Curating Brain Images in a Psychiatric Research Group

SCARP Case Study No.1

Summary and Recommendations

Angus Whyte

Digital Curation Centre, University of Edinburgh

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 eScience 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 the report[1] *Curating Brain Images in a Psychiatric Research Group* introduces four study 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*, 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.

[1] ***The report and annex are available at http://www.dcc.ac.uk/scarp***

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' [2] 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 eScience 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.

[2] 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.