has been to preserve all the data produced by the programmes covered by the WFAU to enable long term reuse. Early in the history of the data centre data storage was relatively expensive so it was decided that image data served to users should be in a 20 times compressed format – leading to some loss of data but not to a significant extent for scientific purposes. More recently, good lossless image compression methods have been developed which will compress images by 3 times – which is now sufficient given that disk space continues to become relatively less expensive with the passage of time. Astronomical data have legacy scientific value and the goal is to maintain accessibility to legacy data. This involves migrating the data between several generations of media. For example, the original SuperCOSMOS image files are currently stored on Exabyte tape and are being transferred to LTO-3 tapes. The majority of WFAU's 19,000 photographic plates and films have not been scanned into digital form and care is taken to keep them in good condition so that the information they store in analogue form is not lost. Other non-digital holdings include the archives of the UK Schmidt Telescope, which are held on a mix of media and some may have heritage value. This intensity of effort is beyond the scope of individual researchers which is why data are centrally curated, which in turn facilitates data sharing for the benefit of astronomy research in general.

Climate science is another field that benefits from the existence of centrally funded data centres. Climate scientists produce large volumes of data including those from model runs, long term monitoring and observation, derived data and data associated with the graphical representation of climate data. The Natural Environment Research Council (NERC) funds the British Atmospheric Data Centre (BADC), which is the natural home for climate science data - though it cannot accommodate all the climate science data produced in the UK. The BADC sets up and maintains long term relationships with many projects, an approach conducive to scalability. There are also some national and international research projects which facilitate and encourage the sharing of datasets, but for the most part data sharing is not widespread. In general, climate scientists spend little time thinking about how to make their datasets available to a wider audience, particularly in the case of model run data. Researchers rarely request model run datasets from other researchers or research groups, though there is some demand for people to share processed, higher order datasets. Datasets that are not taken on by a data centre will typically remain on data creators' local computer systems, normally preventing access by third parties.

Climate science is a broad church and although climate modellers tend not to share data, data sharing is more common in ocean modelling and among researchers dealing with observational data. Within the weather data community, commercial imperatives all but preclude the routine sharing of data. This general predisposition not to share data is compounded by the difficulty of

discovering datasets that might be useful. For the future, NERC's policies to encourage data management planning may, in due course, lead to more data sharing in climate science.

In contrast to NERC, the Engineering and Physical Sciences Research Council (EPSRC) has no policy on data sharing. This has led to some in the field of crystallography to believe that requests for funding that include provision for data management and curation make them look relatively uncompetitive and, therefore, limit the likelihood of them being successful. Not all crystallography research is funded by the EPSRC; a considerable amount of such research conducted in universities is paid for by chemical or pharmaceutical companies. Naturally some of the research commissioned by the private sector will be commercially confidential so the issue of data sharing will not arise. In fact it is thought that more that half the data generated on crystals are not publicly shared. Of the data that is shared, much is deposited in the Cambridge Crystallographic Data Centre (CCDC), the main UK repository for such data. Although the CCDC was funded from the public purse in the past, it is now a self-sustaining not for profit independent organisation which derives a large part of its running costs from the industrial sector. Other relevant data centres are found in Germany, Canada (both are subscription-based services) and in the USA. The latter, the Protein Data Bank, is publicly funded and provides free access to data.

Among the crystallography research community there is increasing pressure for more open access to crystallographic data. Examples of projects pushing this agenda include: the e-Crystals open access database of crystal structures in Southampton; CrystalEye, developed by the Unilever Centre for Molecular Informatics at the University of Cambridge; ReciprocalNet in the USA and the Crystallography Open Database hosted in Lithuania. The community's goal is to complement the work of the CCDC rather than jeopardise its value. Larger crystallographic laboratories normally store the raw and derived data they produce, but this is not always the case for small groups or individual researchers. Crystallographers publish datasets primarily as part of a journal article since this is the main route to recognition and reward, but many datasets are not shared, residing instead in local computer systems until people find the time to deposit them in the CCDC.

So far we have looked at examples of disciplines where a lot of data is shared (astronomy) and examples of disciplines where a reasonable proportion of data is shared (climate science and crystallography). At the other end of the spectrum, a case study examining the work of the Innovative design and Manufacturing Research Centre (IdMRC) in the Department of Mechanical Engineering at the University of Bath, indicates that data is rarely, if ever, shared. Data tends to be used on a short term basis, that is, to support a single paper, report or thesis, rather than to build up a resource over time. Data is kept in order to enable the findings of a paper to be justified if necessary, without any expectation of wider reuse. Some of the raw data that researchers work on is so confidential that the researchers have to use it at the industrial partner's site, and only take off-site aggregated or anonymised derived data. The majority of data used or produced by the Centre *can* be kept on campus, but is still commercially sensitive and cannot be released; some data may not even be shared with other researchers in the same team. Sharing data is impeded largely by commercial sensitivities and associated confidentiality agreements. Where these do not apply, there is a mixed attitude to sharing data, with some groups happy to make data available, some more cautious (perhaps looking to the possibility of a spin-out

company) and some do not consider the possibility of others being interested in reusing the data they have produced.

Physical Sciences Data Discovery, Access and Re-use

The ease with which data in the physical sciences can be discovered, accessed and reused varies considerably discipline by discipline. Some disciplines are relatively well served, while others are not. The encouraging thing to note is that in disciplines where problems with data discoverability are recognised, initiatives are under way to address these issues. In engineering, for instance, the Pilot Engineering Repository Xsearch (PERX) project sought to provide a resource discovery service across a series of repositories of interest to the engineering learning and research communities. The pilot project was designed to explore issues such as the range of relevant repositories, the cultural barriers to adoption, enhancing metadata quality together with a number of related technical issues. The project was funded as part of JISC's Digital Repositories Programme and concluded in October 2007. The experience gained was not lost: input was provided to the TechXtra service run by Herriot Watt University, where people can search for up to four million items across engineering, mathematics and computer sciences. It is relevant to note three reported conclusions from the PERX project. First, there exists community support for a subject-based approach to resource discovery. Second, there are differences between disciplines which should be carefully considered in the evaluation of suitable approaches to resource discovery. Third, the project concluded that the engineering information environment is complex, which makes the task of providing cross repository discovery services somewhat difficult.

Focusing on the engineering case study reviewed for this project, the case study investigator concluded that within the field of engineering there are few repositories of engineering information of any kind, and that there is a lack of repositories dedicated to engineering research data. Some research is being done to consider how to make industrial data more interoperable but, for the time being, data discovery, access and reuse in engineering is difficult. Notwithstanding the absence of suitable repositories or databanks, in cases where research is funded by the commercial sector, the barriers that attend issues of confidentiality come into play. In addition, most of the specialist software used in the engineering department at the centre of the case study is licensed on a time limited basis. This means that the usability of data within and outside the department is dependent upon the continuing availability of the software and the associated licenses and the willingness and ability of the department to continue to pay for it.

The climate science research community is much better served by publicly funded data centres compared with engineering but, even so, the discoverability of datasets is not straightforward added to which researchers' discovery strategies tend not to be systematic or optimal. The British Atmospheric Data Centre offers web-based searching capability and currently lists 219 datasets that are publicly available. Further afield, ACCENT (Atmospheric Composition Change, The European Network of Excellence) offers 244 datasets in a metadatabase and is funded by the European Commission. There are also specialist centres where researchers can access datasets, such as the National Oceanography Centre in Southampton, the European Centre for Medium Range Weather Forecasts, the Program for Climate Model Diagnosis and Intercomparison, and the World Climate Data Program. These are well managed organisations

where long term funding assures their continuity and they typically provide tools for data discovery, reformatting and delivery. At the project or departmental level, a minority of climate science departments in the UK have websites from which people can access datasets, though they are typically not well funded and their long term availability is by no means guaranteed. Some climate science projects have their own data archives. Although these are primarily for internal use, access to external users may sometimes be negotiated. Datasets on researchers' personal computers or institutional servers are not normally discoverable though, where people do find out about them, access might be negotiated with the data creators.

Recognising that the discovery of datasets can be a tortuous and sometimes futile pursuit, the Natural Environment Research Council has been funding the development of the NERC Data Grid. This comprises a series of components designed to facilitate data discovery in the disciplines funded by the NERC. Most relevant here, the NERC Data Discovery Service enables researchers to search across databases held in managed repositories or data centres and other project-based locations using a simple search interface. In practice, while there is demand from researchers for access to datasets that contain monitoring, remote sensing or observational data, there is much less demand for the raw data produced by climate modellers. In greatest demand are the large national or European datasets – though the larger and richer the datasets are, the more computer processing power is required for analyses and to derive new datasets. In this context, demand for the biggest datasets is self-limiting.

For crystallographers the Cambridge Crystallographic Data Centre is the primary resource for finding and accessing datasets. The quest for datasets is made simple by a system of accession numbers which can be cited alone or together with a journal article. The data centre permits the reuse of CIFs (Crystallographic Information Files) for data mining, allowing new kinds of scientific work to be done based on these data, not least in the realms of crystal engineering, supramolecular chemistry, polymorphism and crystal structure prediction. The publicly funded Chemical Database Service enables researchers to search a variety of databases related to crystallography via web interfaces. In addition, *Acta Crystallographica E* publishes items that are effectively datasets only since there is little or no introductory or discussion material in the "articles". It was noted in the data sharing section above that at least half the crystallography data produced is neither discoverable, accessible nor reusable. Given that the demand to reuse crystallographic data (thanks to new tools to mine and re-engineer data) this is clearly a limiting factor. In response, open access collections of datasets have been developed including ReciprocalNet, developed by a consortium of crystallographic laboratories to share their results and facilitate reuse. The CrystalEye project, which offers around 100,000 datasets, harvests CIFs from the websites of journals published by the American Chemical Society, the International Union of Crystallography, the Royal Society of Chemistry plus some individual journals from other publishers. CIFs themselves are not subject to copyright, but publishers' formats are. The American Chemical Society has exercised its copyright over CIFs published in their journals presenting a major barrier to reuse for these data.

While researchers in a number of disciplines in the physical sciences struggle to find data, astronomers have few such concerns. Metadata standards are well established, enabling effective discovery of datasets. Every observation is assigned an identification number and there are established methods for referencing an individual object in the sky. When journal articles are published, datasets may be referenced using their archive accession number. While datasets can

be discovered fairly painlessly, accessing some types of datasets is more troublesome. The datasets created by observational astronomers become freely available shortly after they are collected, certainly after twelve months. Derived datasets, on the other hand, are less easily accessible. Some such datasets, such as those from small observational astronomy projects, are published *ad hoc* on project websites or sometimes not at all. If those data have not been mentioned in the published literature they will be very difficult to find or access. Where access is negotiated with individuals or small groups, it is usual for the data creators to request the opportunity to co-author formal journal articles with the researchers requesting the data. Some astronomers, particularly theoretical astronomers who make a lot of use of derived datasets, would prefer a system where all derived datasets are routinely disseminated, perhaps using arXiv as the locus. Overall, however, there is a strong culture of data sharing and an expectation that datasets exist to be found, accessed and reused by the astronomy community.

This propensity to share data is reflected in the development of the Virtual Observatory (VO). The aim of the VO is to make the world's astronomical data transparently usable. International standards are required to ensure that the rapidly increasing volumes of astronomical data can continue to be found and reused. The doorway to the VO is AstroGrid which provides "a suite of desktop applications to enable astronomers to explore and bookmark resources from around the world, find data, store and share files in VOSpace, query databases, plot and manipulate tables, cross-match catalogues, and build and run scripts to automate sequences of tasks."3

Physical Sciences Data Preservation

The field of climate science does not stand out as one where data sharing is a prominent aspect of the research culture overall, but people do recognise the high and persistent value of raw data collected using observational or remote sensing techniques – not simply because they can be very expensive to collect, but also because they are unique in terms of space and time. Not all such data find a home in a major data centre, representing a continuing challenge for funders and researchers. Data produced by climate modellers, on the other hand, are not traditionally shared or stored where they can be accessed by others and demand for such data is reported to be low. In any case, with an estimated useful life of around five years such datasets are not high on the curation priority list.

The Mesospheric, Stratospheric, Tropospheric (MST) dataset produced by the radar at Aberystwyth is good example of an irreplaceable earth observation record. It cannot be replicated through experimentation or model simulation. The case study focusing on the MST dataset presents a detailed preservation analysis methodology and highlights the importance of the role of an archivist dedicated to curating the dataset. Three preservation strategies are proposed: the first is an emulation strategy; second, the data could be converted to a compatible format, such as from NetCDF files to NASA AMES; third, by adding representation information. Whatever the preservation strategy selected, the key message is that the quality of preservation and therefore the value offered by the dataset is strongly linked to the subject and data management expertise of the person charged with looking after it. This reflects the views of

[3] http://www.astrogrid.org/

many: in order to curate datasets effectively the person or people responsible for doing so should have an understanding of the processes behind the creation of the data, the characteristics of the data, common data formats and other standards in the field and the range of purposes to which the data have application in terms of reuse.

Astronomers historically take a long term view of data collection and handling, perhaps because facilities take so long to build: for instance, a telescope can take up to fifteen years to construct and is expected then to produce data for decades. Most astronomical data are eventually publicly available, though there is a proprietary period following the collection of the data. Even so, the long term storage of data is not always optimal, some of it being kept on reels of tape. There are some big data centres such as those run by NASA and the ESA and the UK data centres maintain their own archives of data, even data collected at other facilities. The Wide Field Astronomy Unit (WFAU) in Edinburgh is one such data centre and has been the subject of one of the case studies covered by this report.

One of the key challenges in working with astronomical data is determining whether a certain feature in the data is real or whether it is a software processing artefact. Although data processing techniques are improving, it remains important to record the versions of all data processing software packages used in the generation of a particular data product. Similarly with original observation data, various problems can occur. For example, instrument settings in particular periods might have been sub-optimal, or data collected under certain observing conditions might be known to suffer particular defects. It is important, therefore, for the people reusing the data to be able to trace the history of a data produce right back to the telescope.

The WFAU data centre has aimed to preserve all the data produced by the programmes they cover to enable long term reuse. Astronomical datasets are normally very large and many of the analyses that astronomers want to run require more data than they can download to their local workstations. The data centre is looking to allow users to upload data analysis syntax which would be run at the data centre itself. Dedicated curators are needed to offer this level of service without risking the security of the datasets.

The long term funding of astronomical data is not guaranteed, though in this sense it is no different to other research disciplines. The Science and Technology Funding Council explicitly support the curation of recent and new data – though usually for one project at a time with tapering funding for data preservation after the project or mission finishes. Because of the e-Science aspects of astronomical research, it is hoped that the Engineering and Physical Sciences Research Council will become involved in funding astronomical data centres. On a smaller scale, individuals or small research groups tend to look after derived datasets on local systems and there is rarely budgetary provision for looking after those data for the long term. There is also concern about the long term accessibility of observational data since the demise of the UK Starlink Project when its funding ceased. The project provided and coordinated interactive data reduction and analysis software and facilities which enabled researchers to make use of curated datasets. This highlights an important issue in the data preservation arena: even where funding for data preservation can be secured, who should pay for the software and hardware required to extract value from those data?

In contrast to the range of data storage facilities available to astronomers, the situation for the engineering research community is very different. The department-focused case study reviewed for this report indicates that the outputs from the department are stored locally on a variety of hard disk drives and removable storage media. There are moves to store project data on a dedicated network drive where project folders will be kept for seven years. Some are protected by conventions which require that data are never overwritten and that cleaned or processed versions of existing datasets are saved separately. The metadata associated with the project folders containing data are very basic.

Crystallographers, on the other hand, benefit from the professional curation practice brought to bear by the big databanks like the CCDC. While the CCDC is financially self-sustaining, as is the pattern with other disciplines, smaller repositories of data are more vulnerable. An example of this vulnerability is provided by the case of the Chemical Database Service based at the Daresbury Laboratory. This held a variety of chemical information along with some data relevant to crystallography which was provided free to UK academics. The facility was liked by crystallographers because of the speed with which they can access data. Even though this was one of the National Services funded by the Engineering and Physical Sciences Research Council, the Council announced the closure of the database towards the end of 2006 despite protest from the user community. Ultimately there was a partial reprieve: crystallography content was spared while all the other data is no longer available. This experience serves to emphasise the vulnerability of datasets and the uncertainty that surrounds their long term viability.

CONCLUSION

This review of a number of case studies which cover various aspects of the way in which research data is discovered, shared and preserved provides useful insights into the disciplinary differences that exist. These differences exist for a wide variety of reasons:

- the heritage and practices of niche research communities
- the type and quantity of data they produce
- the uniqueness of those data and their potential value in terms of reuse
- the propensity of each community to create, adapt or adopt common data formats, metadata schema and other relevant standards
- their willingness to share data in a world where competition for funding looms large
- the policies of funding bodies in relation to data management, sharing and preservation
- the provision of storage infrastructure including national data centres and effective discovery systems

The size of research teams (larger teams can benefit from keeping its own data private)
Some fields are more diverse, so there is less competition on the micro scale
The list could go on but whatever the reasons behind the disciplinary differences, they clearly do exist. This reality gives rise to the question as to what should be done to optimise data curation within each discipline, to the extent that it is deemed valuable to so do. The scale of the task appears, on the face of it, to be large and complex. That said, this review demonstrates that there are disciplines that are currently reasonably well positioned with respect to data curation and which, with continued or additional funding, will continue to offer an adequate or good service. Bringing other disciplines up to speed requires more thought and investment particularly in terms

of whether and how to provide national solutions to curating the most important, valuable datasets in each discipline. This report concludes with some general observations about how progress might be made in light of the information gleaned from the case studies.

Attitudes to data management and sharing

In some disciplines dealing with data is part of the day job: genomics, systems biology, astronomy and crystallography to name a few. In these areas data management is central to the research and this is recognised and reinforced by the policies of most of the relevant funding bodies. In other disciplines, perhaps the majority, data is a stepping stone to the production of articles published in journals. Even where funding bodies have data management policies and facilities to aid and encourage compliance, such stipulations are sometimes viewed by researchers as bureaucratic obstacles to be negotiated in the quest for research funding. At present, researchers' compliance with data-related funding policies in the post-award phase is not especially well monitored and sanctions for non-compliance are rarely applied. In order for the role and value of data management to be internalised by researchers in some disciplines, there needs to be a more explicit link between the effort that is required to manage and share data and career recognition and rewards. For now publications reign supreme in this respect. If more weight was to be given to good data management, patterns of behaviour may shift over time. The widespread adoption and recognition of a system for directly citing datasets may also help change researchers' behaviour with respect to managing the data they produce and use.

Infrastructure for data curation

Researchers are probably not best placed to be responsible for the long term viability of datasets. Their expertise lies in research; the sensible division of labour suggests that responsibility for data curation should rest on other peoples' shoulders. In many disciplines there are data centres which, while their capacity is limited, can professionally curate datasets deemed valuable to the research community. The problem is what to do with the "small science", the datasets produced by individual researchers and smaller research groups which have neither the funds nor the infrastructure to look after datasets beyond the timeframe of their research project. It is often suggested that institutional repositories are the natural locus for such datasets. In the UK the majority of higher education institutions now have repositories in place, though their use for digital research data curation is very limited at present. The benefit of this route to data curation is that the risk of datasets disappearing over time through lack of care or resources is spread across a whole institution and is, therefore, diminished. Looking after the data outputs of their research community is a key strategic challenge for institutional managers and repository administrators, one that is likely to involve changes in organisational structure and culture. Whether the will, skills and resources necessary exist to meet this considerable challenge within individual institutions remains to be seen, but two initiatives are helping to light the path ahead: JISC's Digital Preservation and Records Management Programme[4] is funding research projects in this field, and the UK Research Data Service (UKRDS) is currently planning for the

[4] http://www.jisc.ac.uk/whatwedo/programmes/preservation.aspx

development of a pathfinder service, a step towards building a national shared digital research data service[5].

Expertise in data curation

It is clear from reading the case studies that the quality of datasets themselves and the likelihood of them being viable for the longer term is related to the disciplinary and data-related expertise of the people charged with responsibility for those datasets. They could have discipline-specific backgrounds and be embedded within research groups, departments or stand-alone units (and may be called data scientists or data managers) or they might be information experts employed by a data centre or library with a remit to reach out to and develop close ties with one or two disciplines or departments. Suitably descriptive job titles for these roles are still evolving. These data professionals may need domain-specific, task-specific and cross-cutting, general data curation training and support. Some research councils are able to offer support at a domain level as are many data centres, while at a national level the Digital Curation Centre offers strategic direction, advice on best practice and organises relevant workshops and conferences to bring together distributed, sometimes disparate sources of expertise. Data scientists and data librarians are in short supply but demand for the skills they offer is sure to grow. Work to develop appropriate training course for librarians is under way though there is still some way to go before these efforts bear fruit in sufficient quantity. There needs also to be suitable training and career opportunities and incentives to enable researchers to develop into effective data scientists. If the research and research support services communities are to accommodate the inevitable data deluge, manage it so that the potential of the data is maximised, and preserve it for an appropriate period of time, for many disciplines there remains much to do.

[5] http://ukrds.ac.uk/home

Links to the case studies

The following DCC-sponsored case studies have been reviewed for this project. The summary information about each cast study reflects the information provided on the DCC's website: http://www.dcc.ac.uk/resource/case-studies/

- SCARP Case Study No. 6 - Digital Curation approaches for Architecture

This study highlights choice in how to provide for appropriate care of digital objects, choice in digital curation treatments, as a means of promoting more effective current and future architectural practice and research. The digital assets produced by use of digital tools and from digital methods of working in the teaching, learning, research and practice of Architecture require appropriate curation treatment if the full value of the assets is to be realised.

- SCARP Case Study No. 5 - Roles and Reusability of Video Data in Social Studies of Interaction

The study reviews the curation landscape in several interdisciplinary fields that use video analysis in studies of human interaction. A 5 page Summary and Conclusions **[PDF, 136KB]** is also available. The study primarily focuses on uses of video in ethnographic studies and in eye movement research, and is based on interviews and field study.

- SCARP Case Study No. 4 - Curated Databases in the Life Sciences: The Edinburgh Mouse Atlas Project 13 July 2009 | Elizabeth Fairley

This study scopes and assesses the data curation aspects of the Edinburgh Mouse Atlas Project (EMAP), a programme funded by the Medical Research Council (MRC). The principal goal for EMAP is to develop an expression summary for each gene in the mouse embryo, which collectively has been named the Edinburgh Mouse Atlas Gene-Expression Database (EMAGE).

- SCARP Case Study No. 2 - Curating Atmospheric Data for long term use: Infrastructure and Preservation Issues for the Atmospheric Sciences community 2 June 2009 | Esther Conway

This study engaged with a number of archives, including the British Atmospheric Data Centre, the World Data Centre Archive at the Rutherford Appleton Laboratory and the European Incoherent Scatter Scientific Association (EISCAT). We developed a preservation analysis methodology capable of identifying and drawing out discipline specific preservation requirements and issues. We present the methodology along with its application to the Mesospheric Stratospheric Tropospheric (MST) radar dataset, which is currently supported by and accessed through the British Atmospheric Data Centre. We suggest strategies for the long-term preservation of the MST data and make recommendations for the wider community.

- SCARP Case Study No. 3 - Clinical Data from Home to Health Centre: the Telehealth Curation Lifecycle 28 June 2009 | Tasneem Irshad, Jenny Ure

This study looks at the data curation lifecycle in Telehealth research. Telehealth, or telecare, is an emerging sub-domain of eHealth, and the report profiles current practices in several telehealth pilot projects. Data curation is at an embryonic stage but can draw on related eHealth initiatives and clinical data management practices, and the report considers the infrastructure needed for data curation in this field of research and practice.

- SCARP Case Study No. 1 - Curating Brain Images in a Psychiatric Research Group: Infrastructure and Preservation Issues 14 November 2008 | Angus Whyte

This study involved the Neuroimaging Group in University of Edinburgh's Division of Psychiatry. It combined an assessment of risks to the long-term value of the research group's datasets with field work to understand current data practices in their context. A 5-page Summary and Recommendations [PDF, 170KB] is available. An annex Neuroimaging Data Landscapes [PDF, 1.23MB] provides background on the development of imaging, the nature of the data collected for neuroimaging studies in psychiatry, data repository and curation resources available, and legal and ethical constraints on data exchange.

- CARMEN (Code, Analysis, Repository and Modelling for e-Neuroscience)

  28 November 2008 | Graham Pryor

- Integrative Biology

  7 April 2008 | Martin Donnelly, Victoria Boyd, and Jill Spellman

Computer simulation organ functions based on models offers the potential to increase researchers' understanding of the causes of medical conditions such as heart disease and to work towards developing new drugs and treatments to combat these illnesses. The main aim of the Integrative Biology (IB) project is to realise this potential by developing multi-scale models - spanning the range from genes to whole organs - and to provide data management features for its disparate users including the sharing of data in a secure infrastructure, and enabling the storage and re-use of simulation outputs.

- PrestoSpace 19 March 2008 | Martin Donnelly, Victoria Boyd, and Jill Spellman

Explicit strategies are needed to manage 'mixed' audio visual (AV) archives that contain both analogue and digital materials. The PrestoSpace Project brings together industry leaders, research institutes, and other stakeholders at a European level, to provide products and services for effective automated preservation and access solutions for diverse AV collections. The Project's

main objective is to develop and promote flexible, integrated and affordable services for AV preservation, restoration, and storage with a view to enabling migration to digital formats in AV archives.

- JHOVE 5 April 2006 | Martin Donnelly

Accurate file format information is crucial for preserving access to and the rendering of digital information over time. As such, it is vital that when a digital object is deposited in a repository, the object in question is of the type it purports to be. However, the representation of file formats is easily corruptible - whether accidental or intentional. This is of particular concern to institutions with an interest in preserving digital materials in repositories. The JSTOR/Harvard Object Validation Environment (JHOVE) is an Open Source, extensible framework for the format-specific identification, validation, and characterisation of digital objects.

- Wide Field Astronomy Unit

  8 December 2005 | Martin Donnelly

Two planned SCARP studies on astronomical data could not be completed for health reasons. This older case study was used instead; this may affect the currency of some of the material relating to astronomy in the main body of the report. The Wide Field Astronomy Unit (WFAU) creates and curates astronomical data, serving a large community of data re-users. Linking its databases with the nascent Virtual Observatory, WFAU collections are made available to the entire community. Regular curation tasks include loading catalogues into a database, matching them with prior observations, preparing data for publication via a web interface, and occasional recalibration and replacement. The data products extracted from the WFAU archives are mostly held in the Flexible Image Transport System (FITS) format. The set of keywords used in FITS metadata records is defined only by weak constraints, not by a controlled list. As a result, records can be difficult to interpret beyond the institution which generated them, creating potential problems for data centres which ingest quantities of externally created data.

- Other reports

The full case study reports for the following disciplines and cross-disciplinary fields were published by the Research Information Network in 2008, written by the authors of this synthesis, and can be found via the link to the appropriate RIN webpage below the list:

- Classics
- Rural economy and land use programme
- Social and public health sciences
- Genomics
- Systems biology
- Astronomy
- Chemical crystallography

- Climate science

http://www.rin.ac.uk/our-work/data-management-and-curation/share-or-not-share-research-data-outputs

Work in progress:

The following case studies are currently being finalized, though summaries or drafts were made available for use in this project.

- Curating Engineering Data – synthesis summary. Alex Ball, Colin Neilson, UKOLN, University of Bath