 ISSN 1759-586X

Curating Brain Images in a Psychiatric Research Group: Infrastructure and Preservation Issues

SCARP Case Study No. 1

Angus Whyte
Digital Curation Centre, University of Edinburgh

DCC SCARP INTERIM CASE STUDY REPORT
Deliverable B4.8.2.1

Version No.	1.1
Status	**FINAL**
Date	11 November 2008

Copyright

Catalogue Entry

Title Curating Brain Images in a Psychiatric Research Group: Infrastructure and Preservation Issues
Creator Angus Whyte (author)
Subject Data curation; formats, processes and issues; system development; standards; legal factors; methodology, and problems overcome; human factors
Description Curating neuroimaging research data for sharing and re-use involves practical challenges for those concerned in its use and preservation. These are exemplified in a case study of the Neuroimaging Group in the University of Edinburgh's Division of Psychiatry. The study is one of the SCARP series encompassing two aims; firstly to discover more about disciplinary approaches and attitudes to digital curation through 'immersion' in selected cases, in this case drawing on ethnographic field study. Secondly SCARP aims to apply known good practice, and where possible to identify new lessons from practice in the selected discipline areas; in this case using action research to assess risks to the long term reusability of datasets, and identify challenges and opportunities for change.
Publisher University of Edinburgh; UKOLN, University of Bath; HATII, University of Glasgow; Science and Technology Facilities Council.
Contributor Stephen Lawrie
Date 11 November 2008 (creation)
Type Text
Format Adobe Portable Document Format v.1.3
Resource Identifier ISSN 1759-586X
Language English
Rights © 2008 DCC, University of Edinburgh

Citation Guidelines

Whyte, A. (2008), "Curating Brain Images in a Psychiatric Research Group: Infrastructure and Preservation Issues ", SCARP Case Study 1, Digital Curation Centre, Retrieved <date>, from http://www.dcc.ac.uk/scarp

Contents

Executive Summary

Curating neuroimaging research data for sharing and re-use involves practical challenges for those concerned in its use and preservation. These are exemplified in a case study of the Neuroimaging Group in the University of Edinburgh's Division of Psychiatry. The study is one of the SCARP series encompassing two aims; firstly to discover more about disciplinary approaches and attitudes to digital curation through 'immersion' in selected cases, in this case drawing on ethnographic approaches. Secondly SCARP aims to apply known good practice, and where possible to identify new lessons from practice in the selected discipline areas; in this case using action research to assess risks to the long term reusability of datasets, and identify challenges and opportunities for change. The Neuroimaging Group is involved in several collaborative e-science initiatives to improve data sharing and re-use in their discipline. At the same time a key issue for them is improvement of local infrastructure to address their expanding digital curation needs.

Study Scope and Contents

The first chapter of this case study report introduces four themes: -

- Data policy drivers, enablers and barriers
- Data stewardship practices
- Curation tools and infrastructure
- Preservation of contextual and provenance information

The chapter relates these themes to literature on neuroimaging research in psychiatry and its rationales for data sharing and re-use. The Annex to the report *Neuroimaging Data Landscapes* (Whyte, 2008b), reviews in more depth the development of imaging, the nature of the data and the limited curation resources available, and legal and ethical constraints on data exchange. It also further describes and reflects on the methods used in the case study.

Chapter Two further describes the Neuroimaging Group in this case, and why digital curation is of interest to its investigators. The group researches major psychiatric disorders, and is particularly known for work in schizophrenia. Neuroimaging studies typically follow a case-control design. Study data is mainly observational; relating brain images captured at particular points in time to related clinical and demographic data. Studies of brain function combine these observations with data of a more experimental form, gathered from subjects' responses to stimuli. The Group has been gathering MRI (Magnetic Resonance Imaging) data over a relatively long period and has acquired a wide range of clinical and demographic data, resulting in large data volumes (approx 9TB in several million files, at the time of the study).

Chapter Three reports on how DRAMBORA- a risk assessment approach for digital repositories -was applied along with the OAIS functional model for archival information systems, to help the Group compare their own data management activities with those recommended for a data archive, which the UK currently does not have in this domain. Risks are mapped to identified activities and digital assets. The DCC Curation Lifecycle model is used to take stock of the Group's current measures to address risks to data.

Chapter Four considers and recommends next steps for curation and preservation of the Group's datasets and a phased approach to supporting data documentation, including the scope of that documentation and high-level system requirements. These take

account of the human infrastructure underpinning data sharing and curation in the Neuroimaging Group.

Chapter Five looks further at the local practices of data sharing and re-use, and their role in the socio-technical infrastructure for data preservation. Neuroimaging in psychiatry depends on close interaction between researchers from various disciplinary backgrounds. By interacting 'heedfully' [1] researchers help to ensure that knowledge of datasets and experimental protocols is passed reliably from peer to peer, and from more established researchers to newcomers, enabling continuity in research and flexibility in project membership.

Report Conclusions and Recommendations

This is an interim report from SCARP; its recommendations will be considered by the DCC and appropriate actions taken following discussion of strategy and resource implications. The conclusions and recommendations for DCC and research policy-makers follow the themes below. These acknowledge the limits of any qualitative study of one laboratory (the accompanying report reflects on these in more detail). The particulars of the case illustrate and exemplify themes evident in recent neuroimaging literature, and draw on the participants' knowledge of the neuroimaging community, but they do not seek to make the kinds of generalisation from sample to population that is characteristic of quantitative survey research.

"Think global, act local" to build metadata exchange capabilities

Curation needs human infrastructure and this should be taken into account when assessing curation capabilities. The study shows how researchers and investigators heedful attention to each other's data underpins curation. Neuroimaging involves continuous care of increasingly large and dynamic datasets. Neuroimaging investigators are custodians of millions of images and, to contribute to medical research, these need to be related to richly varied and highly sensitive personal information on research subjects. Some of that data is being shared, including in e-science projects aiming to provide federated data storage and improve data integration (see below). The large majority of data is held at lab level however, with access governed by Principal Investigators under terms set by Research Ethics Committees. Compliance with these terms and protecting personal data is of more immediate concern to researchers than sharing data with independent researchers in other laboratories or fields. Rather, data tends to be shared on a quid pro quo basis both within the laboratory and with external collaborators, when legal and ethical constraints allow it and there is evident benefit to be gained from exchanging access to data and/or analytic methods. It would be more accurate to see this as a form of 'gift exchange' between data custodians than as 'sharing'.

Interest in re-using datasets is mainly in the areas of using novel analysis techniques to identify patterns in images or in the associated clinically-related and demographic data on subjects, and (among the researchers interviewed) less in re-using derived data to replicate previous analyses. Documentation and metadata on research subjects and on analytic protocols is key to any form of re-use, and is encouraged by the ethics compliance regime. Images, associated subject data, and structured contextual and provenance information about these need to be inter-related. Lack of that structured, standardised documentation is a major source of risk to datasets' long-term reusability, yet this is an area that is reportedly under-invested in.

[1] Applying the ordinary sense of this word to work with datasets, i.e. carefully, consistently, purposefully, attentively, studiously, vigilantly, conscientiously.

Standardisation in neuroimaging methods and data documentation is driven by the need for larger datasets to enable studies with higher reliability. This requires larger-scale collaboration and hence wider trading of methods and data. The top-down data sharing policy framework put in place by the MRC and Wellcome Trust needs to be accompanied by further ground-up initiatives to exchange semi-structured data between imaging centres. Neuroimaging research has strong potential to benefit from e-research tools and infrastructure, as the large-scale U.S. investment in BIRN indicates. Borrowing the environmentalist slogan, there is a need for U.K. research funders to "think global and act local" to support the development of data curation in this domain. The UK neuroimaging community is well-placed to further develop models for achieving that, following the examples of Neurogrid, PsyGrid, NeuroPsyGrid and Carmen. However it needs investment in tools to support a gradual transition from inter-personal and study-level sharing of neuroimaging metadata to wider dataset 'trading' and collaborative re-use. Such tools should be simple to deploy and use by neuroimaging researchers. They should enable researchers to structure their study documentation and link it to relevant datasets, and to make the resulting metadata selectively and securely available; and they should enable potential collaborators to easily find relevant studies through metadata aggregation services.

Data integration drives new curation requirements

Multi-centre neuroimaging collaborations are augmenting existing curation capabilities, adding value to datasets by enabling them to be integrated for re-analysis purposes, and fostering innovations in image analysis through transfer of techniques from informatics disciplines. Examples include development of image normalisation techniques to harmonise image data from multiple scanners, and automated analysis of images to enhance productivity. These in turn add to the variety of contextual and provenance information needed to track data as it is integrated from disparate sources and analysed by multiple people and/or centres.

Frequent change in the analytic methods used in neuroimaging makes the need for structured documentation more acute. Community standards for recording provenance and representation information are urgently needed in the neuroimaging community, and transferable techniques are likely to be found across other fields of image-based research. Meanwhile, effective exchange of data and methods is likely to be hampered by inevitable changes in the schemas used to describe these.

Recommendation 1 ~ DCC should further investigate and map provenance information management requirements in neuroimaging and other fields of image based research, to provide better advice on tools and methods to address these requirements.

While novel analysis techniques make retrospective analysis of imaging datasets increasingly promising, this makes appraisal of the value of imaging dataset more complicated. For example Neuroimaging Group researchers have reported achievable benefits from using ontologies to combine MRI datasets across centres, to enable cross-analysis of psychosis and other datasets. Researchers and funding bodies need to make informed decisions about whether greater value is obtained from gathering new data or re-using the old in new ways. This coincides with an increasing need to appraise the value of data amassed from long running longitudinal studies that have been sustained through successive projects and custodians.

Recommendation 2 ~ The neuroimaging community requires further support to assess the viability and usefulness of combining existing MRI data sets on psychosis and other neuropsychiatric disorders.

Recommendation 3 ~ DCC should further investigate and map factors that affect the value of reusing imaging datasets, to enable that value to be measured and support better advice on appraising and valuing datasets.

Recommendation 4~ DCC should develop and provide guidelines, advice and templates for data access policies, using neuroimaging as an exemplar of the challenges of reconciling the requirements for data confidentiality and more open access in medical research. This should be supported by stakeholders such as the MRC Data Support Service, and is in keeping with the recent interim report of the UK Research Data Service Feasibility Study (SERCO, 2008), which identifies a requirement for more advice on practical issues related to managing data, including help producing data management/ sharing plans.

Integrating 'good curation practice' into research training

Neuroimaging labs are interdisciplinary communities of practice whose members need to share data and skills. That is especially so for newcomers, who are required to reuse datasets and research protocols to learn the practical skills of image analysis. Junior researchers learn by participating in colleagues' studies, directly benefit from sharing experimental protocols, and could play an active role in standardising study documentation and collecting metadata. Integrating these tasks into research supervision may benefit students by helping them identify the characteristics of datasets that are essential to re-use, while also alleviating the bottleneck that manual metadata creation is regarded as by senior researchers. Ethical clearance procedures engender thorough documentation of research protocols at the outset of projects, providing an opportunity to link training on these procedures with training on curation lifecycle management, adapted to meet the needs of the neuroimaging field.

Recommendation 5 ~ DCC should support the development of digital curation in neuroimaging and related fields by providing curation lifecycle management training targeted at doctoral or masters students and briefing materials targeted at research supervisors.

Risks to dataset reusability reflect the disciplinary mix in neuroimaging; clinicians and imagers have tended to manage different kinds of data; while clinicians are data custodians concerned with close personal management of demographic data, imagers have historically required network servers and archiving resources to manage larger image datasets. The case for integrating demographic and imaging datasets coincides with growing convergence between the neuropsychiatric and imaging domains, e.g. as imagers have developed capabilities to contribute to the psychiatric domain.

The report demonstrates the need for case studies of how "enablers and barriers" to data sharing, curation, preservation and reuse operate on the ground in particular research communities. For example the current study has documented how the 'lack of standardisation of neuroimaging methods' reported in the neuroinformatics literature affects data sharing between early career lab researchers with differing skills levels or disciplinary backgrounds. A focus on how newcomers attain membership of research communities also helps to address one of the major difficulties of 'immersive' case studies- that they require an understanding of the terminologies and competencies needed to do research in the host research community. Relatedly, if case studies are to benefit host teams they require easily and quickly transferable tools to apply 'best practice' in digital curation. In the current case DRAMBORA needed some adaptations to apply it outside of its main target group of established archival organisations'.

Recommendation 6 ~ DCC should adapt the DRAMBORA risk assessment tool to enable it to be easily used by data custodians at the department or research team level.

1. "Little Big Science"- Neuroimaging in Psychiatry

1.1 Introduction and background to the study

Given the increasing importance attached to curating and preserving digital research data for informed reuse, further study is needed of researchers' practices and how these vary across disciplines (Borgman, 2007). A recent Research Information Network report makes broad disciplinary comparisons and concludes:

> "In developing their policies, research funders and institutions need to take full account of the different kinds and categories of data that researchers create and collect in the course of their research, and of the significant variations in researchers' attitudes, behaviours and needs in different disciplines, sub-disciplines and subject areas..." (Research Information Network [RIN], 2008).

The SCARP case studies, funded by the JISC, contribute to this area with a focus on a range of disciplines including medical and social sciences; and on four themes:

1) *Policy drivers, enablers and barriers:* organisational and institutional factors including different skill levels, preservation policies and arrangements, willingness to use these, and relationships to incentives and reward structures.

2) *Stewardship practices:* how the research process and methods relate to the primary data created and external sources, how these are reused and linked to publications, attitudes to doing this, the usefulness of prior data, and the sustainability of collected digital information.

3) *Tools and infrastructure:* tools and facilities used to collect, deposit, find, cite, discuss and annotate the data, and to ensure persistence and preservation.

4) *Preserving context:* how communities of practice and their knowledge bases can be characterised, and how lineage and provenance is or may be documented.

The study aimed to be "immersive", using a qualitative approach combining ethnographic field study in the research context with "appreciative intervention" to facilitate change, drawing on action research traditions (e.g. Karasti, 2007). Field study data was gathered using 20 semi-structured interviews with a cross-section of Group members, and by observing meetings over five months. In parallel, a data risk assessment was facilitated using the DRAMBORA approach (Digital Curation Centre [DCC]/Digital Preservation Europe [DPE], 2007) and the Digital Curation Lifecycle (DCC, 2008), leading to recommendations for new measures to address risks.

This first chapter surveys the 'landscape' of neuroimaging as background to the case. The chapter is a brief summary of a more extensive literature review is available in the accompanying SCARP report *Neuroimaging Data Landscapes* (DCC, 2008), which reviews the development of imaging as a technique for psychiatric research, the nature of imaging studies, the range of primary data collected, and the techniques used to derive analytic data. That report describes national and international collaborative efforts to share neuroimaging data and harmonise data analysis techniques and terminology, together with legal and ethical constraints on the exchange of data between researchers. It also contains a brief description of the methods used in the case study.

The Neuroimaging Group in Edinburgh University's Division of Psychiatry researches major psychiatric disorders, and is particularly known for schizophrenia research.

Neuroimaging studies typically follow a case-control design; subject groups with a positive diagnosis are compared with groups at high risk, plus healthy controls (Lawrie et al., 2005). The Group has unusually large and rich datasets. For example the longitudinal Edinburgh High Risk Study (Johnstone, Russell, Harrison, & Lawrie, 2003) includes social and economic classification data, information on family history and life events, and on alcohol and drug use for over 200 subjects. Clinical and behavioural data includes diagnoses and case history, psychiatric assessment, performance in IQ and other cognitive tests. The majority of participants were seen on several occasions over up to ten years. Subjects in this and other studies have also given genetic data to illuminate the heritable characteristics of psychiatric disorders.

Access to the Neuroimaging Group was agreed with Stephen Lawrie, Professor of Neuroimaging in the Division of Psychiatry, after earlier contacts had established some synergy between the Group's work in e-science projects and SCARP's aims. What was particularly interesting from the latter point of view was that large volumes of medically-related data were being handled; that there was awareness of data curation as a specialist task and current issue; and that SCARP was seen as an opportunity to review the Group's existing practices.

The study boundaries were the Edinburgh group's curation practices, rather than those of e-science projects they collaborate in, although these play an important role in developing local curation practices. This might be expected since e-science or 'eResearch' has, according to Day (2007) "accelerated the collaborative nature of science, and large-scale collaborations are no longer just typical of traditional 'big science' disciplines like high-energy physics or astronomy, but have become an important part of recent initiatives in chemistry, bioinformatics, healthcare and other disciplines". Collaboration, trust and the organisational and technical infrastructures to support it are also regarded as key to the development of digital curation and a focus for recent work in scholarly communication (Fry 2002, Borgman 2007).

1.2 Neuroimaging and Psychiatry

Neuroimaging in psychiatry focuses on finding neurobiological explanations of psychiatric disorder (Lawrie, Weinberger, & Johnstone, 2005). The rationale is that imaging techniques can depict differences at one point in time between groups of patient and control brains, or sometimes changes over time in brains, which may then be correlated with a range of measures of behavioural, social and clinical phenomena.

The SCARP study took place against a background of medical research funders' interests in improving data curation and sharing. The Medical Research Council and Wellcome Trust, major UK funders of neuroimaging research and of psychiatry, both of which are relative UK research strengths, recently published policies on documentation and sharing of medical research outputs (Medical Research Council [MRC], 2007). These establish principles for grant holders and roles of data creators and custodians; to curate datasets throughout their lifecycle, make them available with few restrictions, and with sufficient information for informed reuse. Custodians are called on to provide transparent access policies, while complying with the research ethics approval process, which places limits on the kinds of data that may be gathered, their processing and retention. An important factor in studies involving (psychiatric) patients is that any risk of the loss of medical confidentiality must be minimised (MRC, 2007).

The MRC also funds e-science projects in the UK to permit data sharing by providing an infrastructure to integrate neuroimaging datasets. While various imaging techniques have been used in psychiatric research, MRI (Magnetic Resonance Imaging) has

become predominant. MRI has provided a means to investigate brain structure without surgical or even X-Ray exposure and, with the introduction of "functional" MRI, to couple that with studies of brain processes (Pekar, 2006).

MRI 'slice' images are acquired from scanners by radiographers, and commonly stored in the DICOM standard medical image file format[2]. A structural MRI image highlights the spatial distribution of brain tissue components; enabling structure to be mapped against standard templates and potentially tracked through repeated scans. Three-dimensional images of the brain are "reconstructed" from individual "slices" of the head, captured digitally from scanners that subject the research participant to intense magnetic pulses. Functional (fMRI) studies measure the flow and oxygenation level of blood in the brain, which change in response to task "stimuli" participants/ subjects are asked to respond to inside the scanner. fMRI scanning sacrifices some spatial image resolution for the added dimension of time, building up a movie-like sequence (Pekar, 2006).

Structural MRI studies are *observational*, being concerned with capturing individuals' brain anatomy at particular points in time. Functional MRI studies on the other hand also encompass e*xperimental* methods, since brain functions are analysed in relation to a task stimulus hypothesised to affect them.

At the design stage *ethical approval* must be obtained from the relevant NHS Research Ethics Committee, to carry out the research while safeguarding the interests of the human subjects it intends to involve. The subjects – whether patients, healthy relatives or control groups - must of course also be recruited and their informed consent obtained to take part in the study. These legal and ethical requirements mean that studies have to be designed to a high degree before any data is actually acquired.

Various *pre-processing* steps are performed on images before they may be used in analysis (Toga 2002, Van Horn 2004). A common first step is *anonymisation* to remove metadata identifying the subject, included in the DICOM file header of each scan, before *reconstruction* of the three-dimensional brain volume from the slice data. Subsequent processing is directed at reducing the amount of 'noise' in this data. Small movements of the head are typical and image-processing software is used to correct for this. Because of the additional time factor in functional studies, differences in slice timing and other motion-related effects also need to be compensated for. Since functional scans are of relatively low resolution, they are aligned or *co-registered* with structural scans obtained at the same time.

To compare images between individuals, the next step is to map the structural (or co-registered functional and structural) images to a common set of spatial coordinates for the brain. This is referred to as spatial *normalisation* or 'brain warping', and normally involves reference to a brain atlas known as the Talairach system (Toga, *ibid.*). The result is, for each brain volume, a set of x, y and z coordinates for each voxel[3], matched to a known neuroanatomical region with a degree of statistical significance. As a final step before statistical analysis, smoothing algorithms may be applied to improve the signal-to-noise ratio (Van Horn, 2004). Statistical analysis typically follows three stages;

[2] The DICOM (Digital Imaging and Communications in Medicine) Standard is published by the US- based National Electrical Manufacturers Association (http://dicom.nema.org/). The standard encompasses a file format that includes patient metadata in the file header, and a TCP/IP based network communications protocol, enabling scanners to be linked to other hardware devices across a network, and integrated into a Picture Archiving and Communications System (PACS).

[3] The term *voxel* is used in imaging as an abbreviation of 'volumetric pixel'. Where a pixel is a unit of two-dimensional space visualised on a computer display, a voxel is a unit on a three dimensional grid used to represent a volumetric dataset or object (Kaufman et al, 2003).

firstly to find individual differences between subjects, and secondly the differences within the study's groups of subjects. The third level is analysis of any statistically significant differences between these groups that would test the study hypothesis, or lead to further hypotheses.

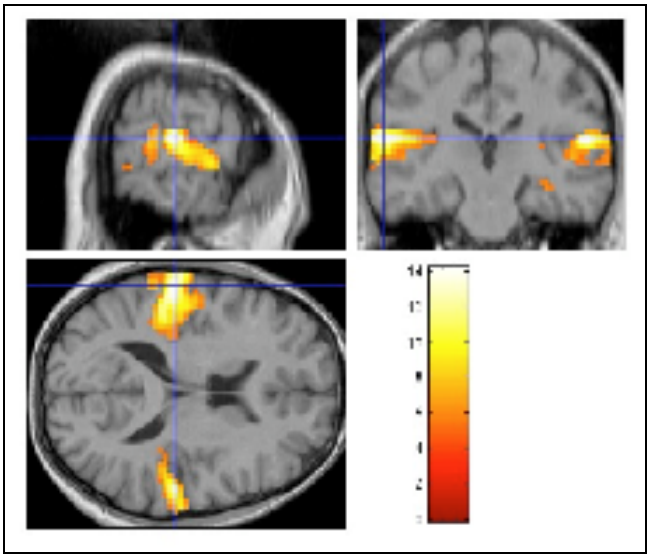

Figure 1. Example of overlays produced by the SPM 5 neuroimaging analysis software (Ashburner et al 2008 p.197) © Wellcome Trust Centre for Neuroimaging

Neuroimaging analysis software typically provides *overlays* like those shown in Figure 1 - maps of the analysed brain volumes that use colour gradations to highlight statistically significant changes in brain activity. Researchers typically include these in publications along with tables of relevant statistics on individual or group differences, as the final results of data reduction (Van Horn 2004).

1.3 Data Sharing Resources and Risks

Neuroimaging is used in a wide variety of neuroscience research involving humans and other animals. In psychiatry in particular, MRI scan data is of little use without detailed information about the person scanned. Since neuroimaging studies aim to find variables that explain differences in their brain structures and activity, a great deal of quite complex data is gathered on the research subjects or participants (Keator et al, 2006) The data gathered includes *demographic* data – typically that with some bearing on brain size and physiology, such as age, sex, height, and handedness. Also gathered are data with some known relation to mental health; for example the Edinburgh High Risk Study (Johnstone et al 2003) includes social and economic classification data, information on family history and life events, and on alcohol and drug use. In addition to this, *clinical and behavioural* data gathered includes genetic data (as some psychiatric disorders are held to be inherited), any diagnostic or case history, current psychiatric assessment, and performance in IQ and other cognitive tests.

Neuroimaging researchers are increasingly seeking to integrate datasets from different centres through collaboration in multi-centre studies, to improve the statistical power and reliability of research findings from larger study populations than single centres could feasibly recruit. Integrated datasets provide a wider range of clinical, behavioural and demographic data to identify and correlate variables. Dataset integration is a prime target of e-science projects such as the UK-based Neurogrid and NeuroPsygrid and US-based BIRN (Biomedical Informatics Research Network). The cost efficiencies of multi-centre studies are a further incentive: the possibilities of retrospective meta-analysis

underpinning work on effective data mining (Keator, Gadde, Grethe, Taylor, & Potkin, 2006; Ure et al., 2007).

Various factors however confound data integration: scanners vary in magnetic field and image intensity, centres may recruit from markedly different populations, and adopt any of a number of different scales to measure (for example) psychotic symptoms. Also there is wide variation in image analysis tools - hence projects increasingly focus on standardised tools to harmonise methods, normalise scanner output, coordinate quality assurance, and bridge symptom scales (Ure et al., 2007).

There are obstacles to sharing neuroimaging data apart from the barriers to integration, including concerns about disclosure of confidential data. A key issue for Gardner et al. (2003) is that neuroimaging data reuse is relatively straightforward, but susceptible to misinterpretation with insufficient representation of the original experimental context. As a result;

> "…the scope of shareable data may legitimately vary depending upon the standards and practices of different fields or techniques, and may thus include or exclude any or all of 'raw', partially processed, processed or selected datasets. Ideally shareable data should be defined as the combined experimental data and descriptive metadata needed to evaluate and/or extend the results of a study" (Gardner et al., 2003, p.291).

This indicates the early stage of standards for experimental context metadata, dataset structure and content (Gardner et al., 2003) reflecting the rapid pace of change in this field. Neuroimaging laboratories tend not to have invested in database technologies, and according to Geddes et al. (2006) data curation in neuroimaging research tends to be poor. Large-scale curation and publication of datasets have however been embarked upon by US and international collaborations, including fBIRN (Keator et al., 2006). Some databases provide canonical reference data: web-based brain atlases and coordinate systems, and statistics representing norms of brain structure or function. Other databases provide primary data or derived results from studies to support meta-analysis (Toga, 2002). While the UK currently lacks established data centres to support domain archiving, the MRC-funded e-science projects are developing services intended to be sustainable (although it was not the study's remit to assess that). The MRC is also establishing a data support service, and supporting the Mental Health Research Network's *Cohort Dataset Directory* (Mental Health Research Network [MHRN], 2007).

The need to safeguard patient confidentiality is paramount in arrangements for data sharing. Research councils provide specific guidelines on the levels of anonymisation required by Research Ethics Committees. However neuroimaging raises particular concerns regarding image identifiability. While personally identifying metadata are easily removed, three-dimensional reconstructions of the head are potentially recognisable from photographic databases of known individuals, including by automatic facial identification techniques (Kulynych, 2007). Collaborative neuroimaging projects therefore tend to stress provision of variable levels of access, for example PsyGrid (Ainsworth et al., 2007) and SINAPSE (Rodriguez et al, 2008) limit access to approved collaborators using role-based approaches. Access limitations are characteristic of medical domains and according to Lowrance (2006) "open access" may refer to data that is open to *application* for access. Determining which applications are legitimate may involve various considerations including confirmation of professional competence, and screening of the scientific merit of proposed collaborations. One of the challenges for medical e-infrastructure is to manage the range of access rights needed; Lowrance (2006) identifies confidentiality and anonymisation as one of the "issue clusters" most in need of attention for data sharing in medical research.

2. Neuroimaging Group: Curation Drivers and Context

2.1 Introduction

Digital curation features in the Neuroimaging Group's research agenda because of the potential value that can be derived from its datasets; and because funding bodies expect datasets to be retained beyond the funded period of a project (for example up to 20 years in the case of the MRC) and made available for re-use. Research data acquired in longitudinal imaging studies are continually revisited. The rewards of improved data description are potentially very high; re-using data with novel multivariate analysis techniques would enable studies that would otherwise be uneconomic. For example this could enable follow-up studies to analyse data collected from individuals who may have participated in various studies over many years. The Group is also internationally recognised for its work in developing leading edge analytic techniques. More re-use value could be realised by providing these as web services, to enable remote analysis by external researchers. Steps in this direction have been taken in the NeuroGrid e-science project (Neurogrid, 2007).

To explore data reusability, the initial study phase was framed as a set of interviews aimed at documenting data curation issues and opportunities. A more 'immersive' phase followed, central to which was a self-assessment of risks to data, which we return to in Chapter 3.

2.2 Research Roles and Stakeholders

Research in the Neuroimaging Group has been funded by the Medical Research Council, the European Commission, the Chief Scientist Office in Scotland, the Wellcome Trust and a number of smaller charitable trusts; the Health Foundation, the Dr Mortimer and Theresa Sackler Foundation, the Shirley Foundation and the Stanley Medical Research Institute (SMRI).

Neuroimaging Group is one of four groups in the Division of Psychiatry headed by Professor Eve C Johnstone. Digital data curation is partly the systems administrator role, yet much of the work to acquire, manage and add value to research data is of course embedded in the work of the Group's researchers. It is a multi-disciplinary group comprising (at time of writing) 24 people. There are four broad research roles: -

1) *Clinicians*: psychiatrists who combine academic research with NHS practice are invariably the Principal Investigators on projects and custodians of the data acquired. They are responsible for identifying hypotheses, designing studies, obtaining ethical approval, ensuring compliance with ethics committee and statutory requirements, identifying subjects and coordinating their recruitment, coordinating visits for scanning and clinical interviews, collating and managing the non-imaging data acquired, assessing the results of analysis and producing publications of these.
2) *Psychologists and research nurse*: who design and carry out cognitive tests on research subjects, provide the data acquired from these tests to clinicians; participate in its analysis, and in publication of results.
3) *Imaging researchers*: who quality-assess acquired image data, process and analyse it, perform statistical analysis of it and associated data provided by clinicians, visualise the results and co-author their publication.

4) *Support roles*: including the Systems Administrator who coordinates the acquisition of imaging data with hospital-based radiographers and medical physicists, manages and archives imaging data and carries out curation activities identified later in this chapter, plus Administrative and Secretarial staff who assist in recruiting research subjects and scheduling interviews and scanning visits.

These roles are highly inter-dependent; while only those with psychiatric, nursing or psychologist positions have direct contact with subjects, recruitment relies on administrative support. And while clinicians in lecturing posts tend to be involved in all study stages, imaging researchers and psychologists collaborate closely on statistical analysis and publication. This reflects the wide disciplinary backgrounds of researchers, and the growing value of novel imaging techniques to clinical research.

The Group works with various stakeholders within the institution, notably:-

- The *Centre for Clinical Brain Sciences* is a 'virtual' research centre comprising the Department of Psychiatry, plus the *Department of Clinical Neuroscience*; a partner in the Neurogrid project (see chapter 1) and location of the SFC (Scottish Funding Council) *Brain Imaging and Research Centre* (BIRC). Housed at Edinburgh's Western General Hospital, this provides MRI scanning as a service to Neuroimaging Group and other research and clinical groups, as well as collaborating on research into psychiatric disorders.

- The Psychiatric Genetics Group in the Division of Psychiatry, and collaborators in the *Department of Clinical Genetics* provide genetic data obtained from blood samples, also taken at the Western General Hospital. This data is provided as alphanumeric strings; DNA sequences are derived from samples, and small fragments of these - single nucleotide polymorphisms or SNP (pronounced 'snip') - representing genetic variations that are suspected of being correlated with psychiatric disorders, are provided to the clinical researchers.

- In the School of Informatics the *Neuroscience Doctoral Training Centre* (DTC) provides a programme of interdisciplinary PhD research in neuroinformatics and neuroscience, and has been a source of research students to the Group.

- The Edinburgh Compute and Data Facility (ECDF) provide a high-performance cluster of servers and storage used for parallel processing. This has been deployed recently to perform analyses that would otherwise be impractical, for example processing scans from a schizophrenia study in 28 hours, as opposed to an estimated 469 days with serial processing on machines normally available.

2.3 Studies and Datasets

Aside from these institutional stakeholders the main group with an interest in the outcomes of neuroimaging research are psychiatric patients and their families; some major psychiatric disorders have a genetic component that means family members are categorised as 'high risk'. This partly explains the Group's large and in some cases unique datasets and their high research value; Scotland's population has a relatively high propensity to schizophrenia and relatively low recent immigration. Stable family networks have also helped Neuroimaging Group to recruit to longitudinal cohort studies of schizophrenia and other disorders. The Group pioneered the use of imaging to demonstrate physical deficits in the brains of people with schizophrenia, and has datasets spanning more than 10-years.

Their value stems partly from their relative size and time span, since they illustrate disease progression in individuals, and partly from the combination of clinical/ neuropsychiatric, behavioural and genetic data that allow explanatory factors in disease progression to be identified. Neuropsychiatric, demographic, cognitive and genetic data are acquired and related to image data from the same individuals. This data is recorded from a variety of sources; clinical records and questionnaire-based instruments for assessing a person's mental health (neuropsychiatric data), their physical and social characteristics including family history of psychiatric ailments (demographic data), their performance in IQ and other psychological tests (cognitive data) and from blood samples (genetic data). Major projects have investigated schizophrenia, bipolar disorder, autism, and learning disabilities, and include the following in addition to the Neurogrid and Neuropsygrid projects outlined in Chapter 1.

Edinburgh High Risk Study (EHRS): Begun in 1994, this extensive longitudinal study of schizophrenia has tracked more than 200 young people who are and are not at enhanced genetic risk of schizophrenia. Individuals have been scanned at 18-24 month intervals for periods up to 10 years, providing around 250 fMRI scans, 500 sMRI scans and associated demographic, clinical and behavioural data.

The Edinburgh Study of Co-Morbid Learning Disability: Also funded by the MRC, this study has involved around 300 young people who are and are not at enhanced risk of psychiatric disorder (schizophrenia, autistic spectrum disorder) as they have IQs of 60-80. The study also concerns genetic and neuropsychological aspects.

Brain Function in Relatives of People with Bipolar Disorder: Funded by the Sackler Institute for Psychobiology and the Health Foundation, this relates structural and functional scanning to genetic, cognitive and life-history data from families who have and do not have a relative with bipolar disorder, to identify genetic factors in its development. This multi-centre study also involves Glasgow, Aberdeen and Dundee Universities.

Calibrain: Funded by the Scottish Government's Chief Scientist Office, the project aims to facilitate large clinical studies, common in other areas of medical research, by refining data analysis methods for multi-centre structural and functional MRI (s&fMRI), to address scanner differences that are currently an obstacle to data integration.

2.4 Drivers for Curation: Deeper Analysis and Wider Datasets

A key element of digital curation is 'adding value' to data. Data is expensive to acquire (typically £4-500 per scan) and of potential future research and clinical value, although as research data it is not storable in clinical records. In Neuroimaging Group data curation is being driven by steps to add value; these steps potentially enable larger datasets to be built and described, mainly by enabling existing datasets to be joined up and/or re-analysed. Most of this chapter is concerned with outlining these steps and potential consequences for the local group's data curation and storage, as part of the backdrop for considering local risks to reusability.

The Group is already sharing data with other research groups through the MRC funded e-science projects outlined in Chapter 1; Neurogrid and NeuroPsyGrid. These projects aim to give researchers in the psychosis domain access to datasets that integrate image data from various sources, along with their associated neuropsychiatric, demographic and genetic data. Participation in multi-centre studies enables greater statistical power and reliability of research findings by providing larger study populations than single centres could feasibly recruit. As Chapter 1 identified, e-science projects build the infrastructure for this and add to collaborating partners' data curation capabilities. So Neuroimaging Group has for several years pursued a strategy of building digital curation

and infrastructure through these collaborative efforts to develop remote analysis capabilities, metadata and ontologies.

Locally, meanwhile, a range of disciplinary factors makes *data storage* an ongoing concern for the Group; at the time of the study their 9Tb approx of image data had grown by 29% in 6 months. This partly reflects recent growth in the number of active researchers in the group. It also reflects trends from recent innovations. Figure 1.2 and the descriptions following it illustrate how these augment each stage of a study, while also presenting challenges for curation and preservation, especially in terms of storage and information on provenance and context.

New data sources and innovation in analytic techniques add value to existing datasets by providing new opportunities to correlate progressive changes in subjects' brains with other data. This counteracts the tendency for innovation in scanner technology to *reduce* the value of older and lower resolution image data, and means that relatively little 'old' image data is considered of low enough value to be archived offline. Image data is only rarely disposed of; normally only when quality assurance finds it to be inadequate for the research purpose. The norm is to retain all primary data; brain images are stored and retained in 'raw' and processed form, but derived data from statistical analysis is normally only retained until it has been published, on the principle it can be recreated from the parameters used.

Relevant innovations include:

a) Adopting new scanning methods and data sources

Functional imaging experiments provide an increasing proportion of the scan data; with each subject scanned providing 500-600Mb this increases storage requirements by a factor of 10-20 compared with structural scans, depending on the scanner sequence and the tasks subjects are asked to perform. *Genetic data* has also been increasingly used; for example at the end of the Edinburgh High Risk study new funding was obtained to gather genetic information from the young people involved. This has added new value to this dataset, providing new findings on genetic correlations with psychosis (Hall et al, 2006). The Group has also been able to take advantage of higher MRI *scanner resolution*; the field strength of scanners available to the Group has increased fourfold (0.75T to 1.5T and recently to 3T) in the last decade, each increase multiplying the image resolution and therefore the data storage requirements.

b) Anonymising images for privacy

Compliance with data protection and ethical approval requirements entails routinely removing identifying metadata such as names from scans, as soon as their identity has been checked. However reconstructed scans show recognisable physical features that could also identify an individual; and the challenge is to irreversibly remove them before enabling any access by collaborating researchers. In the Neurogrid project the Edinburgh group has developed 'skill stripping' software that partially automates this, removing the ears and face (Ure at al, 2007). Some curation issues arise with this; how to track the *provenance* so as to document what image processing was carried out when and by whom, and how to ensure the processing is not reversible. Also storage requirements are increased since the original and anonymised scans both require local storage.

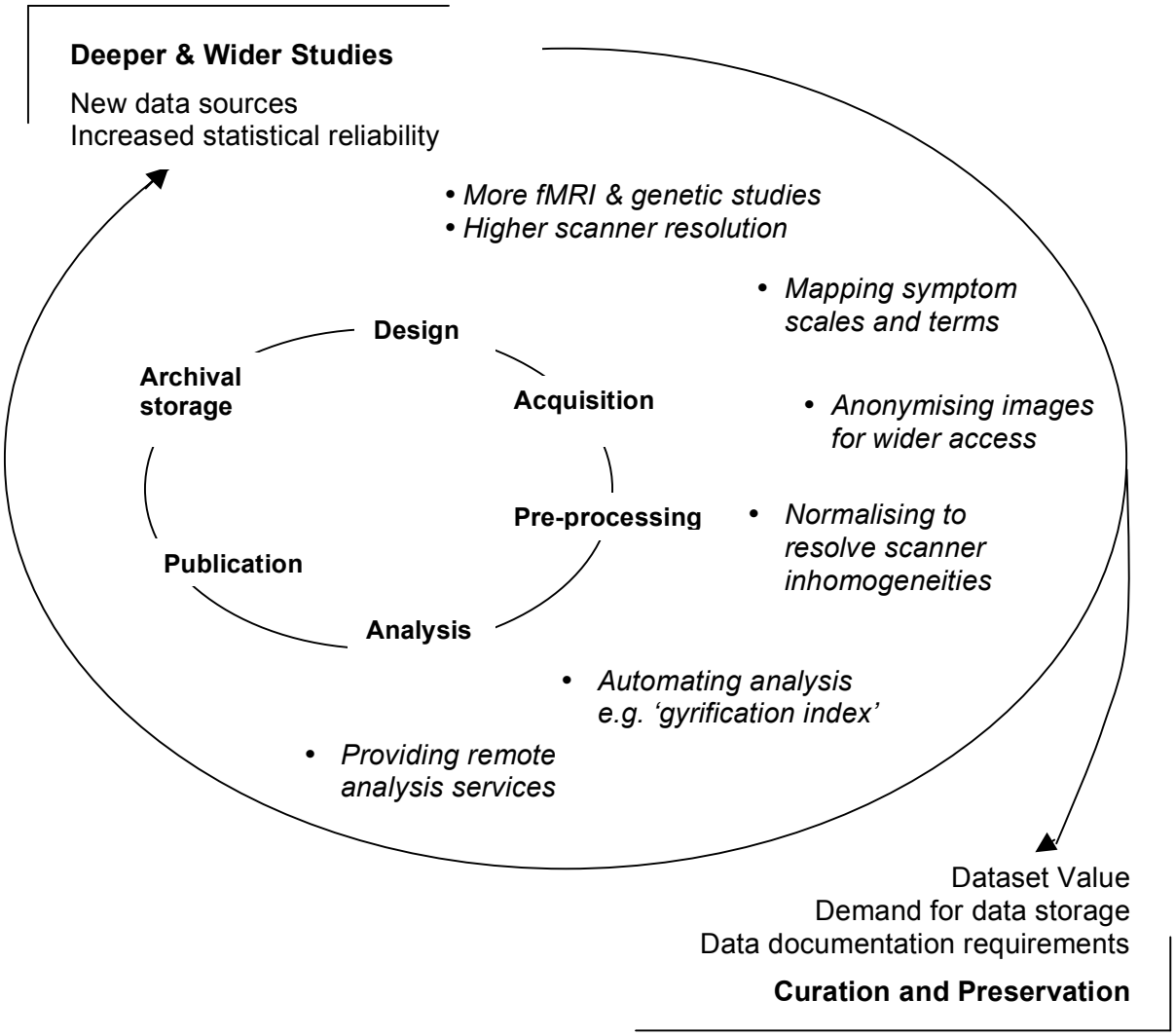

Figure 2.1 Innovation in the research cycle drives curation and preservation needs

c) Mapping symptom scales and terms to a common ontology

The collaborative ontology developments in Neurogrid, NeuroPsyGrid and BIRN are seen as vital to multi-centre studies. As mentioned in Chapter 1, design and analysis in these calls for integration of data collected using different instruments, and requires *shared terminology* for describing the datasets. One aspect of this collaboration is to address changes and cross-site differences in the measures of psychotic symptoms (Kola et al 2008). For example, the PANSS (Positive and Negative Syndrome Scale) and PSE (Present State Examination) assessment scales use terms with similar semantics, suggesting they could be at least partially mapped in an ontology.

While both PANNS and PSE are based on questionnaire responses they use scales with different structures; PSE is a three-level categorisation of symptoms, PANNS a seven-level numeric scale. PSE provides a wider variety of terms for the resulting diagnosis. Clinicians may be inclined to use the scales differently for research and clinical purposes, leading to systematic differences between centres – both within and between scales. The presence of both scales in the Edinburgh High Risk Study data offers opportunities to detect and resolve such differences.

Progress to identify and resolve inter-centre differences depends on information about how symptom data were derived from original clinical interview instruments; and in any subsequent merging of scales the original measures may need to be preserved. These represent new requirements for *provenance information*. Also, integration of different scales helps to maintain the value of older datasets and hence the need for *online storage*, since retrospective analysis of data acquired using different scales at different periods becomes a more realistic possibility.

d) Normalising for scanner inhomogeneities

Research in the multi-centre Calibrain project is using structural and functional imaging and phantom data acquired using three MRI scanners from different centres in Scotland. Scanner noise is being assessed and addressed on the hypothesis that smoothness equalisation will enhance intra-scanner and inter-scanner agreement. The Neurogrid collaboration also addresses scanner normalisation; here the scan data from multiple scanners is being pooled so as to measure scanner variability relative to a common template. That is, by 'warping' the images to make them look virtually identical, statistical data on the residual variability between the images can be analysed, and that due to variations in scanner contrast modelled (Geddes, 2006). The aim is enhanced pre-processing to provide for the seamless integration of data from different scanners, with consequent demands to store and manage *larger datasets* and new requirements to integrate *provenance metadata* on the various source scanners and the processing steps taken to normalise the data.

e) Automating analysis e.g. 'gyrification index'

Neuroimaging analysis has tended to be a highly labour-intensive process, and increasing focus has been placed on automating aspects of it (Poliakov et al, 2007). Automation enables larger numbers of brain images to be used in analysis, with the attendant benefits of progressively greater accuracy and efficiency; coupled with *higher demand on storage* and computational resources, and again new demands to document the steps taken in processing as *provenance information* (*ibid.*).

The analysis of 'gyrification' is one example developed by Neuroimaging Group. Gyri are the ridges visible in brains and *sulci* the furrows between them. The patterning of these across the cortex, measured by the 'Gyrification Index' is of interest because cortical folding abnormalities in various brain regions are indicative of psychiatric disorders. Gyrification Index (GI) is an established measure of cortical folding and is assessed as the ratio of traced lengths between the entire cortical surface and the superficially exposed cortical surface (Moorhead et al 2006). The Group has led development of an automated technique (A-GI) that substantially reduces the time costs involved in this analysis. Previously this has relied on researchers tracing out the patterns by hand from sMRI images, so A-GI which is implemented as an add-on to the SPM toolkit, makes it economic to analyse many more scans than previously.

f) Providing remote analysis services

Work in NeuroGrid has developed the principle of presenting algorithms as Grid services (Geddes, 2006). The service-oriented approach used in this has enabled remote analysis to be carried out by Oxford University collaborators on a combined dataset held in federated data storage. The approach claims several benefits; firstly it is seen as 'lowering the barriers to entry', i.e. novel analysis techniques can be provided to collaborating neuroimaging centres as and when they require it and with less duplication of effort. And secondly analysis functions can be made available more efficiently this

way; rather than the expensive investment needed to develop a new all-embracing toolkit, researchers can share their algorithms with collaborators using an 'algorithm-wrapping portal'. The principle is that to make a new software available a researcher would upload their script to the portal, and specify the parameters required to use it. The necessary java code and XML service specifications are then automatically generated. The script could then be registered with an index service to make it discoverable (Neurogrid, 2005). So remote analysis also creates new demands for *tracking provenance*, in terms of identifying which datasets have been processed by which tools and when. It is also likely to increase local demand for *storage of derived data*, since remote analysis expands the range of datasets available to each participating centre.

The innovations highlighted here evidently contribute to data curation to the extent that they add to datasets' research value. In doing so they also imply new requirements for data storage and documentation, especially since much locally generated datasets are likely to remain locally stored, whether or not the alternatives prove sustainable. So far this report has not considered how data curation is handled at the lab level, which the next chapter now considers.

3. Risks to Data Reusability

3.1 Introduction

The Neuroimaging Group was already experienced with risk analysis as a means of prioritising infrastructure development needs; as their systems administrator remarked "one of the challenges for anyone coming from a digital curation point of view…a lot of it is about risk management rather than delivery of x or delivery of y" (int. 7). Risk assessment was seen as a means to involve clinical researchers in setting priorities. In the SCARP study the approach used to assess risks to dataset (re)usability was the *Digital Repository Audit Method Based on Risk Assessment*, or DRAMBORA (DCC/DPE, 2007). Figure 3.1 outlines the main stages of the approach, which the rest of this chapter details.

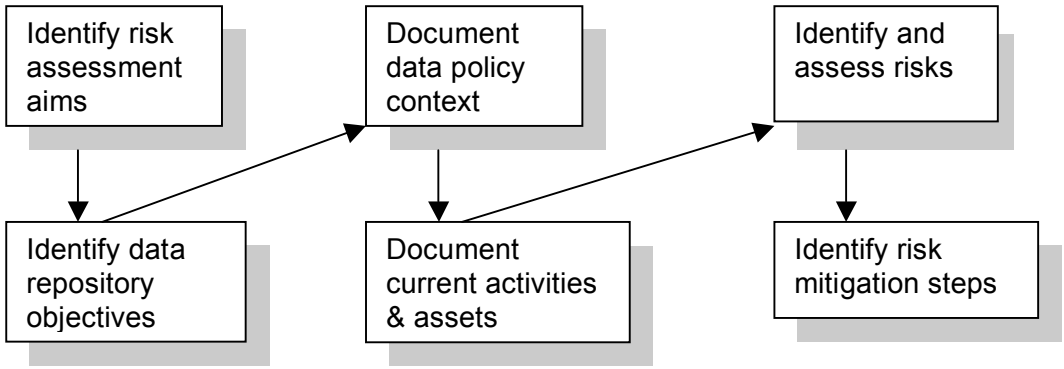

Figure 3.1 Risk assessment steps

3.1 Risk Assessment Aims and Repository Objectives

The aims of the risk assessment stem from Professor Stephen Lawrie's interest in curation as a means to unlock the large 'hidden' value in the datasets by developing a framework for describing them, to enable the data to be mined. The risk assessment was seen as a step towards achieving the Group's aim of developing an integrated MRI and associated data facility for that purpose.

The DRAMBORA approach: "begins with the self-auditing organisation specifying its mandate. From this starting point, a hierarchy of fundamental objectives and activities is identified" (op.cit., p.38). Neuroimaging Group is an active data creator and does not have an institutional or externally defined mandate to act as a *repository*, except in the sense that the University and funding bodies give them custodial responsibility for research data. So as a basis for the self-assessment a statement of repository objectives was drafted as follows:

"Repository objectives are to contribute to the Neuroimaging Group research aims by:

- *Securely managing the Group's digital data assets to support the activities required to meet those research aims*
- *Preserving and adding to the digital data assets' long-term value as a resource for retrospective and secondary analysis by the Group, its research partners, and the wider research community."*

3.2 The Data Policy Context

The first two chapters in this report have documented the data policy context in terms of the Group's research objectives and activities, and how these relate to factors driving digital curation in the neuroimaging community. The regulatory need to protect confidentiality of personal data has already been mentioned. Neuroimaging research data largely relates to individuals who are psychiatric patients, members of their families and others recruited as healthy controls; and is deemed sensitive personal data under the Data Protection Act (DPA). Compliance with this is of prime concern to Principal Investigators. While the University of Edinburgh's governance framework and administrative support is centralised in its Records Management Office, responsibility for ensuring compliance is largely devolved to Heads of School. In practice it falls to PIs, as custodians of the data gathered on research subjects to ensure that all data handling practices comply with the DPA, since they are also given responsibility for that through the NHS Research Ethics Committee approval process.

The Group has a data management policy drawn up several years ago by the previous systems administrator, and covering data management and curation tasks, good practice guidance on where and how to store raw, pre-processed, and processed image data, plus the linking of this to associated clinical, cognitive and demographic data and processing scripts. Guidance is given on directory organisation, file naming and on secure data storage. The policy is considered at an early stage of development, because resource constraints are seen as stifling the development of systems that would provide researchers with enough of a 'carrot' to police their own behaviour.

Funding body policies and contractual terms also govern curation and preservation; the Group is committed to documenting and maintaining datasets for up to 10 years post-funding, which for longitudinal studies of 5 to 10 years in effect means up to 19 years from acquisition (as this typically begins one year into a project). The MRC and Wellcome Trust policy statements on data sharing and access are regarded as positive developments and resonate with the Group's research objectives; their involvement in multi-centre grid projects is designed to facilitate data integration and sharing, yet a database infrastructure to maintain and add value to *locally held* data remains in the early stages of planning; a situation that is seen as typical of centres in the UK and attributed to a historic lack of infrastructure funding for the neuroimaging field.

3.3 Current Curation Activities

Interest in developing informatics-related means of adding value to existing datasets is widely shared in the neuroimaging community. However the UK has no established data archiving service in this domain to parallel U.S. initiatives such as the BIRN Human Imaging Database. It and other international efforts gave examples of data repository functions that are relevant here, but a more comprehensive guideline is the OAIS reference model (CCSDS, 2002) of the functions an 'open archival service' should provide. The case study therefore used it for comparison with current activities identified in interviews with Neuroimaging Group members, as a basis for considering how to further develop curation capabilities.

The main curation activities, assets and risks identified are described below for each of seven broad functions derived from the OAIS functional model:-

1) Data Acquisition & Ingest
2) Data/ Metadata Management

3) Archival Storage
4) Preservation Planning
5) Data Access
6) Systems Administration & Services
7) Management Support

3.4 Data Acquisition & Ingest

- *Provide appropriate storage capability to receive imaging and associated data and metadata relating to research projects and their subjects, and control access to the data received.*
- *Provide appropriate quality assurance to validate the integrity, authenticity and usability of data received.*
- *Transform received data to formats appropriate to research application and archival storage.*

The term 'ingest' is used here to refer to the processes from acquisition by researchers in the Group to its local storage. Currently the Group has no specific arrangement to submit datasets to a data repository or archive, so has no experience of archival ingest processes through which datasets are submitted by data creators and prepared for discovery by other data users.

The process differs for imaging and demographic (genetic and neuropsychiatric or cognitive) data. *Demographic and clinical* data are acquired by clinical researchers in questionnaire form and transcribed into Microsoft Excel spreadsheets or Access databases, and provided to the PI who curates this data personally; on secure server space or on individual hard drives. Clinical interviews are also *videotaped* and held securely in a video library. Sequenced genetic data is provided directly from the Western General Hospital to the PI/custodian in charge of a project, and held in numeric form in spreadsheets.

Imaging data is physically acquired at research scanner sites in hospitals. Images originate in the DICOM format, and are sent by hospital medical physicists across the academic network. These are received into an 'incoming' area of the server reserved for this purpose, transferred into the corresponding project directory, and converted to other formats by image analysts, primarily Analyze[4] and NIFTI[5].

Quality assurance was being upgraded during the case study and comprises three main checks on the data acquired; its transfer, study, and scan integrity, *Transfer* integrity checks involve automatic notification that files sent from hospital-based medical physicists are received. This overlaps with *study integrity* checks, i.e. that the files are as expected, for example that the number of folders corresponds with the number of scanning sequences run, that brains have been fully scanned, and that the correct people have been scanned. A new format for 'master files' has been introduced to

[4] Analyze biomedical image format is published by the US medical education centre Mayo Clinic and available at: http://www.mayo.edu/bir/Software/Analyze/AnalyzeTechInfo.html
[5] The format is an acronym for Neuroimaging Informatics Technology Initiative (NIfTI), sponsored by the US National Institute of Mental Health and National Institute of Neurological Disorders and Stroke and available at: http://nifti.nimh.nih.gov/nifti-1/

standardise the details that the lead researcher records about a scanning session and the subjects participating. *Scan integrity* checks are carried out using image analysis scripts to automatically check the image intensity levels are within acceptable limits, and that certain known artifacts are not present. When unacceptable images are relatively few in number, the errors can be corrected using averaging techniques. Aside from the above QA scripts and procedures, the main *digital assets* associated with acquisition are scanner sequences. The main *risks* identified were that: -

- *Image and other files sent do not match those received*; this risk is being addressed through ongoing collaboration with the Brain Imaging Research Centre to implement a file hashing script to enable checks at both ends of transmission.
- *Archived data cannot be traced to information about receipt*, this may arise when the relationship between the files acquired in a session is lost- this link is only implicitly identifiable from the files' original location in the file structure, or in a block of server storage volume.
- *Images contain errors that are undetected until later stages of image analysis*; this actually occurred during the case study, although the image artifacts were detected early in image analysis and QA scripts mentioned earlier were quickly developed and implemented to address this.

3.5 Data/ Metadata Management

- *Provide services and functions for populating, maintaining, and accessing both descriptive information, which identifies and documents data held, and administrative data used to manage the repository.*
- *Maintain and apply appropriate data and metadata schemas to ensure the integrity, authenticity and usability of data held.*
- *Perform database updates and queries to generate result sets, and produce reports from these result sets.*

Datasets are documented in various ways. Study protocols are documented in *ethics committee approval applications*, and describe the experimental design, the proposed subjects and the data on them to be used and gathered; *'master files'* are spreadsheets used to record details of the recruited subjects who take part in scanning sessions; *experimental protocols* document the processing steps taken, and *readme files* and *script comments* are more ad-hoc notes of processing results and file locations.

Some standardisation of master files has been introduced, although no metadata schema is employed, nor is database management software used to record metadata or to retrieve image data/metadata attributes consistently across datasets, although the imaging files held number in the millions. Metadata is embedded in the DICOM image file headers data, including scanner attributes and individual subject details (removed before processing). Other data relating to clinical, cognitive or genetic data is also recorded in spreadsheets and SPSS files.

Clinical researchers (principal investigators) have historically managed this data largely on individual hard disks, although in one recent large study a database was developed in secure server space. The separation of image and demographic data reflects several disciplinary and historical factors; demographic data is managed on a per study basis by clinical PI's who have personal indemnity insurance to mitigate risks of its loss; also file sizes tend to be an order of magnitude smaller for this data than for images. Systems administration is largely focused on image files for disciplinary and pragmatic reasons; since image analysts have informatics-related backgrounds they have tended to carry

out systems administration and infrastructure development 'on top of' other research work, and have focused on managing the ever-scarce storage resources.

Scripts used to process images can also be considered a form of data; they are normally saved by Group members who perform the analysis, and may be re-used when there is a need to re-run the same analysis. The derived data from image analysis is also stored, although intermediate steps are not. Scripts are documented in the form of code comments and 'readme' files, as are coding changes. A software versioning system has recently been implemented to improve script change-tracking.

The digital *assets* identified here are the dataset contents, their organisation, the scripts used to process them, documentation of the data and the research protocols, readme files, and software licences. In terms of the funding used in their creation alone, the datasets have a financial value of millions of pounds.

The main *risks* identified with data and metadata management were: -

- *Loss of integrity of information*, i.e. the possibility that datasets cannot be retrieved because the data from their constituent parts cannot be reliably or accurately inter-related, or linked to study details.
- *Context information lost or unrecorded*: the possibility that datasets cannot be used because it is not clear who collected what data, for what purpose, or how it has been described.
- *Provenance information lost or unrecorded*: the possibility that datasets cannot be used because it is not clear when changes were made, by whom or with what processing steps, and therefore whether the available version meets the purpose. The possibility was judged to be higher for demographic data than for image data. Risks also arise from the variety of software packages, file formats and versions of both that are used in neuroimaging, and the strong inter-dependencies of study results and the software and hardware used for analysis. Replicability of analyses therefore depends on documenting technical metadata of the computational environment used to render them (or *representation information*), as well as of the analysis process.
- *Identifier providing referential integrity is compromised*: the possibility that datasets cannot be retrieved because a unique identifier is lost or can no longer be used, meaning that data from their constituent parts cannot be reliably or accurately inter-related.

Each of these was considered more of a barrier to the aim of enabling retrospective analysis by others, and to datasets being reused by newcomers to the group or independent researchers; rather than as barriers to current members' use and re-use.

3.6 Archival Storage

- *Store, maintain and retrieve archived digital objects efficiently and effectively.*
- *Manage a storage hierarchy, refresh storage media on which archive holdings are stored, and perform error checking routinely and as required.*
- *Provide disaster recovery capabilities.*

There is a steady increase in demand on the systems administrator's time to manage online storage resources and archival tasks, which consist mainly of ensuring data is securely backed-up and easily retrievable. This demands more scarce time resources than does acquiring new image data, since scans are obtained as a service from clinical research facilities, in effect meaning that older datasets cannot be moved to cheaper offline storage (e.g. LTO tape) as quickly as new data is acquired. This makes it

essential to be able to retrieve backup sets easily and reliably in the event that storage volumes are damaged. A backup storage system has been developed in-house to address this need, together with procedures to mitigate risks of media failure and allow data to be quickly and easily recovered from tape.

Storage media have recently been refreshed, although there is no policy in place to do this at regular intervals. Disaster recovery capabilities are seen as basic but reasonably effective; tapes are stored in several different locations, although not currently off-site. The backup system is the main digital asset related to archiving, aside from the backup datasets themselves, and associated documentation. The system mitigates against risks that might be expected for the archival function, although two were identified:-

- *Extent of what is within the archival object is unclear*: a possibility that datasets cannot be used effectively because it is not clear which archived files contain which items of data. Currently the system records filenames and can recover these, but relies on the availability and knowledge of the data custodian to relate filenames to a study and the other data associated with it.
- *Destruction or non-availability of repository site:* a possibility that datasets are not usable as they have been damaged through physical destruction of the repository, or are temporarily unavailable at a critical time, for example through fire or flood.

3.6 Preservation Planning & Action

Collaborative projects that use federated storage and provide remote analysis services

- *Evaluate the contents of the repository, periodically recommend updates to migrate current holdings, and develop detailed migration plans, software prototypes and test plans to enable implementation of Systems Administration's migration goals.*
- *Develop recommendations for repository standards and policies.*
- *Monitor changes in the technology environment and in the designated user community's service requirements and knowledge base.*

carry out some of these functions during the projects' lifetime. The risk is more to locally held datasets, which are the majority, and was identified as: -

- *Inability to evaluate the effectiveness of preservation*: the possibility that datasets cannot be retrieved or used because of a failure of preservation actions that would be avoided if these were assessed on explicit criteria and measures.

3.7 Data Access & Dissemination

Provide services and functions to: -

- *Communicate with the designated user community to receive requests.*
- *Apply controls to limit access to specially protected information.*
- *Coordinate the execution of requests to successful completion, generate responses (digital objects, result sets, reports) and deliver the responses to users.*

The term 'designated community' is OAIS terminology for those people who *should* be able to understand the contents of an archive, assuming that documentation of the data's origination, provenance and context is made available to independent researchers with the knowledge to be able to repeat a neuroimaging analysis. At present access is limited to members of the Neuroimaging Group and collaborators, but with a view to extending that to the wider psychiatric research community; as currently done by sharing data through for example the Neurogrid project.

Locally, image data may be found by navigating the server file structure. Directory and files are labelled according to a naming convention, and are mostly organised according to the studies they have been collected for. Study directories are organised according to processing stages. Users also have 'data' directories for storing images and 'home' directories for documents, spreadsheets and scripts relating to work-in-progress on their analysis. The file structures and naming convention are the main digital asset here, and the main risk identified was:-

- *Finding/searching tools are not sufficiently effective or usable*: the possibility that relevant datasets do not get used because they cannot be found.

The risk is that current file searching tools and subject identification methods provide very limited abilities to meet the aim of retrospectively analysing subjects *across* datasets. For current studies where analyses are on single datasets, the participating group members did not see the file structure as a major barrier to finding image files. Nor is finding demographic data seen as a major difficulty. However this was recognised as the most 'at risk' since loss of a file containing demographics for a study would require extensive effort to reproduce that data from paper originals. Since the demographic and clinically related data is vital for analysis, loss of availability would reduce value of all the corresponding scan data.

Since neuropsychological test data relates to standard questionnaire scales and commonly used abbreviations for these, clinical researchers did not see it as a problem to interpret data field headers in each other's spreadsheets. There is also much tacit knowledge shared about who has what clinical data, and for what purposes it has been used; so neither was it seen as a major risk for group members to understand the context of datasets they had not worked on themselves. As discussed in Chapter 4, local data sharing is an aspect of the strong collaborative culture and the fact that few of the lead researchers responsible for completed studies have departed the Group. Data dissemination *outside* the group is also discussed in Chapter 4.

3.8 Systems Administration & Services

- *Solicit and negotiate submission agreements with data creators, auditing submissions to ensure that they meet repository standards.*
- *Maintain configuration management of system hardware and software, and provide system engineering functions to monitor and improve operations, and to inventory, report on, and migrate/update the repository contents.*
- *Establish and maintain standards and policies and provide user support.*

Formal submission agreements are not currently relevant, in the absence of a data archive, but would become so if one were to be established. The deposit processes of other data archives such as BIRN Human Imaging Database indicate the scope of what would be required. Currently the focus of systems administration is on supporting and maintaining the Group's software and hardware. Much of this infrastructure has been replenished in the past two years. A more recent addition to that role is to support the use of high-performance parallel computing to enable new forms of analysis that would not be practical with conventional serial processing, e.g. connectivity analyses.

The main risks identified under this heading were:-

- *Obsolescence of hardware or software*: the possibility that datasets cannot be retrieved or used because it is no longer feasible to read the data format using available software, or to replace some hardware device needed to read it.
- *Media degradation or obsolescence*: the possibility that datasets cannot be retrieved, used or understood because the media they are stored on cannot be reliably used with available hardware or software.
- *Loss of encryption key*: the possibility that datasets that are needed cannot be accessed because an encryption key is lost or decryption fails.

3.9 Management Support

- *Provide and periodically review the mandate of the repository.*
- *Maintain a commitment to digital preservation services and fitness of the organisation to provide it.*
- *Ensure that the operation of repository services complies with legal & regulatory requirements.*

The 'mandate' for a repository normally refers to a formalised requirement for a library or archival function of an institution to provide a repository service to a designated community of users. Although this is not so relevant to the present context, the research funders, University, and Neuroimaging Group's management provide the mandate for the stewardship and archiving of the Group's data, in terms of the objectives identified earlier in this section and through supervision of projects, staff roles and their performance, including compliance with legal and regulatory requirements.

Foremost of the risks identified is that, while the Group is active in several key e-science infrastructure initiatives to facilitate data integration and sharing in the neuroimaging domain, the funding available to set up and maintain *local* data infrastructure has not scaled up to match the increasing volumes of data collected or the requirements for metadata and documentation. Previous sections have shown the provision of local *computing* infrastructure is high but better *information* infrastructure is needed locally to ensure the long-term usability of datasets and extract the most value from them. Three risks associated with that were identified: -

- *Preservation failure*: the possibility that, if no further action is taken to preserve them, one or more of the Group's highly valued datasets will become unusable- either completely or for the kinds of analysis sought. This could arise through a combination of lower-rated risks whose impacts would have a knock-on effect on the overall probability of failure e.g. loss of demographic data held on individual hard disk, combined with loss of key people, continued lack of a structured framework for documenting or storing/retrieving metadata, or backlog of data that needs 'cleaned' to fit into one, continued storage mgmt problems; or a fire destroying archives held on site.
- *Loss of key member(s) of staff*: a risk that several key investigators, researchers or support people leave or are absent for a long period, and as a result datasets cannot be retrieved or used effectively.
- *Data integration unmanageable*: a possibility that inconsistencies between data structures, names or values prevent cost-effective importing of datasets into a database. Since each project has demographic and clinical data stored by different custodians in different record formats inconsistencies are likely (it was not possible in the case study to quantify their extent).

3.10 Addressing Risks across the Curation Lifecycle

Monitoring is a key aspect of any risk assessment methodology. *Lifecycle* approaches to managing data are considered an aid to planning successful curation and long- term preservation (Higgins, 2008). The DCC *Curation Lifecycle* model (ibid.), illustrated in Figure 3.2, is informed by OAIS and other standards intended for a wider variety of organisations, so provided an appropriate guide to considering risk mitigation steps[6]. Table 3.1 summarises the current risks to datasets and steps being taken to address these by Neuroimaging Group, and maps these to each of the lifecycle actions. The table highlights the risks to data that current work addresses, and those parts of the lifecycle where further steps are needed. The investment in risk mitigation at the later stages of the curation lifecycle is not currently matched at earlier stages of the lifecycle where the balance of risk is highest, i.e. from *Create or receive* to *Store*, although the ontology and schemas used to describe data will eventually benefit the *Ingest* stage by providing the kinds of metadata that a domain repository would need.

While these development are scheduled to bear fruit in the next few years, applying that work to existing locally-held datasets is likely to take time beyond that. Meanwhile there is a need for 'quicker wins' to address the risks identified. These include seeking funding to support the data curation infrastructure and the further steps recommended in the next chapter.

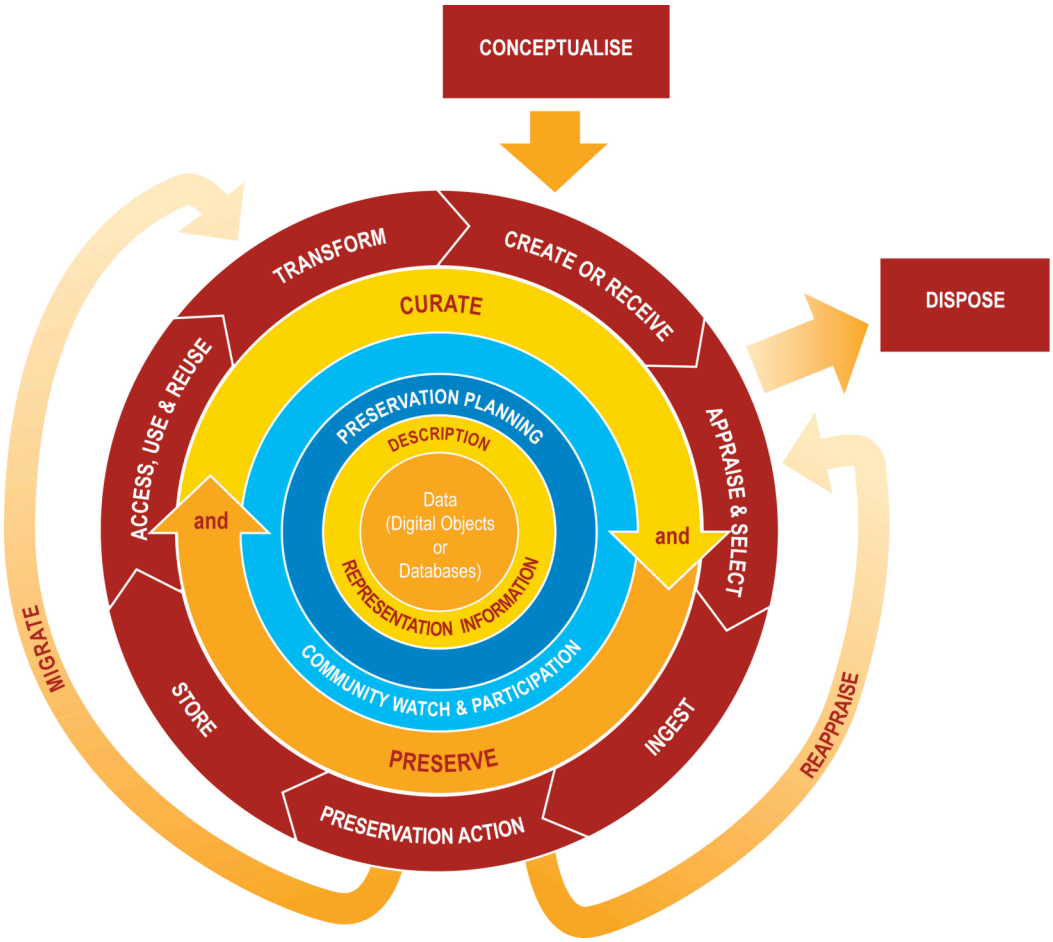

Figure 3.2 DCC Curation Lifecycle Model (Higgins, 2008)

[6] The risk assessment in this study was completed before the lifecycle model was finalised.

Table 3.1 Curation Lifecycle Risks and Mitigation Steps

Full Lifecycle Actions

Actions	Recommended Scope	Main Risks Identified	Mitigation Steps in Progress (*= study recommendations)
Description and Representation Information	Assign administrative, descriptive, technical, structural and preservation metadata, using appropriate standards, to ensure adequate description and control over the long term. Collect and assign Representation Information required to understand and render both the digital material and the associated metadata.	• Loss of integrity of information, i.e. links between dataset elements & study docs.	• Ontology to describe data • Standard 'master file' documentation. • Data documentation system to link and describe study files and representation info. *
Preservation Planning	Plan for preservation throughout the lifecycle of digital material	• Preservation failure • Loss of key member(s) of staff • Funding bodies may misperceive the level of infrastructure for local curation	• Seeking funding for preservation activity • Develop data policy to include dataset appraisal & migration *
Community Watch and Participation	Maintain a watch on appropriate community activities and participate in the development of shared standards, tools and suitable software.	• None	• Active participation in e-science consortia, multi-centre projects and professional networks
Curate and Preserve	Be aware of, and undertake actions to promote curation and preservation throughout the lifecycle.	• Data integration unmanageable	• Establish extent of data cleaning needs, using sampling approach*

Table 3.1 Curation Lifecycle Risks and Mitigation Steps (continued)

Sequential Actions			
Conceptualise	Conceive and plan the creation of digital material, including capture method and storage options.	• Identifier providing referential integrity is compromised	• Data integration to enable retrospective & multicentre studies • Link subjects across studies by unique identifier *
Create or receive	Create digital material including administrative, descriptive, structural and technical metadata.	• Data integration unmanageable • Files sent do not match those received	• Ontology to map different assessment scales • QA to check integrity of files received
Appraise and Select	Evaluate digital material and select for long-term curation and preservation. Adhere to documented guidance, policies or legal requirements.	• Inability to evaluate the effectiveness of preservation	• Appraisal process/ criteria* • Data documentation sys. *
Ingest	Transfer material to an archive, repository, data centre or other custodian. Adhere to documented guidance, policies or legal requirements.	• Archived data cannot be traced to receipt • Privacy breach	• Data documentation sys. * • Anonymisation to strip images of identifying data
Preservation Action	Actions to ensure long-term preservation and retention of the authoritative nature of digital material. Ensure material remains authentic, reliable & usable while maintaining its integrity. Actions include validation, assigning preservation metadata & representation info, ensuring acceptable data structures, file formats.	• Context information lost or unrecorded • Provenance information lost or unrecorded	• Ontology to describe data • Data documentation sys. * • Policy on renewing backup technology *
Store	Store the data in a secure manner adhering to relevant standards.	• Extent of what is within archival object unclear • Destruction	• Standard master file format • Data documentation- share metadata * • Offsite storage of backups *

Table 3.1 Curation Lifecycle Risks and Mitigation Steps (continued)

Occasional Actions			
Access, Use and Re-use	Ensure that digital material is accessible to designated users and re-users on a day-to-day basis. This may be in the form of publicly available published information. Robust access controls and authentication procedures may be applicable.	• Finding/searching tools are not sufficiently effective or usable	• Ontology to describe data • Dataset sharing • Analysis automation & remote services • Data documentation- share metadata *
Transform	Create new digital material from the original, for example - by migration into a different form - by creating a subset by selection or query to create newly derived results, perhaps for publication.	• Data integration unmanageable	• Ontology to map terms • Normalisation to correct scanner inhomogeneities
Dispose	Dispose of material which has not been selected for long-term curation and preservation in accordance with documented policies, guidance or legal requirements. Typically data may be transferred to another archive, repository, data centre or other custodian. In some instances data is destroyed. The data's nature may, for legal reasons, necessitate secure destruction.	• Privacy breach	• Introduce assured deletion process *
Reappraise	Return data which fails validation procedures for further appraisal and reselection.	• None	• QA process to check integrity of image & study data
Migrate	Migrate data to a different format. This may be done to accord with the storage environment or to ensure the data's immunity from hardware or software obsolescence.	• Obsolescence of hardware or software • Media degradation or obsolescence	• Migration policy for moving datasets to new storage formats or media *

4. Supporting the Curation Lifecycle

4.1 Introduction

This chapter elaborates on risk mitigation steps to help meet Neuroimaging Group's objectives. Three main sections describe the next steps identified: -

- Further develop a *data policy* and *core metatada set* as described in section 4.2. The data policy should cover the curation lifecycle steps, e.g. to identify how datasets are archived, according to what appraisal criteria, and how often to migrate them to newer storage media and technologies. The core metadata set should identify what information and documents to link to datasets.
- Develop a *data documentation system* in incremental steps towards a curated database integrating imaging, associated clinical data, and study metadata. Section 4.3 describes the requirements at a high level.
- Take a *phased approach to development* as summarised in section 4.4 introducing more systematic documentation, systems to manage metadata and documentation, and then link this to integrated imaging and clinical data.

4.2 Developing the Data Policy

The following steps would complement the Group's Data Policy, and follow the 'sequential actions' of the lifecycle model:

Dataset Conceptualisation

Defining and implementing a *unique identifier* for data on study participants/subjects would help develop capabilities for more data driven and retrospective analyses. This would mitigate risks to the referential integrity of the datasets, and build on recent steps to standardise 'master files' recording the subjects scanned for each project.

Appraisal and Selection

Explicit criteria and procedures for appraising datasets would clarify when datasets should be moved from on-line to (cheaper) off-line storage, freeing up online storage and systems administration time. Criteria proposed by Hilder (2005) would be a useful starting point:-

1) Vital for continuity of medical research and survival of line of research
2) Important for continuity of medical research, research projects will fail if lost, or a statutory requirement
3) Not keeping will put a research project at risk, loss of replicability
4) Not keeping will be an inconvenience, but desirable to keep for history of science research
5) May never be needed - little or no effect if not kept
6) Don't need

Key steps would be to: -

a) List the datasets and the main custodian of each

b) Decide what characteristics of the datasets determine their current value, i.e. what should the criteria be, and how often does that change i.e. how often should value be re-appraised.
c) Decide what actions follow from the criteria, e.g. move dataset offline on tape
d) Apply the criteria to all datasets
e) Validate the results- i.e. check that the criteria are applied consistently and after doing so the results are desirable, or if not redo the criteria until they are.

Disposal

Disposal can entail transferring data to another archive (see Ingest below) or deletion. The Group currently deletes very little imaging data but this is likely to change as datasets age, and an important corollary of the appraisal process is that: -

- Data marked for disposal is deleted using an assured deletion process, especially where it contains personal data.
- Storage media are securely erased when they (or a machine they are built into) are earmarked for disposal.

Ingest

Domain archives are not yet established in the neuroimaging domain in the UK, although if the Group decided to transfer data to one, or to another custodian, it is party to several collaborations that might be in a position to accept it. Examples of the regulatory considerations and metadata requirements can be found from (e.g.) the BIRN Data Repository, whose *Data Contributor's Agreement*[7] and *Submission Guidelines*[8] are useful references for planning any future transfers of datasets.

Preservation action

The Group needs better capabilities in the areas of *dataset documentation*, including records of data relationships and dependencies. Rigorous documentation of the Edinburgh High Risk Study has been a key step in enabling further value to be extracted from it, for example in the Neurogrid project. More needs to be done to meet the aim of extracting further value through restrospective analysis and to ensure datasets are usable and interpretable by independent researchers in (say) 20 years time.

Specifically there is a need to implement a *core metadata set*, and at time of writing progress on this was underway. Any schema chosen at this stage is likely to change; the various multi-centre collaborative projects the Group is party to are developing data and metadata schemas that, if not yet community standards, might well acquire that status in the near future. For example, various schemas associated with BIRN are currently being harmonised with each other and the fMRIDC's (according to Marcus et al 2007). This makes an incremental or 'agile' approach appropriate; starting with a very simple schema and working to define the level of detail needed to support emerging query requirements should help the group identify those requirements and gain experience before embarking on a larger scale project.

As the core metadata set and any neuroimaging community standards are developed further for archival needs they should aim to include the types of information the OAIS standard (CCSDS, 2002) recommends that an archive should package with the content of the digital objects/datasets in its charge (though it avoids the term 'metadata'). These are:

[7] BIRN BDR *Data Contributor's Agreement* available at: - http://nbirn.net/bdr/overview_of_submissions.shtm
[8] BIRN BDR *Overview of Submissions* available at: - http://nbirn.net/bdr/overview_of_submissions.shtm

- *Representation Information*: this is the metadata used to render data in a human or machine understandable form, e.g. the structural and semantic information that software needs to detect, read and display a file format, and any dependencies the output of that software has on the computational environment – script or compiler versions, libraries, hardware architectures etc.
- *Descriptive information:* enables users to identify, find and retrieve objects, e.g. by using controlled vocabularies for classification, and providing links to subject-related resources.
- *Preservation Description Information*: is that needed to preserve digital objects and actions undertaken to preserve them, and includes:-
    - *Context information*: comprising links to information about the creation of the digital objects, administrative documents and metadata about study aims, ethics approval, access rights, intellectual property rights etc., or any other documentation needed to explain the study rationale and setting.
    - *Provenance information*: is documentation and metadata about the history of an object from its creation through subsequent processing and any preservation actions taken (e.g. migration to a new file format).
    - *Reference information*: provides the means to cite or link to a digital object, e.g. a persistent identifier or a url.
    - *Fixity information*: assures the authenticity of a digital object, e.g. the checksum calculations applied to data items.

It is important to note that *provenance* is used in the archival community in a more restricted sense than is common in neuroimaging, e.g. Mackenzie-Graham et al (2008) define provenance in neuroimaging to include some aspects of context and representation information. The last section in this chapter makes some suggestions as to what should be documented under these categories.

Storing data

Data is stored for archival purposes, and storing copies of archival tapes *offsite* would mitigate the risks of on-site fire or flood. Neuroimaging Group's innovative backup method minimizes the time costs of restoring data in the event of failure, and in other respects exceeds the current 'best practice' of storing at least two copies of all data, using at least two file formats, and in two places.

Access, Use and Reuse

The Group's policy governing access to datasets by independent researchers could be developed further, informed by the range of factors mentioned earlier (p.18-19).

Data Migration

The Neuroimaging Group already uses well-documented open file formats. The group has also recently invested in the local technology infrastructure; i.e. servers, workstations and backup technology. To safeguard the data used on these there should be a *migration strategy* defining how often to replace storage media and technologies and move backup data to the 'refreshed' media.

The tape used for backups should be refreshed with new generation (version) tape every 5 years. LTO tape is regarded as the 'safest' media for long-term storage, since it has a lifespan of 30 years when stored in ideal conditions (20 degrees C, 40% humidity).

In practice however technical obsolescence is a significant risk. The LTO specification requires that tape drives can read tapes two generations back. For example, LTO-3 drives introduced in 2005 will read LTO-1 tapes, but the new (2007) LTO-4 drives will not. Since a new generation is introduced every 3 years, the risks of obsolescence rise steeply if tape is not refreshed using the latest generation every 5 years.

Preservation of the Group's library of *clinical interview videos* would best be addressed by migrating these to 'digital master' copies and storing these on LTO tape. Current costs are approx. 30 Euros per hour of recorded video. Best practice for archiving video is to digitise it at as high resolution as possible- a practical solution for DVD quality results would be to digitise at 4Mb/second to MPEG 2 file format, maintaining a 'digital master' for each video. For secure online access and distribution, where smaller files are sufficient, MPEG 4 is a widely used current standard. When digitised 300 VHS recordings of 20 minutes duration would require 180Gb of storage if digitised at 4Mb/second to MPEG 2, and 22.5 Gb of storage for MPEG 4.

4.3 Towards a Data Documentation System

A relatively simple documentation repository would be a useful first step towards an integrated database. A starting point for that is to outline the activities involved in using a documentation system, without pre-determining a technical solution. Those activities need to take into account the interests of funding councils in making information about datasets more widely available to the research community:

> "…whatever detailed description of selection, measurements, validation, coding, programmes, sample handling and so on – including the changes over time – that *a stranger to the data set* would have to know in order to conduct a respectable independent analysis." (Lowrance, 2006, p. 10, emphasis added).

As the previous chapter discussed there are different degrees of 'stranger' and newcomers to the Neuroimaging Group are already guided through datasets and their processing by more senior researchers and their documentation. One way of expressing the need for a data documentation system is to help such 'strangers' by building on these current practices to support students. That need could be stated as:-

> *A system to contribute to the Neuroimaging Group's curation objectives by enabling researchers to share study documentation and learn how study aims, design, data and analysis are used in producing the study results.*

The 'system' also refers to the human activities necessary to make an information system work; for example the PI's roles as data custodians. Researchers also have a collaborative ethos that they value and which is practised in (for example) their weekly group meetings, where PIs encourage peer discussion and guide more junior researchers on productive lines of enquiry. So the system should *support local collaboration*, and enhance the benefits that researchers already get from documenting their data and its analysis, i.e. that doing so saves time when writing-up for publication; and that sharing experimental protocols helps researchers learn new skills.

Figure 4.2 below gives a high-level model of activities involved in gaining familiarity with the Group's datasets. Modelling at this broad level is concerned only with depicting human activities to be supported rather than how they would be supported. The activities have been worded so they may equally refer to the provider or reader of the documentation. Activities 9-11 are 'control' activities that relate to all of the others.

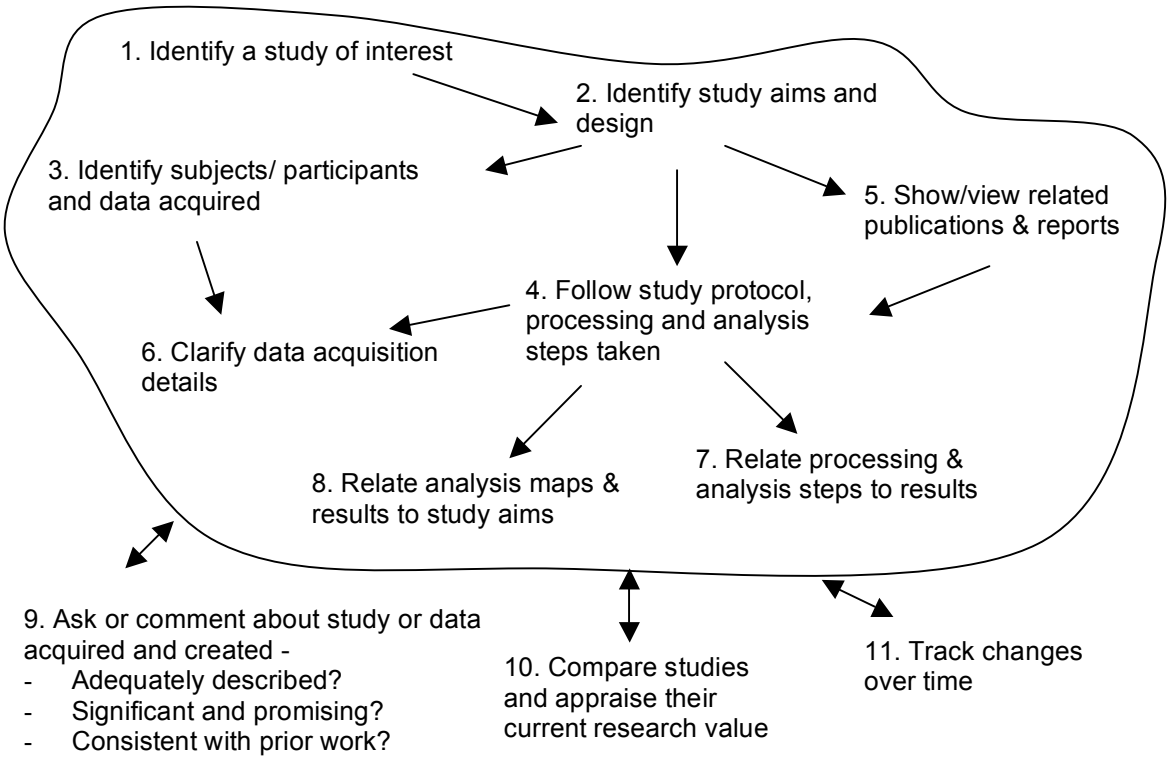

Figure 4.2 High-level model of Data Documentation System

A 'root definition' of the system (Checkland and Scholes, 1999) is given in Table 4.2.

Customers	Junior researchers, new members of group, potential collaborators, the public (eventually, and with no access to subject data)
Actors	Junior researchers, Lead researchers on each study, Principal Investigators/ Custodians, Senior researchers, Others to comment
Transformation	Researchers unfamiliar with studies -> researchers sufficiently familiar with studies to understand how to re-use.
Worldview/ assumptions	Junior researchers contribute to documenting existing studies and datasets since it is in their interests to learn about their scope and the specific data acquired, the ethics requirements, and experimental protocols applied. Senior researchers and PI's contribute to documenting past and present studies as it is in their interests to demonstrate regulatory compliance, and the supervision of research student's documentation. Data documentation activities can evolve to become increasingly standardised, and progressively more open to comments from colleagues initially, and then from others in the neuroimaging community. Doing so and recording the changes made can 1) address risks that datasets lose value through lack of consistent documentation, and 2) better enable retrospective analysis by providing the basis for a standard approach to describing studies and their datasets.
Owner	Dataset custodians, Head of Neuroimaging Group

Table 4.2 'Root definition' of a system for Data Documentation

Environment	1) Research policy interests in making dataset documentation more widely available to promote re-use. 2) Ongoing development of neuroimaging data schemas by collaborative project consortia the group is aligned with. 3) Subject data is highly confidential and can only be shared (internally or externally) in anonymised form. In practice this limits what can be shared even internally. There is also some reluctance to sharing study protocols until these are at final draft stage.

Table 4.2 'Root definition' of a system for Data Documentation (cont.)

4.4 A Phased Approach

A collaborative system for sharing data documentation represents a shift in practice for Neuroimaging Group, since documentation is not systematically shared across the whole group at present, and that would need to be embedded 'internally' before making metadata more widely available online. This section describes short and medium terms towards that, then outlines 'roles and goals' for a system.

Phases of Data Curation and Repository Support

A recent approach developed by Treloar and colleagues at Monash University envisages data curation as a set of 'continua' shown in Table 4.3.

Object		
Less metadata	⟷	More metadata
More items	⟷	Fewer items
Larger objects	⟷	Smaller objects
Objects continually updated	⟷	Objects static/ derived snapshots
Management		
Researcher manages	⟷	Organisation manages
Less preservation	⟷	More preservation
Access		
Mostly closed access	⟷	Mostly open access
Less exposure	⟷	More exposure

Table 4.3 Data curation continua (Treloar et al, 2007).

In the 'continua' approach research communities can be positioned on various dimensions that characterise how they access and manage digital objects. These can be grouped into domains or phases of curation, with the repository support that would typically be appropriate for each, and so provide a migration path (Treloar and Harbroe-Ree, 2008).

Domain characteristics	Typical Repository
Private research: less metadata, more items, larger objects that are often continually updated, researcher management of the items, less preservation, mostly closed access and less exposure.	File system, spreadsheet-based system, or repository management system (e.g. Fedora, Dspace, ePrints)
Shared research: more metadata, fewer items, smaller objects that are usually static or derived snapshots (rather than actively updated data), researcher management, possibly more preservation, and less restricted (but not open) access.	Collaboration support system, e.g. structured wiki or blog-based electronic lab notebook
Publication: more metadata than the collaboration domain, fewer items again, smaller objects that are almost certainly static or derived snapshots, organisational management, more preservation, open access and exposure of metadata for harvesting	Institutional or domain-based data repository

Table 4.4. Curation domains, adapted from Treloar and Harboe-Ree (2008).

The Neuroimaging Group's capabilities are typical of the 'private research domain', according to the characterisation in Table 4.4. Building systems to support data documentation, and the Group's repository objectives, involves a shift in curation capabilities towards shared research and data publication. The two phases outlined below should enable that.

Phase 1: Documentation to Support Group Collaboration

This aims firstly to ensure the file system is structured so there is a clearly identified place for the items comprising the core metadata set – and for other documents required. This lays the groundwork for deploying off-the-shelf collaboration technologies to make documentation easier to track and manage across the Group. This would mitigate the risks that, although files are easily retrieved from archives, the links between related files that are needed to understand the relevance of particular image or data files may become lost or forgotten.

A core metadata set defines what information needs to be recorded, and gives a structure to enable that information to be recorded consistently. Table 4.5 suggests metadata for *projects* and for *analyses* that would be extracted from existing documents and linked to these. This is partly based on metadata that BIRN require for datasets submitted to the BIRN Data Repository (BIRN, 2008)

Access to metadata and documentation could be shared locally on a collaborative wiki or blog, and packaged together with the study file directories for backup to tape. Various blog and wiki platforms provide the capability to define templates, and some also provide simple tables and database functionality. These may be used to provide Group members with a consistent structure for documentation, i.e. forms based around the core metadata set.

Projects		
Metadata element/ fieldname	Description	Links to documents/files
1. Study Ref	Unique study reference	-
2. Study Name	Official title of study, plus any other names	-
3. Brief description	Summary from funding proposal	Funding proposal
4. Funding body	Full name	-
5. Start date	As given in contract	-
6. End date	As given in contract	-
7. Principal Investigator	Name of current Chief/ Principal Investigator(s) and organisation	-
8. REC approval	Name of REC	Application form Approval letter and subsequent correspondence on changes
9. Approved activities	List e.g. types of data that may be collected, whether GP may be contacted, follow-up contact etc.	Informed consent brochure
10. Data sharing criteria	Conditions for sharing with partners and independent researchers	Policy statement for public release
11. Project partners/ centres	Names of organisations if multi-centre	Url links to websites
12. Media coverage	Description e.g. name of media outlet and date	Url links to websites
13. Presentations	Name of event	Presentation slides
14. Publications	Bibliographic references	Pubmed citations, source files or separate bibliography
Analyses		
1. Study Ref	Unique study reference	-
2. Analysis title	Working title of analysis/ experiment	-
3. Minimum Data retention period	End date of project plus number of years stated in contract	-
4. Custodian/ Primary contact	Name, If different from PI	-
5. Researchers involved	Name of each contributor, and organisational affiliation	-
6. Dataset Appraisal	Custodians' choice of e.g. vital/ important/ minor	-

Table 4.5. Neuroimaging Metadata and Documentation

7. Subject groups	Key characteristics of groups, cohorts	Master files(s)
8. Visits	Brief description, dates	
9. Assessments	Names of assessments used e.g. PSE Location references for paper or video records of clinical interviews	
10. Image Acquisitions	Description of acquisition provenance- scanner, sequence, orientation etc.	Spreadsheet
11. Image file formats	Name(s) of format and version	
12. Other Acquisitions	Description of other data collected e.g. Cognitive; DNA; Demographic; Developmental milestones	Spreadsheet
13. Workflow and lab notes	Protocol brief description Software/script name, function, version and brief description Environment, Options Input & Output files Binary configuration System configuration	Protocol document(s) Readme file SPSS set up files Overlay files/ statistical maps Link to version control system
14. Dataset Location(s)	Paths for study directories - image & clinical data	N/a

Table 4.5. Neuroimaging Metadata and Documentation (cont.)

Enlisting Support for Implementation

Building up the documentation system needs the support of junior and senior researchers, beginning with those who are already 'good documenters' of their studies. It may be appropriate to start with a small group, at first with access to each study restricted to the research students and their supervisors. Building up several exemplars would allow a demonstration to the wider group and enlist further involvement.

Whichever technology is used needs to be capable of restricting access to the group and enabling specified pages to be accessible to specified individuals. The initial group could be involved in the choice of blog or wiki technology. In both cases it would be essential to find a working match between the structure imposed by the system (blog 'posts', wiki 'pages' or 'topics'), the metadata set, and the documents concerned. Useful implementation experiences can be found in other fields, for example the 'Laboratory Blog –Book' for documenting biological chemistry experiments developed in the Combechem project at the University of Southampton (Frey et al, 2008).

Roles and Goals for a Documentation System

A well-established approach to requirements definition is to identify *use cases* or interactions a system should support. A first step is to define the 'roles and goals' required (Cockburn, 2001), as shown in Table 4.6 below.

Users	Roles	Goals
New researchers/ masters & doctoral students	Ask questions about documented studies. Document previous studies that may help conceptualise/ design new study	Learn about scope of existing studies and types of datasets acquired.. Learn what methods have been applied to datasets in order to re-use datasets or methods.
Principal investigator - Data custodian, responsible for compliance - Research supervisor	Add top-level documentation about studies e.g. funding bid & ethics approval Appraise research value of study dataset and analysis. Comment on other studies.	Ensure dataset understandable in future Ensure documentation does not disclose personally identifiable data Comply with research council requirements for data preservation & access Guide researchers by checking and comment on whether e.g. a study is adequately described, consistent with prior work, promising analysis, significant results etc. Respond to comments or queries
Lead researcher	Add details of data acquired and study protocols, analyses, presentations, publications etc. Contribute notes on contributions to colleagues' studies Respond to comments Comment on other studies	Account for progress on study Use notes to help draft publications Gather notes from colleagues who contributed Assist colleagues, especially new ones, visitors and students, by showing/ describing a study through the stages of design, acquisition, analysis and publication.
All	Find interesting studies Compare studies on selected attributes Track changes to documentation, when these were made and by whom.	Gain & maintain familiarity with past and current studies Identify possible new lines of research from study similarities and differences

Table 4.6. Data documentation system users, roles and goals

Phase 2: Integrating Datasets and Documentation

Phase 2 builds on phase 1 by further structuring the metadata fields and where it is possible to do so consistently across studies, mapping those fields to the data structures already present in the server filing system and spreadsheets. The aim is to support secondary analyses, both by providing information (the documentation) to guide researchers on the kinds of analyses that may be fruitful – and enabling that analysis.

Integration with Phase 1 which focused on managing semi-structured text for documentation purposes is needed since the distinction between documentation/ metadata and the data itself is not absolute, especially where subjects, visits and acquisitions are concerned. Descriptive notes commenting on this data or that process in a given study will likely contain data that could be used in analysis, but would be more usable with more structure.

A database suited to the purpose will therefore use a combination of ontologies, xml and relational database technologies, as appropriate for neuroimaging data structures. There is a need to combine data that is: -

- Hierarchical and semi-structured, (e.g. protocol documentation; suited for xml storage and retrieval, and for browsing using wikis, blogs or 'electronic lab notebook' software.
- Hierarchical and formally structured, (e.g. neuroanatomical structures); suited to 'knowledge representation' technologies such as ontologies and RDF (Resource Description Framework).
- Associative and formally structured, (e.g. subject assessments) –suited to relational database storage and SQL (Structured Query Language) retrieval.

To progress beyond the roles and goals in Table 4.6 the next step would be to define detailed use cases and scenarios specifying the interactions required with a database. The choice of technologies and the architecture for an integrated database and documentation system would stem from those use cases, and take account of similar developments by other labs. Although that stage is beyond the scope of the SCARP study some characteristics of a phased approach to building a data repository or archive have been identified in this chapter.

5. Human Infrastructure for Curation

5.1 Introduction

This chapter relates social, cultural and legal aspects of Neuroimaging Group's working practices to previous studies of 'human infrastructure' in research communities of practice. Apart from offering a better appreciation of how the Group shares data, this more discursive chapter helps address an underlying question that emerged from the risk analysis; how has Neuroimaging Group 'got by'- managing millions of files for many years without major data loss? The suggesting here is not that Neuroimaging Group is exceptional, but that ordinarily un-remarked aspects of their practices are important to curation, especially as they mitigate risks to data reusability.

Drawing on interviews with Group members and observations of their work, Section 5.2 summarises Group views on data sharing, how these relate to the regulatory requirements for anonymisation of brain images, and the preferred approach of limiting access to known collaborators. The section points out ways that the regulatory environment encourages sharing, as well as the enforcement of constraints on access.

Section 5.3 focuses on the theme of 'stewardship' of data and how routine working practices involve continuous care of data. There are points in the research process where data documentation offers researchers direct benefits, i.e. rather than the more indirect one of data preservation. Data description, though not formalised, is a normal occurrence- for example descriptions for the benefit of newcomers about how and why data has been acquired and used, and why a dataset may be worth re-using. Interaction between group members is 'heedful', part of the *human infrastructure* that previous studies of neuroimaging e-infrastructure have seen as critically important.

5.2 The regulatory environment: constraint or enabler?

Neuroimaging Group members' efforts to advance data integration and sharing between research centres were described in the previous chapter. Neuroimaging group members shared the concerns noted in the literature review; over the need to observe legal and ethical limits on data sharing, and over the misinterpretation of results by researchers distanced from the study context. While they were in favour of limiting data access, they also saw the regulatory requirements as an opportunity for innovation.

Anonymisation and the identification of medical images

Members interviewed had concerns about the consequences of disclosing personal data. Nevertheless they believed their measures to protect research subjects' identity were enough to make that unlikely. These include developing 'skull stripping' image processing techniques to remove the potentially recognisable ears and facial features, as mentioned earlier.

Imaging data is de-identified after quality assurance has determined that the person scanned matches the person expected. After that, the study PI is gatekeeper of any identifying demographic details, while senior imaging researchers ensure that scans are also de-identified. However multi-centre projects raised complex issues about balancing individual privacy by pseudonymising shared data, with the recruiting centres' clinical research need to be able to re-identify it, for example to carry out repeat scans in longitudinal studies.

The sensitivity of neuroimaging data was regarded as an *increasingly* difficult area, as advances in image analysis exploit their fingerprint-like qualities, and enable easier prediction of psychiatric disorders. Researchers thought it probable that some neurophysiological features identifiable from brain scans– cortical folding patterns for example - are unique to the individual. The Neurogrid project has developed for QA purposes the capability to automatically compare scans to assess with high level of accuracy whether they come from the same person.

The probability of identity disclosure is currently very low since it would require a high level of neuro-scientific knowledge to match an anonymised scan from one source with one from another source linked to personally identifying data. But this can be expected to change with automation, with increasing numbers of people scanned, and more especially if these are shared on public databases. It is likely to become easier to identify when two brain images come from the same person - and to predict illness from those scans. This raises the possibility of brain images becoming regarded as biometric data in the near future; different from iris or fingerprint images only in the relative difficulty and cost of matching them to the person they were acquired from. Currently biometric identifiers must legally be removed before depositing imaging data in the BIRN Data Repository (under the U.S. HIPAA Privacy Rule), suggesting potential legal difficulties for public sharing models in the future. Group members' views on sharing could be characterised as cautious, but supportive of innovation that would allow controlled access, as for example in the Neurogrid project mentioned previously.

Data reuse: trading access and skills

Obtaining new brain scans from patients or healthy volunteers requires ethical approval, which in turn demands the time and effort of senior researchers to ensure that research students do not propose ill-considered research that could cause distress to patients. Students, especially those studying for Masters degrees by research, may not be with the Group long enough for an ethical approval application to even be considered. However they can demonstrate their worth as novice imagers by producing interesting results from old data. Neuroimaging Group postgraduate students are therefore *internal re-users* of datasets. This allows students to gain practice in imaging on anonymised datasets and, with the guidance of senior researchers, to obtain new and valuable results from those datasets. As well as providing these potentially mutual benefits to student and research group, this arrangement is *necessary*, as otherwise the time and effort needed for students to gain practice would be too great.

Neuroimaging Group members expected that metadata would be shared with *collaborators* in multi-centre projects. Researchers supported the principle of linking metadata about studies to publications derived from the data, but had lower expectations about making this information publicly available. Concern to protect the research value of the data held is as much a factor in wider data sharing as subject privacy; data sharing whether within the Group or more broadly is seen as a form of trade – a 'give to get' rather than 'give away' approach as one member put it. Internal sharing and reuse depends on senior researchers and students recognising some mutual benefit, with some prior weighing-up of the others' capabilities; access to data is not given wholesale or to all-comers.

Research council policy requires custodians to balance the benefits of sharing with safeguarding of privacy and patient welfare. However at time of writing there was little practical advice from the research councils on how to strike that balance, for example in drafting data access policies. Legal/ethical risks of sharing neuroimaging data are likely to change in line with the advances in diagnostic and data integration capabilities. For neuroimaging data custodians, the balance appears to be between high risks of

disclosure and the (by definition) uncertain benefits of publicly sharing data with unknown others.

To summarise, the legal and ethical frameworks that demand de-identification of medical images, and are typically seen as a *constraint,* can also be an *incentive* for sharing (e.g. for newcomers to neuroimaging to practice), and for innovations in the neuroimaging field. Collaboration might for example be stimulated through making research centres study metadata discoverable by trusted members of their research community. Collaborations are negotiated, and depend on would-be collaborators' mutual awareness of their domain knowledge and shared trust in their commitment to some common enterprise.

3.3 'Heedful Interaction' and Curation

> *"It is essential that decisions on selection for… preservation and curation form part of an organisational process and are not made on an ad hoc basis. This requires ongoing processes for care, and selection for retention or disposal."* (Beagrie, 2007, p.7)

> *"Being careful is a social rather than a solitary act. To act with care, people have to envision their contributions in the context of requirements for joint action. Furthermore, to act with care does not mean that one plans how to do this and then applies the plan to the action. Care is not cultivated apart from action. It is expressed in action and through action."* (Weick and Roberts, 1993, p.369)

> *"Successful neuroimaging laboratories tend to be those in which there exists an active and dynamic interaction of …specialities… data-sharing with peers, both within and between scientific disciplines, is an inherent and necessary component in the science of neuroimaging."* (Van Horn, 2001, pp.1323-1324).

It seems uncontroversial to say that data curation or 'stewardship' implies *continuous care*; Buneman at el (2008) point out that curation derives from the Latin verb *curare* 'to care for' and describe models for ensuring the continuous care of databases as they evolve through comment and critique. Dictionaries associate curation more with 'control' or 'organisation', which is in line with the first quotation above, suggesting that 'ongoing processes of care' require formalised managerial processes. Much of the data curation literature stresses formal, explicit processes and some of the recommendations in the last chapter echo that (for example formal appraisal criteria for datasets).

On the other hand SCARP is also concerned with *disciplinary practices*. The quotation by Van Horn associating successful neuroimaging labs with 'dynamic interaction between specialities' points to the need to look beyond formal planning processes to the more day-to-day interactions in a research group. In doing that, this section takes up organisational sociologist Karl Weick's work on the social nature of care, and more specifically the 'heedful interaction' identified in the Neuroimaging Group's weekly meetings, arguing this is a form of human infrastructure for curation.

The importance to the Group of weekly meetings, and their relevance to curation, can be seen in responses to a questionnaire, which were used to inform the risk analysis and posed the following scenario:

> " You need to find and use the data from a study that is several years old and was worked on mostly by the PI and researcher who are not available to help. You have been told there is a project folder on the server that will have the scans in it, and there must be spreadsheets somewhere with the demographics, clinical and

behavioural tests that were carried out. You want to bring everything together, and repeat the analysis that was done but with some new variables. How helpful would you find each of the following".

The most highly rated approach from 6 alternatives was "raising it at a weekly meeting to see if anyone knows more" with "asking the systems/data manager" a close second[9].

Observations from attending the Group's meetings began earlier in the study, and were recorded in field notes[10] between November 2007 and March 2008. These meetings were primarily the weekly 'structural' and 'functional' meetings at which colleagues discussed their work, each chaired by a senior psychiatrist specialising in sMRI and fMRI studies respectively. The aim of attending these was firstly to gain some understanding of the nature of the group's work and terminology used, and then to focus on 'data practices', i.e. what group members reported they did with their data, and any reported issues relating to accessing, sharing or reusing it. It quickly became clear that Group meetings were themselves relevant as a 'data practice'. Data-related work was done in and through the meetings. To unpick how, this section uses two related perspectives from sociological studies of collaboration in communities of practice; 'heedful interaction', and 'legitimate peripheral participation' as a framework to describe three main observations:-

- Weekly reporting meetings accomplished activities that can be considered part of the curation lifecycle.
- The 'heedful' nature of the interactions mitigates risks to datasets by sustaining coordination of data storage tasks with the systems administrator, and encouraging researchers to describe and learn about otherwise unfamiliar datasets.
- The Group's collaborative working includes an informal apprenticeship style of learning about imaging methods and datasets. Considering this as a kind of 'data description for relative strangers', such as visiting scholars and students, was useful in identifying how and when to gather preservation information, as a basis for a more standardised approach to data documentation.

As a first step, we can characterise some aspects of the Group meetings in terms of the DCC Curation Lifecycle Model (Figure 3.2). While the specific details discussed in meetings are not the concern of this report, the summary descriptions below relate some of the topics discussed to lifecycle actions:

- *Conceptualise*: topics relating to study design included the utility of particular models, experimental protocols likely to provide interesting results, and how other centres' published research related to the Group's ongoing studies.
- *Create or receive*: members frequently gave updates on scans they had acquired; other topics included the possibilities of gaining access to new forms of instrumentation to acquire new forms of data.
- *Appraise or select*: meetings reported over several weeks on the status of a scanner fault, the resulting artefacts found in some scans, and (later) on how to detect and fix them.
- *Ingest*: discussion included which datasets to make available on a federated database, the methods for making it available and the rationale for doing so.

[9] The questionnaire and responses are detailed further in the 'Neuroimaging Data Landscapes' annex to this report.

[10] For confidentiality reasons it was not appropriate to record the meetings. Field notes described discussions of Group members' unpublished research results and how these related to other centres' work.

- *Store*: members occasionally reported on the volumes of data they had acquired or planned to acquired, and what they might need to ask of the systems administrator.
- *Access, Use and Reuse*: e.g. the possibilities and benefits of junior researchers performing specific kinds of analysis on previously acquired datasets.
- *Transform*: one meeting focused on the possibilities and challenges of mapping psychosis symptom scales to enable datasets to be joined.

It is not surprising that meetings dealt with aspects of the 'curation lifecycle'; it is after all intended to be relevant to any organisation managing data. It is more useful to consider *how* group members interacted in these meetings, and how that related more generally to mutual awareness of the data gathered and what was done with it.

A high level of cooperation between Group members was clear from early in the study; from observing weekly meetings and from other aspects of the working environment; for example researchers rarely closed their doors and colleagues would often drop in announced during interviews. The rest of this section tries to unpick what it was about their cooperation that sustains the reliable exchange of knowledge about datasets and what to do with them – in other words 'data description' as it is currently practised.

Heedful interaction and organisational reliability

By the second month of the study it was already apparent that while data storage was an ongoing problem, data loss was not thought a major issue. To an outsider it seemed strange that the Group had managed in over a decade of studies of international repute to have few experiences of dataset becoming lost or unusable, yet working with sizeable datasets spread over various disparate filing systems and little to separate them from disastrous data loss (backup procedures notwithstanding). While field notes and interviews were written up, re-listened to and transcribed, the nature of interactions between Group members emerged as a recurring theme. In early interviews where members talked about their roles in the group they often referred to the richness of the datasets they worked with, and what they valued about the interdisciplinary environment. When observing their meetings, meanwhile, the variety and density of professional terminology made following these extremely difficult. This was discussed in later interviews to explore how newcomers overcome these difficulties.

The term 'heedful interaction' is used to describe the relating of one's work processes to others' work process carefully, critically, consistently, purposefully, attentively, vigilantly, and conscientiously (Weick and Roberts,1993). The importance of this to curation is that heedful interaction, as distinct from habitual interaction or repetitive behaviour, is associated with *reliable* performance. Heedful interaction, according to Weick and Roberts, rests on the notion that individuals create the social forces of a group when they act as if there are such forces. In doing this, they construct their actions (contribute) while envisaging a social system of joint actions (represent), and interrelate that constructed action with the system that is envisaged (subordinate). It is possible to analyse interactions as more or less 'heedful' – and collectively 'mindful' - to the extent that these three aspects of contributing, representing and subordinating are done with care: -

> "The more heed reflected in a pattern of interrelations, the more developed the collective mind and the greater the capability to comprehend unexpected events that evolve rapidly in unexpected ways. When we say that a collective mind "comprehends" unexpected events, we mean that heedful interrelating connects sufficient individual know-how to meet situational demands." (*ibid* p.367)

Originating in studies of flight-deck operations on aircraft carriers, the approach has also

been used to study reliable software development in engineering teams (McChesney and Gallagher, 2004). While the 'collective mind' concept is not essential for our purposes, the properties Weick and Roberts identify with careful action were useful in exploring aspects of Neuroimaging Group's practices that sustain mutual awareness of their datasets and expertise in analysing them, beginning with the inter-dependence of their work.

The growing inter-dependence of imaging and clinical work

Research in the Group involves highly inter-dependent skills and practices. This is true of neuroimaging generally; the field is historically dependent on reliable communication between disciplines. Cohen and Baird (1999), in a study applying Galison's notion of scientific innovation as a 'trading zone' (Galison 1997), describe how early MRI images served as a 'pidgin' language between the scanner engineers and clinicians.

Miscommunication between clinical and engineering disciplines has historically had unfortunate clinical consequences; in the mid-1980's for example there were many misdiagnoses of spinal disorders due to clinicians mistakenly interpreting image density in scans. The cause of this was an imaging artefact unknown to clinicians, though well known to MRI physicists and engineers, and eventually brought to the medical community's attention by an individual with clinical and engineering knowledge (Cohen and Baird, op.cit.). Cohen and Baird provide further examples demonstrating the need for "persons who bridge instrument engineering and clinical uses of MRI instrumentation" (*ibid* p.250). Using Galison's terminology again, this inter-disciplinary development continues to happen through individuals from different disciplines crafting an 'inter-language' or 'creole' that allows them to trade expertise while still working within their different theoretical spaces, through actions which commit them to a shared set of goals (*ibid.*) or, as Weick might put it, 'heedfully' inter-relating and subordinating their contributions to the purposes of an envisaged system.

Research in Neuroimaging Group has accomplished this inter-disciplinary working in a similar manner. A senior imaging researcher in the Group described the increasing inter-dependence, which has resulted in researchers from engineering and informatics backgrounds becoming first authors of published articles in medical journals. It is seen as important by imagers to 'always have a psychiatrist on top of the analysis' while that analysis involves clinicians and engineers learning each other's language: -

> "…we all have to learn new things when we come here otherwise we can't really contribute.. we have to learn a lexicon of terms which are entirely new… and there's a lot of my language creeping in to the language here.… the learning process for me was really centred around a paper that I wrote within a few months of starting…In the process of writing that paper I learned a great deal about brain imaging and how that is analysed… the sort of norms of analysis… there's a whole language there and very soon I realised that because of how this group is positioned in relation to other research groups everything you do does have an effect….(CN, senior imager, int.6)

The image analysts depend on their clinical colleagues for the experimental hypotheses, the clinical data/metadata acquired from research subjects, and their coordination of the acquisition of scans. While clinical researchers perform some image analysis themselves, the post-doctoral researchers (who have engineering and informatics backgrounds) do much of it in an internal consultancy or master-apprentice role, and so the working relationships between different specialties have much to gain from a collaborative ethos.

Systems administration in the Group also relies on willingness of members to cooperate.

Critically for the issue of storage management, the systems administrator depends on researchers anticipating when they will need large amount of server storage space. Since this happens often, the allocation of space requires coordination with them:

> "If people start chewing up huge amounts of disk space very quickly then I'll notice… and I'll speak with them… and all someone has to do is come to me and ask for more disk space- but you know its less manageable if they just decide to chew up and fill a volume up for instance … it then becomes an issue for other people because y'know it's a shared resource… and that's worked well because people come and ask me for space now and what happens is that I allocate the areas and tell people how much they've got and they then use it and I keep an eye on it" (TH, systems administrator, int. 7)

Other members identified close cooperation with colleagues as one of the distinctive features of Neuroimaging Group, particularly for those new to it, a "very big feature of how things seem to work" as one put it.

Learning from others' documentation

A 'chain' of mutual help is notable as a feature of interaction between those new to the Group and its relative 'old timers', as well as between the established imaging and clinical researchers. Newcomers to the Neuroimaging Group, whether as research staff or pre- and post-graduate students, learn about datasets as an integral part of learning how imaging is practised in the Group. The process is highly coordinated and structured for all its apparent informality, and weekly meetings (which we return to shortly) are key to that for the researchers interviewed. In terms of documentation, also important were MRI analysis software manuals, and *other people's* study protocols and 'readme files'. Protocols are documented in the design phase; for PhD and post-doctoral researchers they are part of the ethical compliance regime, since ethics committees need to know how data will be used. Readme files are usually shorter notes on software parameters used, the software output, and where results are stored.

The experience of being a newcomer is described here by one of the doctoral students in the group: -

> "There's not an official guide into how you would go about doing your analysis… it's a lot of feeling your way …the nuts and bolts of how they organise studies and in terms of what you have to do with imaging data for it to make any kind of sense was all completely new… In terms of how studies work and how research works in the department weekly meetings [helped] … its one of those things you just have to expose yourself to continually and all of a sudden you realise you've got an understanding and you've not sat down to learn about it it's just- especially when you get new masters students in and they're asking about things and you realise you do understand a lot more about how the department works that you've just picked up by osmosis really" (SQ, research student, int.36)

Knowledge of datasets and analysis techniques is passed from students who are relative 'old hands' to newcomers. This is not so much a matter of adopting standardised work routines as of how to *adapt* others' protocols and software scripts that are close to what is needed. As noted in Chapter 1, processing techniques in neuroimaging are constantly evolving and so acquiring these is more about learning how to inter-relate local preferences than about absorbing standard operating procedures: -

> " …there are certain common things that always come up… yes though the other thing I find tricky with it is… if I compare it to previous research work I've done… 'this is how you do this technique' you get shown how to do it and it works… with imaging its not really that there's a right way and a wrong way, there's different ways of doing things … people have their own methods that work for them, they

> like to strip in Matlab, they like to use a user interface… it's achieving the same thing but there are different ways of doing it and when you're new to it you don't know what you're preference is going to be." (SQ, research student, int.36)

The 'chain' of knowledge exchange involves learning how to draw on others' varied sets of skills and experience; while one researcher may be more experienced in looking at datasets and what 'sensible' data looks like, others will have more experience with scripting in particular software versions, or have greater engineering knowledge of how scanners work. Sharing documentation is an important part of this. Researchers' notes in 'readme files' and 'master files' are a reference for others; recording (for example) the scanner sequences, software parameters, or script versions they have used.

Learning as 'legitimate peripheral participation'

The 'chain' of learning in the Group can be described as a form of 'legitimate peripheral participation' (LPP), a perspective on learning developed by Lave and Wenger (1991) from studies of the enrolment of newcomers in communities of practice. From this perspective, newcomers gain membership of a community initially by taking peripheral roles in the activities of expert members. Newcomers gain expertise and membership by observing experts and participating in progressively more complex tasks as they become familiar with the terminology, routines and organising principles of the community, gradually becoming adept enough to be considered experts themselves (ibid, Wenger 1998).

Neuroimaging Group postgraduate students' re-use of datasets involves learning through participation in others' projects. This allows students to gain practice in imaging and, with senior researchers' help, to obtain new and valuable results from those datasets.

> "What I tried to do is before I had my own data coming in was try to use other people's datasets or get a little bit involved in their analysis …getting a little bit involved in what other people are doing – for example saying like I could do the reliability for the tracing you're doing on that bit of imaging – just means you're involved for that bit of the project and you pick up other bits of projects and you see how somebody manages their dataset" (SQ, research student).

As well as providing mutual benefits to new and senior members of the group, this arrangement also reduced risks to students and patients; a consequence of the need (mentioned earlier) for ethics committee approval, a time-consuming process relative to the duration of a Masters course. Research proposed by inexperienced students would be more likely to fail this hurdle, since ill-considered research could cause distress to psychiatric patients.

The *weekly meetings* are a site of legitimate peripheral participation; and not only because they familiarise newcomers with imaging, neuropsychiatric and statistical terminology. Senior members of the group steer the proceedings so that members are made mutually aware of what datasets they are working on and for what purposes. In the course of these discussions, principal investigators frequently encourage pre- and post-doctoral students to contribute to each other's work; often suggesting who to interact with and which analysis techniques are "good to learn".

Weekly meetings feature informal presentations in which a group member will talk through some newly derived data, or an article from recent literature related to the Group's work. Principal investigators (PIs) then lead others, mainly the post-doctoral researchers, in offering constructive critiques of their colleagues' work, often positioning it against that of other imaging labs by pointing out similarities between 'their problems'

and ‘our problems’, the superiority of the Group’s imaging techniques, or inconsistencies in the statistical analysis reported in other labs’ work.

Presentations of newly derived data and work reported in the neuroimaging literature provide opportunities for senior researchers to demonstrate the range of competencies that are expected of group members, and how to apply them. For example one meeting, highlighted journal articles by doctoral students in another imaging lab reporting apparently inconsistent statistical results from the same dataset. In the next meeting senior researchers involved in two studies on the same dataset described how such comparing of notes on their study protocols would avoid the risk of such inconsistencies.

Presentations of interim analyses are routinely used to identify potentially promising lines of enquiry, in which post-graduate students are encouraged to work together and are praised when they do. For example one meeting focused on issues of ‘subject habituation’ as a confounding factor in a student’s analysis of fMRI study on subjects’ responses to a series of photographs. The PI suggested looking more closely at what was going on in the time series, recommended ‘interacting with’ another (absent) student on this, then answered questions about which brain regions might be a good place to look, leading another doctoral student to suggest factorial analysis as a good statistical means to identify what was going on. This ‘excellent suggestion’ was followed up several weeks later by a presentation of the factorial analysis where both students gave an account of their joint efforts towards the first’s thesis. Talking through a succession of print-outs showing the task stimuli, the data acquired on the subjects and control groups, and three levels of analysis, led to increasingly excited senior researchers pointing to overlay maps and tables (showing the statistically significant correlations between brain activation and experimental variables), and excited talk of these being ‘just where you would expect them to be’.

The point here is not that collaboration ended in ‘good’ results, nor to comment on the particular techniques used. Rather the course of events demonstrated several *unexceptional* aspects of the Group’s practice; that in their interactions they interleave judgements about what is sensible data, what lines of enquiry and techniques are worth pursuing, who it is worth pursuing them with, and why some results are ‘dodgy’ and others ‘exciting’. While doctoral researchers and relatively junior Masters students are often passive observers to the proceedings, they are encouraged to contribute and when they present their own analyses these undergo an informal, constructive and rigorous form of peer review in these meetings.

Acquisition and appraisal: improving QA through heedful interaction

During the study an unusual and potentially serious scanner fault was the subject of discussion in weekly meetings. This fault led to improvements in quality assurance being quickly identified, in a manner that illustrates the relevance to data curation of the ‘legitimate peripheral participation’ and ‘heedful interaction’ found in weekly meetings.

As mentioned in chapter 3, newly acquired scans are routinely ‘eyeballed’ or checked visually for image artefacts or ‘noise’ such as banding which make subsequent analysis of relative brain matter density difficult or impossible. At one of the weekly meetings a doctoral student (‘SK’) reports she has noticed artefacts in some of the scans she has acquired, and is discussing it with one of the post-doc researchers (‘ML’); an engineer and scanner expert, as a matter of concern. The problem quickly escalates; at subsequent meetings the unresolved scanner problem is the main topic of discussion; it has become a ‘major worry’ since SK reported that image artefacts have not only re-occurred, some are invisible to the naked eye so it is not clear for how long the scanner has been malfunctioning. The exchanges are more intense than usual; senior clinicians

resolve to raise this with colleagues in the hospital concerned, senior imagers and doctoral students give accounts of collective efforts to scrutinise scans. The imaging specialists give their assessment of the risks that datasets using the various scanning modalities have been affected. Their discussion draws in the array of normally unmentioned specialists who sustain the service and are relevant to fixing it; medical physicists at the hospital, administrators who manage the service, engineers from the manufacturer who are reportedly on site. Forthcoming scanning sessions are called into question; some of these involve arranging transport for 'hard to get' patients who live hundreds of miles away. These are quickly prioritised to anticipate the effects on patients and on colleagues' studies.

By the next meeting the issue is partly resolved; although the scanner is still out of action and the PIs have decided to postpone all scanning sessions, SK mentions she has written a script, with the help of ML and the hospital's scanning manager, to identify the artefacts in new and existing datasets and in some cases correct errors found. It turns out that only a few scans are affected. At the next weekly meeting SK demonstrates the script. Using printouts of the previously unnoticeable artefacts, and intermediate steps in processing them, she talks the group through a technique of analysing structural features detectable in the non-brain parts of the image to identify various categories of artefact she has found. Over the course of several weeks the senior imagers and systems administrator work with SK and ML to combine their script with that of another doctoral student to construct a new QA toolkit, and the Group's head coordinates new procedures to routinely use this on all newly acquired scans.

Human infrastructure for data curation

The 'scanner artefacts' episode, and the apprenticeship-like interactions of research students with each other and senior researchers, highlight practices that are not identified *as* curation but nevertheless involve the sharing, description, critique, re-use and care of datasets. This final section makes some brief comments on how these practices can be considered 'human infrastructure' for curation

The notion of human infrastructure has previously been highlighted in the study of fBIRN by Lee at al (2006), and draws on previous work by Star and Ruhleder (1996) on the relationship between human and technological infrastructures. Their influential work identified eight aspects through which social practices and institutions lend infrastructure its significance (see Edwards et al, 2007). One of these is that infrastructure is learned as a part of membership, in that technical and organisational arrangements come to be taken for granted by members. A key aspect identifiable in this study is that data sharing and documentation are *learned as part of membership* of Neuroimaging Group; protocols and arrangements for sharing access to data are used by newcomers to work across sub-disciplines and skill levels. Other aspects listed above have been highlighted; the weekly meetings are a convention of practice, affording opportunities to learn the interdisciplinary terminology. The organisational and technical arrangements for handling scanner artefacts became visible when they became a problem.

Star and Ruhleder's dimensions do not separate human and technical infrastructure, rather they are meant to highlight that they are interwoven; infrastructural is temporal and occurs when the tension between local and global is resolved (ibid). Paraphrasing Lee et al (2006), the point of highlighting 'human infrastructure' is to pay attention to the ways that organisational arrangements for curation become infrastructural, rather than assume that formal processes or technologies to support it can be overlayed on practice and 'just work'. The study has illustrated the role of 'dynamic interaction' between specialities in sustaining successful imaging labs (Van Horn, 2001) and that the 'heedful' interaction evident in the Group's research practices already contributes to ensuring that data is accessible, properly acquired, properly shared, and productively re-used.

5. Conclusions

The study with University of Edinburgh Division of Psychiatry's Neuroimaging Group has drawn attention to their curation practices and recommended steps in the direction of improved data documentation to mitigate risks to the informed re-use of their datasets. The conclusions in this section highlight implications and recommendations for DCC and research policy-makers, following the three main themes below. These acknowledge the limits of any qualitative study of one laboratory, and briefly reflect on the limits of this particular study. The report has summarised Neuroimaging Group members' reported views, practices, and experiences of caring for their data, and drawn inferences about how curation practices might change to their benefit. Some the particulars of the case illustrate and exemplify themes evident in recent neuroimaging literature, and of course draw on the participants' knowledge of the neuroimaging community, but they do not provide a basis for the kinds of generalisation from sample to population that is characteristic of quantitative survey research.

"Think global, act local" to build metadata exchange capabilities

Curation needs human infrastructure and this should be taken into account when assessing curation capabilities. The study shows how researchers and investigators heedful attending to each other's data underpins curation. Neuroimaging involves continuous care of increasingly large and dynamic datasets. Neuroimaging investigators are custodians of millions of images and, to contribute to medical research, these need to be related to richly varied and highly sensitive personal information on research subjects. Some of that data is being shared, including in e-science projects aiming to provide federated data storage and improve data integration (see below). The large majority of datasets are held at lab level however, with access governed by Principal Investigators under terms set by Research Ethics Committees. Compliance with these terms and protecting personal data is of more immediate concern to researchers than sharing data with independent researchers in other laboratories or fields. Rather, data tends to be shared on a quid pro quo basis both within the laboratory and with external collaborators, when legal and ethical constraints allow it and there is evident benefit to be gained from exchanging access to data and/or analytic methods. It would be more accurate to see this as a form of 'gift exchange' between data custodians than as 'sharing'.

Interest in re-using datasets is mainly in the areas of using novel analysis techniques to identify patterns in images or in the associated clinically-related and demographic data on subjects, and (among the researchers interviewed) less in re-using derived data to replicate previous analyses. Documentation and metadata on research subjects and on analytic protocols is key to any form of re-use, and is encouraged by the ethics compliance regime. Images, associated subject data, and structured contextual and provenance information about these need to be inter-related. Lack of that structured, standardised documentation is a major source of risks to datasets long-term re-usability, yet this is an area that is reportedly under-invested in.

Standardisation in neuroimaging methods and data documentation is driven by the need for larger datasets to enable studies with higher reliability. This requires larger-scale collaboration and hence wider trading of methods and data. The top-down data sharing policy framework put in place by the MRC and Wellcome Trust needs to be accompanied by further ground-up initiatives to exchange semi-structured data between imaging centres. Neuroimaging research has strong potential to benefit from e-research tools and infrastructure, as the large-scale US investment in BIRN indicates. Borrowing

the environmentalist slogan, there is a need for UK research funders to "think global and act local" to support the development of data curation in this domain. The UK neuroimaging community is well-placed to further develop models for achieving that, following the examples of Neurogrid, PsyGrid, NeuroPsyGrid and Carmen. However it needs investment in tools to support a gradual transition from inter-personal and study-level sharing of neuroimaging metadata to wider dataset 'trading' and collaborative re-use. Such tools should be simple to deploy and use by neuroimaging researchers. They should enable researchers to structure their study documentation and link it to relevant datasets, and to make the resulting metadata selectively and securely available; and they should enable potential collaborators to easily find relevant studies through metadata aggregation services.

Data integration drives new curation requirements

Multi-centre neuroimaging collaborations are augmenting existing curation capabilities, adding value to datasets by enabling them to be integrated for re-analysis purposes, and fostering innovations in image analysis through transfer of techniques from informatics disciplines. Examples include development of image normalisation techniques to harmonise image data from multiple scanners, and automated analysis of images to enhance productivity. These in turn add to the variety of contextual and provenance information needed to track data as it is integrated from disparate sources and analysed by multiple people and/or centres.

Frequent change in the analytic methods used in neuroimaging makes the need for structured documentation more acute. Community standards for recording provenance and representation information are urgently needed in the neuroimaging community, and transferable techniques are likely to be found across other fields of image-based research. Meanwhile, effective exchange of data and methods is likely to be hampered by inevitable changes in the schemas used to describe these.

Recommendation 1 ~ DCC should further investigate and map provenance information management requirements in neuroimaging and other fields of image based research, to provide better advice on tools and methods to address these requirements.

While novel analysis techniques make retrospective analysis of imaging datasets increasingly promising, this makes appraisal of the value of imaging dataset more complicated. For example Neuroimaging Group researchers have reported achievable benefits from using ontologies to combine MRI datasets across centres, to enable cross-analysis of psychosis and other datasets. Researchers and funding bodies need to make informed decisions about whether greater value is obtained from gathering new data or re-using the old in new ways. This coincides with an increasing need to appraise the value of data amassed from long-running longitudinal studies that have been sustained through successive projects and custodians.

Recommendation 2 ~ The neuroimaging community requires further support to assess the viability and usefulness of combining existing MRI data sets on psychosis and other neuropsychiatric disorders.

Recommendation 3 ~ DCC should further investigate and map factors that affect the value of re-using imaging datasets, to enable that value to be measured and support better advice on appraising and valuing datasets.

Recommendation 4~ DCC should develop and provide guidelines, advice and templates for data access policies, using neuroimaging as an exemplar of the challenges of reconciling the requirements for data confidentiality and more open access in medical research. This should be supported by stakeholders such as the MRC Data Support Service, and is in keeping with the recent interim report of the UK Research Data Service Feasibility Study (SERCO, 2008), which identifies a requirement for more

advice on practical issues related to managing data, including help producing data management/ sharing plans.

Integrating 'good curation practice' into research training

Neuroimaging labs are interdisciplinary communities of practice whose members need to share data and skills. That is especially so for newcomers, who are required to re-use datasets and research protocols to learn the practical skills of image analysis. Junior researchers learn by participating in colleagues' studies, directly benefit from sharing experimental protocols, and could play an active role in standardising study documentation and collecting metadata. Integrating these tasks into research supervision may benefit students by helping them identify the characteristics of datasets that are essential to re-use, while also alleviating the bottleneck that manual metadata creation is regarded as by senior researchers. Ethical clearance procedures engender thorough documentation of research protocols at the outset of projects, providing an opportunity to link training on these procedures with training on curation lifecycle management, adapted to meet the needs of the neuroimaging field.

Recommendation 5 ~ DCC should support the development of digital curation in neuroimaging and related fields by providing curation lifecycle management training targeted at doctoral or masters students and briefing materials targeted at research supervisors.

Risks to dataset reusability reflect the disciplinary mix in neuroimaging; clinicians and imagers have tended to manage different kinds of data; while clinicians are data custodians concerned with close personal management of demographic data, imagers have historically required network servers and archiving resources to manage larger image datasets. The case for integrating demographic and imaging datasets coincides with growing convergence between the neuropsychiatric and imaging domains, e.g. as imagers have developed capabilities to contribute to the psychiatric domain.

The report demonstrates the need for case studies of how "enablers and barriers" to data sharing, curation, preservation and reuse operate on the ground in particular research communities. For example the current study has documented how the 'lack of standardisation of neuroimaging methods' reported in the neuroinformatics literature affects data sharing between early career lab researchers with differing skills levels or disciplinary backgrounds. A focus on how newcomers attain membership of research communities also helps to address one of the major difficulties of 'immersive' case studies- that they require an understanding of the terminologies and competencies needed to do research in the host research community. Relatedly, if case studies are to benefit host teams they require easily and quickly transferable tools to apply 'best practice' in digital curation. In the current case DRAMBORA needed some adaptations to apply it outside of its main target group of established archival organisations.

Recommendation 6 ~ DCC should adapt the DRAMBORA risk assessment tool to enable it to be easily used by data custodians at the department or research team level.

Acknowledgements

Many thanks are due to Professor Stephen Lawrie and all members of the Neuroimaging Group, Division of Psychiatry, University of Edinburgh who contributed to the study for their support, comments, generosity and patience.

References

Ainsworth, J., Harper, R., Bridges,L., Whelan, P. Vance, W. and Buchan, I. (2007) 'The Challenges of Clinical e-Science: Lessons Learned from PsyGrid' *Proceedings e-Science All Hands Meeting 2007*.

Ashburner, J. Flandin, G. Henson, R. Kiebel, S. Kilner, J. Mattout, J. Penny, W. Stephan, K., and Hutton, C. (2008) SPM5 Manual Wellcome Trust Centre for Neuroimaging, Institute of Neurology, University College London. Available at: http://www.fil.ion.ucl.ac.uk/spm/

Beagrie, N. (2007) 'e-Infrastructure Strategy for Research: Final Report from the OSI Preservation and Curation Working Group' National e-Science Centre Available at: http://www.nesc.ac.uk/documents/OSI/preservation.pdf

Beaulieu, A. 'Images Are Not the (Only) Truth: Brain Mapping, Visual Knowledge, and Iconoclasm' *Science, Technology, & Human Values*, 27(1), pp. 53-86.

BIRN (2008) *BDR Submissions Overview* Available at: http://nbirn.net/bdr/overview_of_submissions.shtm

Borgman, C. (2007). *Scholarship in the digital age: Information, infrastructure, and the Internet*. Cambridge, MA: MIT Press.

Buneman,P., Cheney, J., Tan, W., Vansummeren. S. (2008) 'Curated databases'. Proceedings PODS 2008 June 9–12, 2008, Vancouver, BC, Canada. pp.1-12 ACM: New York

CCSDS (2002) 'Reference Model for an Open Archival Information System (OAIS). Consultative Committee for Space Data Systems' January 2002. Retrieved 30 June, 2008, from: http://public.ccsds.org/publications/archive/650x0b1.pdf

Cockburn, A. (2001) *Writing Effective Use Cases* London: Addison Wesley

Cohen, M., and Baird.,D. (1999) 'Why Trade?: How zones of trade support epistemic stability.' *Perspective on Science* 7 pp. 231-254.

DCC (2008) *Sharing Medical Data- The Legal Considerations* available at: http://www.dcc.ac.uk/resource/legal-watch/

DCC/DPE (2007): Digital repository audit method based on risk assessment (DRAMBORA), version 1.0, Digital Curation Centre, Digital Preservation Europe, Available at: http://www.repositoryaudit.eu/

Day, M. (2007) Curating our Digital Scientific Heritage: a Global Collaborative Challenge, *paper presented at the 3rd International Digital Curation Conference*, Washington, D.C., December 11-13, 2007

Edwards, P., Jackson, S., Bowker, G. and Knobel, C. (2007) 'Understanding Infrastructure: Dynamics, Tensions, and Design' National Science Foundation, available at: http://hdl.handle.net/2027.42/49353

Frey, J., Hale, J., Milsted, A., Wilson, S., and Neylon, C. (2008) 'The Laboratory Blog-Book: How a laboratory blog notebook has developed to support, and in turn has been

influenced by, experimental laboratory practice' Fourth International Conference on e-Social Science University of Manchester, June 18th-20th, 2008 Available at: http://www.ncess.ac.uk/events/conference/programme/workshop1?ref=/programme/fri/4cfrey.htm

Galison, P. (1997) *Image and Logic: A Material Culture of Microphysics.* Chicago: University of Chicago Press.

Gardner D, Toga A, Ascoli G. et al (2003) 'Towards Effective and Rewarding Data Sharing' *Neuroinformatics* 1(3):289-95

Geddes, J., Mackay, C., Lloyd, S., Simpson, A., Power, D., Russell, D., Katzarova, M., Rossor, M., Fox, N., Fletcher, J., Hill, D., McLeish, K., Hajnal, J. V., Lawrie, S., Job, D., McIntosh, A., Wardlaw, J., Sandercock, P., Palmer, J., Perry, D., Procter, R., Ure, J., Bath, P., and Watson, G. (2006). 'The Challenges of Developing a Collaborative Data and Compute Grid for Neurosciences'. In *Proceedings of the 19th IEEE Symposium on Computer-Based Medical Systems (June 22 - 23, 2006)*. CBMS. IEEE Computer Society, Washington, DC, 81-86. DOI= http://dx.doi.org/10.1109/CBMS.2006.156

Gorman, M. (2002) 'Levels of Expertise and Trading Zones: A Framework for Multidisciplinary Collaboration' Social Studies of Science, Vol. 32, No. 5/6. pp. 933-938.

Higgins, S. (2008) 'The DCC Curation Lifecycle Model' *International Journal of Digital Curation* 3(1)

Hilder, W. (2005) Medical Research Council (MRC) 'IT Data Storage/Preservation Group Remit' available at: http://www.dcc.ac.uk/events/mdb-2005/hilder.php

Johnstone, E., Russell, K., Harrison, L, Lawrie, S. (2003) 'The Edinburgh High Risk Study: current status and future prospects' *World Psychiatry*. 2003 February; 2(1): 45–49

Karasti, H. (2001) 'Bridging Work Practice and System Design' *Computer Supported Cooperative Work* (10) pp. 211–246.

Kaufman, A, Cohen, D, and Yagel, R. (1993) 'Volume Graphics' IEEE Computer, 26(7) July 1993, pp. 51-64.

Keator D, Gadde S, Grethe J, Taylor D, Potkin S, FIRST BIRN. (2006) 'A general XML schema and SPM toolbox for storage of neuro-imaging results and anatomical labels' *Neuroinformatics*. 4(2), pp.199-212.

Kola, J., Harris, J. Lawrie, S., Rector, A. and Goble C. (2008) 'Towards an Ontology for Psychosis' Preprint submitted to Cognitive Systems Research

Kulynych, J. (2007) 'The Regulation of MR Neuroimaging Research: Disentangling the Gordian Knot' *American Journal of Law & Medicine* (33)3 pp.295-317

Lawrie, S. Weinberger, D., and Johnstone, E. (2005) *Schizophrenia: From Neuroimaging to Neuroscience* Oxford: Oxford University Press.

Lee, C., Dourish, P. and Mark, G. (2006) 'The Human Infrastructure of Cyberinfrastructure' *Proceedings CSCW'06* New York: ACM

Lowrance, W. (2006) *Access to Collections of Data and Materials for Health Research;*

A report to the Medical Research Council and the Wellcome Trust available at: www.wellcome.ac.uk/accessreport and www.mrc.ac.uk/research_collection_access

Lyon, L. (2007) *Dealing with Data: Roles, Rights, Responsibilities and Relationships*. Consultancy Report. University of Bath Technical Reports & Working Papers (UKOLN) Available at: http://hdl.handle.net/10247/412

MRC (2000) *Personal Information in Medical Research* available at: http://www.mrc.ac.uk/PolicyGuidance/index.htm

MRC (2007) *Policy and Guidance* available at: http://www.mrc.ac.uk/PolicyGuidance/index.htm

McChesney, S. Gallagher, S. (2004) 'Communication and co-ordination practices in software engineering projects' *Information and Software Technology* (46) 473–489

Mackenzie-Graham A, Van Horn J, Woods R, Crawford K, Toga A. (2008) 'Provenance in neuroimaging' *Neuroimage*. 42(1) pp. 178-95

Moorhead T, Harris J., Stanfield A., Job D., Best J., Johnstone E., Lawrie S. (2006) 'Automated Computation of the Gyrification Index in Prefrontal Lobes: Methods and Comparison with Manual implementation' *NeuroImage* 31(4) pp. 1560-6.

Neurogrid (2005) 'Work Programme' Available at: http://www.neurogrid.ac.uk/work-programme.htm

OECD Working Group on Neuroinformatics (2003) Neuroscience Data and Tool Sharing A Legal and Policy Framework for Neuroinformatics 1, pp. 149–166

Pekar, K. (2006) 'A Brief Introduction to Functional MRI' *IEEE Engineering in Biology and Medicine* March/April 2006, pp. 24-26

Poliakov A, Hertzenberg X, Moore E, Corina D, Ojemann G, Brinkley J. (2007) 'Unobtrusive Integration of Data Management with fMRI Analysis' *Neuroinformatics*. 5(1) pp. 3-10.

RIN (2008a) 'Research Data Principles and Guidelines' Research Information Network Available at: http://www.rin.ac.uk/data-principles

RIN (2008b) 'To Share or not to Share: Publication and Quality Assurance of Research Data Outputs' Available at: http://www.rin.ac.uk/data-publication

Rodriguez, D. Carpenter, T. van Hemert, J and Wardlaw, J. (2008) 'E-Infrastructure for Data Sharing in the SINAPSE Project' Proceedings e-science All Hands Meeting 2008, available at: http://www.allhands.org.uk/programme/index.html

SERCO (2008) 'UKRDS Feasibility Study Interim Report' Available at: http://www.ukrds.ac.uk/

Star, S.L. and Ruhleder, K. (1996) 'Steps Toward an Ecology of Infrastructure: Design and Access for Large Information Spaces' *Information Systems Research*, 7(1), 111-134.

Toga, A. (2002) 'Neuroimage Databases: The Good, the Bad, and the Ugly' *Nature reviews. Neuroscience*. 3(4): 302-9.

Treloar, A., Groenewegen, D. and Harboe-Ree, C., (2007) 'The Data Curation Continuum: managing data objects in institutional repositories', Dlib, September/October 2007. Available at: http://www.dlib.org/dlib/september07/treloar/09treloar.html

Treloar, A. and Harboe-Ree, C. (2008). "Data management and the curation continuum: how the Monash experience is informing repository relationships". Proceedings of VALA 2008, Melbourne, February (in press).

Ure, J. et al (2007) *Data Integration in eHealth*: A Domain/Disease Specific Roadmap in: Proceedings of HealthGrid 2007, Geneva: IOS Press

Van Horn, J., Grethe, J., Kostelec, P; Woodward, J; Aslam, J., Rus, D., Rockmore, D.; and Gazzaniga, M. (2001) 'The Functional Magnetic Resonance Imaging Data Center (fMRIDC): The Challenges and Rewards of Large-Scale Databasing of Neuroimaging Studies' *Philosophical Transactions: Biological Sciences*, Vol. 356, No. 1412, (Aug. 29, 2001), pp. 1323-1339.

Van Horn, J., Grafton, S., Rockmore, D., and Gazzaniga, M. (2004). 'Sharing neuroimaging studies of human cognition' *Nature Neuroscience*, 7(5), 473–481.

Wellcome Trust (2007) *Policy on Data Management and Sharing* available at: http://www.wellcome.ac.uk/About-us/Policy/Policy-and-position statements/WTX035043.htm

Weick, K. and Roberts, K. (1993) 'Collective mind in organisations: Heedful interrelating on flight decks'. *Administrative Science Quarterly* 38 (3), pp. 357-382

Wenger, E. (1998) *Communities of practice: Learning, meaning, and identity*. Cambridge University Press: Cambridge, MA.

Whyte, A. Job, D. Giles, S. and Lawrie, S. (2008) 'Meeting Curation Challenges in a Neuroimaging Group' International Journal of Digital Curation, 3(1), Available at: http://www.ijdc.net/ijdc/article/view/74

Whyte, A. (2008) *Neuroimaging Data Landscapes* SCARP Case Study Report 3 Literature Review and Appendices. Digital Curation Centre, Available at: http://www.dcc.ac.uk/scarp

Yin, R. K. *Case Study Research, Design and Methods*, (2003) 3rd ed. Sage Publications: Newbury Park CA (US)