

# The International Journal of Digital Curation

Issue 1, Volume 3 | 2008

## Meeting Curation Challenges in a Neuroimaging Group

Angus Whyte

Digital Curation Centre,  
University of Edinburgh

Dominic Job, Stephen Giles, Stephen Lawrie,

Division of Psychiatry, School of Molecular and Clinical Medicine,  
University of Edinburgh

July 2008

### Summary

The SCARP project is a series of short studies with two aims; firstly to discover more about disciplinary approaches and attitudes to digital curation through ‘immersion’ in selected cases; secondly to apply known good practice, and where possible, identify new lessons from practice in the selected discipline areas. The study summarised here is of the Neuroimaging Group in the University of Edinburgh’s Division of Psychiatry, which plays a leading role in eScience collaborations to improve the infrastructure for neuroimaging data integration and reuse. The Group also aims to address growing data storage and curation needs, given the capabilities afforded by new infrastructure.

The study briefly reviews the policy context and current challenges to data integration and sharing in the neuroimaging field. It then describes how curation and preservation risks and opportunities for change were identified throughout the curation lifecycle; and their context appreciated through field study in the research site. The results are consistent with studies of neuroimaging eInfrastructure that emphasise the role of local data sharing and reuse practices. These sustain mutual awareness of datasets and experimental protocols through sharing peer to peer, and among senior researchers and students, enabling continuity in research and flexibility in project work. This “human infrastructure” is taken into account in considering next steps for curation and preservation of the Group’s datasets and a phased approach to supporting data documentation.



## Introduction: SCARP Themes and Approach

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:

*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.

*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.

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

*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 preservation 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. The broader lessons are summarised in Conclusions to this article, which begins with an overview of neuroimaging in psychiatry. Then challenges and risks identified in the Neuroimaging Group study are described, with mitigation steps acknowledging the role of ‘human infrastructure’ in sharing knowledge between researchers of different skill levels and specialisms.



## 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 introduced here (Whyte, [in press](#)) 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 eScience 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](#)). 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](#)).

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.



### ***Challenges in Data Integration: Wider and Deeper Studies***

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 eScience projects such as the UK-based Neurogrid and NeuroPsygrid and U.S.-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](#)).

A number of factors however confound image and other 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](#)).

### ***Data Sharing Resources and Risks***

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 U.S. 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, [2001](#)). 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](#)). Levels of access are therefore highly variable, for example PsyGrid limits it to approved collaborators using a role-based model (Ainsworth et al., [2007](#)).

Access limitations are characteristic of medical domains, for example, Lowrance ([2006](#)) notes “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.

### **Challenges and Risks from the Lab Perspective**

The SCARP research site was a single neuroimaging centre in contrast to recent studies of neuroimaging, which adopt eScience collaborations as their research site (Ure et al., [2007](#); Lee, Dourish, & Mark, [2006](#)). The latter’s study of fBIRN however concludes that viewing collaborations as virtual organisations or “disembodied” infrastructure disregards the local alignments needed to make them work, i.e. they can better be understood as ways to blend *local* concerns, organizational relationships and arrangements, including those for access to data. So, although locally focused, the SCARP study spanned Neuroimaging Group researchers’ activities in both local and wider collaborations, including Neurogrid and NeuroPsygrid.

At the study’s outset, interviews with a cross-section of the group’s researchers identified that curation was seen in terms of managing the groups ever increasing needs for secure storage, and integrating local datasets. The study then facilitated a preservation risk assessment to understand the background to these issues, and as a means to identify a ‘way forward’ to address curation and risks to data preservation.

#### ***Assessing Curation and Preservation Risks***

Risk assessment used the DRAMBORA methodology (DCC/DPE, [2007](#)). Although intended for larger and formally established data archives or institutional repositories, it was used prospectively here to consider the range of activities that a data archive could entail, given that the UK has no established archiving service in this domain. The DRAMBORA approach has three main stages (DCC/DPE, [2007](#)): *firstly* the organizational context is characterized in terms of formal mandates and objectives, policy influences and community best practices. This stage identified a draft statement of curation and preservation objectives. The activities currently undertaken to pursue the latter were also identified with the digital assets considered of value.

The *second* stage involved standard risk assessment steps, identifying risks with the relevant activities and digital assets, then assessing the probability and impact of



each risk. Interviews and a questionnaire for researchers were used here. Probability was assessed as the likelihood of the risk event in a given period. Impact was rated in terms of loss of dataset usability and value. The *third* stage identified how risks are mitigated and possible additional measures.

Neuroimaging Group's main digital assets are its datasets, valued for their large number of subjects, time span, and range of associated data. Methods are also valued as they provide new ways to analyse this data. Other assets comprise the local technology infrastructure, where high performance parallel computing enables analyses that would otherwise be infeasible. The systems administration role manages many of these assets. As one would expect of an active research group, much curation and preservation activity is embedded in other research roles, particularly Principal Investigators who as custodians are responsible for clinical data management and security.

Despite these differences from an established data archive it was helpful to use the OAIS functional model (Consultative Committee for Space Data Systems [CCSDS], [2002](#)) to map current activities to the seven main functions in that model as a basis for envisaging more formalised procedures. The scope of current activities and relevant risks were identified from interviews and from risks previously identified by repositories using the DRAMBORA process, and the candidate list rated for possibility and impact. This provided a risk register, reusable by the Group to monitor risks periodically. The outcomes were then elaborated as recommendations for changes to the Group's data policy to support data documentation and preservation, to which we return below.

### ***New Challenges to Preservation from Innovations in Integration***

The Curation Lifecycle Model (DCC, [2008](#)) was used to consider new measures to mitigate risks to data, highlighting the curation steps where a fresh focus would be advantageous. Much of the Group's current work on infrastructure targets the data integration issues identified earlier. It has also widened the range of data it collects and developed new forms of image analysis.

The lifecycle model helps to draw attention to the effect of measures emerging from the eScience infrastructure for dataset integration, i.e. that these address risks in some steps, while adding to curation requirements in others. In Neurogrid the Group's "skull stripping" scripts automatically remove the identifiable faces and ears, enabling images to be shared with collaborators, and the project also provides scripts as grid services, enabling remote access to analyses. Neuroimaging Group also contributes to scanner inhomogeneity correction, enabling data to be pooled from different MRI scanners. Also, in NeuroPsyGrid, ontologies of psychosis symptoms are being developed to bridge the assessment scales used in different centres and over different periods of data collection.

These developments each *add value* to datasets by enabling new, shared and more reliable analyses. Each also however implies new *provenance* metadata requirements; e.g. details of who stripped which images and when, details of who used what grid services, and of which assessment scales were originally used by which centre and for what purpose. Provenance metadata are needed to ensure the accuracy, reproducibility and reusability of results (MacKenzie-Graham, Van Horn, Woods, Crawford, & Toga,

2008) and is broader than that needed and currently gathered for solely local use. Also, since these innovations entail new forms of derived image data, *storage requirements* are greater. The Group uses a variety of image analysis software packages and also developed its own techniques to automate identification of anatomical changes. This has increased not only the number of ways any one scan can be processed but also the capacity to process large numbers of scans simultaneously; again adding to storage demands. Indeed, storage is already at a premium given the increasing proportion of studies that use functional imaging, which produces large volumes of images.

The curation lifecycle steps where these measures address significant challenges are the later ones; dataset access and use, and the transforming of datasets for new purposes. Storage requirements continue to grow as each of the earlier steps need additional measures (see Figure 1 and further details in DCC, 2008). The Neuroimaging Group and other labs hold most of their data locally and in server file store rather than databases integrating image and associated clinical data. Decisions on dataset appraisal are complicated by the additional value that new techniques and data (for example genetic data) provide for retrospective analysis of old datasets. These largely remain in online storage, for which the Group has developed innovative backup solutions to minimise file recovery time and the possibility of data loss.

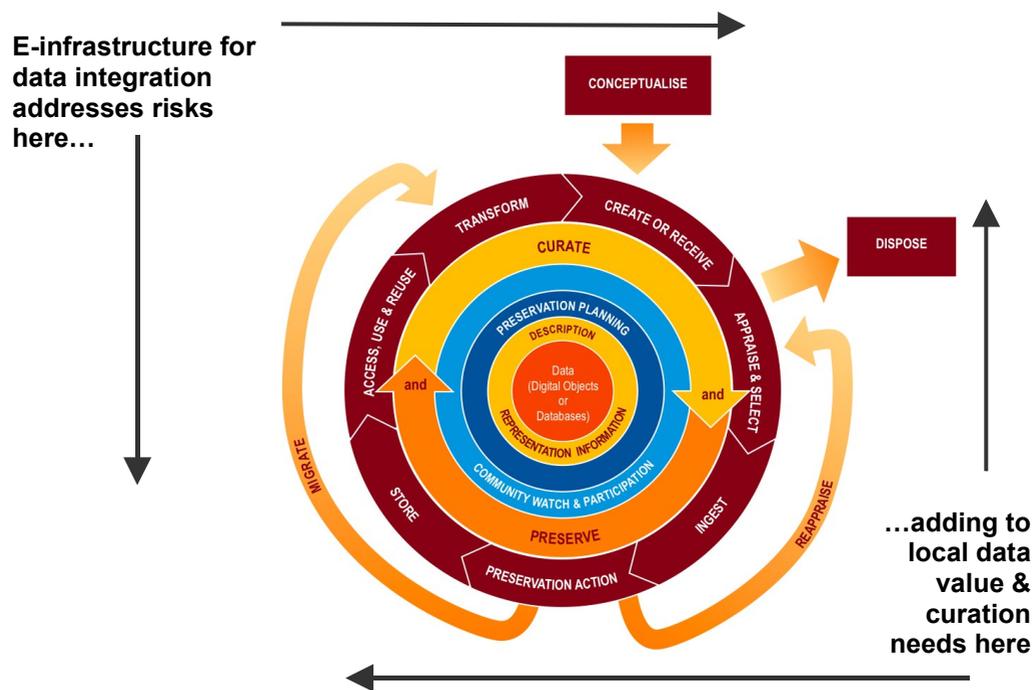


Figure 1. Measures to improve digital curation across the lifecycle.

The areas this study found most in need of additional resources were in preservation planning and action, i.e. to assign metadata, and to appraise datasets and migrate them accordingly to cheaper storage or disposal. As a first step the Group is defining a locally relevant “core schema” and standard set of study-related files; taking



account of the neuroimaging community's development of various metadata schema for provenance and study context. Progress towards more comprehensive data publication requires investment in an evolutionary approach to move from inter-personal sharing of data documentation, using the core schema as a basis for more structured collaboration support (see also Treloar & Harbroe-Ree, [2008](#)). Additionally, to address risks identified at the "create/ receive" step of the lifecycle, quality assurance is being strengthened and locally standardised.

### ***Human Infrastructure: Curation as Learning***

While there was support among the Group for measures to standardise documentation, they reported very low incidence of major data loss and the identified risks were mainly seen as having a low probability. The most valued resource for resolving any issues of understanding unfamiliar datasets was not a set of formal procedures but rather the Group's informal weekly research meetings. Observations of these meetings during the study sought to understand their role in research practice as a form of "human infrastructure", the importance of which has been acknowledged in US studies, including in the neuroimaging domain (Lee et al., [2006](#)). These drew on Star and Ruhleder ([1996](#))'s influential analysis of infrastructure, emphasising that this is learned as a part of membership, and both shapes, and is shaped by, the conventions of a community of practice.

Group meetings involve senior and junior researchers in an informal form of peer review, in which data and interim results are presented, carefully and constructively critiqued, and problems addressed through "heedful interaction" (Weick & Roberts, [1993](#)). Group meetings aid mutual awareness of ongoing work, continuity of research strands, and enable senior researchers to recommend areas of collaboration. They complement the close inter-personal interaction between clinical and imaging researchers (largely from engineering and neuroscience backgrounds) and help junior researchers learn the field's interdisciplinary terminology. Their learning process is highly participatory, requiring students to contribute skills to others' projects, and to reuse datasets so they may gain sufficient experience to acquire their own data. The interdependencies set up a "chain of learning" from newcomers to experienced researchers, which involves sharing experimental protocols and notes, but is nevertheless under strain as the Group expands. Data documentation directly benefits the learning process for new researchers, as well as contributing to the continuity and replicability of research. On that basis, postgraduate learning in this interdisciplinary domain may have a key role in curation.

## **Conclusions**

Multi-centre neuroimaging collaborations target the need for data integration and foster innovation in image analysis. This in turn adds to the need to record information about the context of studies and track the provenance of data that have been integrated from disparate sources and analysed by multiple people and/or centres. Innovations in analysis also place new demands on archiving, increasing the demand for online storage by making analysis of more images practicable, which in turn increases storage requirements for secondary data. New analysis techniques also highlight the need for active appraisal of datasets. They make retrospective analysis of neuroimaging datasets increasingly fruitful, while the timespan of longitudinal studies lengthens with the maturity of the field, and datasets are sustained through successive projects and custodians.



Standardisation in neuroimaging methods and data documentation is driven by the need for higher reliability in studies that also require larger-scale collaboration and hence wider trading of methods and data. The study demonstrates the need for a nuanced view of “enablers and barriers” to data sharing, curation, preservation and reuse. For example, the lack of standardisation in neuroimaging methods is a barrier to data sharing. However it also means that to learn methods and perform studies, lab researchers must share access to, and descriptions of, their data with others who have differing skills levels or specialities. Junior researchers learn by participating in colleagues’ studies, directly benefit from sharing experimental protocols, and could play an active role in building study documentation to serve research group needs.

The study illustrates that neuroimaging in the psychiatry domain involves continuous care of large and dynamic datasets. However investment in data documentation and development of integrated data management facilities at the lab level is required to mitigate preservation risks. Initial steps are being taken to identify a core metadata schema and group collaboration support technologies appropriate to a shift from inter-personal and study-level sharing of documentation to Group-level and wider data publishing.

### Acknowledgements

The study would not have been possible without the generosity and patience of colleagues in the Neuroimaging Group, Division of Psychiatry, University of Edinburgh. SCARP is funded by the JISC.

### References

- Ainsworth, J., Harper, R., Bridges, L., Whelan, P. Vance, W., & Buchan, I. (2007). The challenges of clinical e-science. *Proceedings e-Science All Hands Meeting 2007*. September 10-13, 2007, Nottingham, UK.
- Borgman, C.L. (2007). *Scholarship in the digital age: Information, infrastructure, and the Internet*. Cambridge, MA: MIT Press.
- Consultative Committee for Space Data Systems. (2002). *Reference model for an open archival information system (OAIS)*. Blue Book, January 2002. Retrieved 30 June, 2008, from: <http://public.ccsds.org/publications/archive/650x0b1.pdf>
- Digital Curation Centre. (2008). *The DCC curation lifecycle model* Retrieved 30 June, 2008, from <http://www.dcc.ac.uk/docs/publications/DCCLifecycle.pdf>
- Digital Curation Centre / Digital Preservation Europe. (2007). *Digital repository audit method based on risk assessment (DRAMBORA)*. Retrieved 30 June, 2008, from <http://www.repositoryaudit.eu/>
- Gardner et al. (2003). Towards effective and rewarding data sharing. *Neuroinformatics* 1(3), pp. 289-95.



- Geddes et al. (2006). The challenges of developing a collaborative data and compute grid for neurosciences. *Proceedings of the 19th IEEE Symposium on Computer-Based Medical Systems*.
- 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), pp. 45–49.
- Karasti, H., Baker, K.S. & Schledit, K. (2007). Digital data practices and the long term ecological research program. In *Third International Digital Curation Conference, December 11-13, 2007*. Washington, DC, USA.
- 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.
- 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., & 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*. Retrieved 30 June, 2008, from <http://www.wellcome.ac.uk/About-us/Publications/Books/Biomedical-ethics/WTX030843.htm>
- MacKenzie-Graham, A., Van Horn, J., Woods, R., Crawford, K., & Toga, A. (2008). Provenance in neuroimaging. *NeuroImage* 42, pp. 178-195.
- Medical Research Council. (2007). *Policy and guidance*. Retrieved 30 June, 2008, from <http://www.mrc.ac.uk/PolicyGuidance/>
- Mental Health Research Network. (2007). MRC MHRN cohorts database. Retrieved 30 June, 2008, from <http://www.mhrn.info/index/about/mrc-mhrn-cohorts-database.html>
- Pekar, K. (2006). A brief introduction to functional MRI. *IEEE Engineering in Biology and Medicine*. March/April 2006, pp. 24-26.



- Research Information Network. (2008). *To share or not to share: Publication and quality assurance of research data outputs*. London: RIN.
- Star, S.L., & Ruhleder, K. (1996). Steps toward an ecology of infrastructure: Design and access for large information spaces. *Information Systems Research* 7(1), pp. 111-134.
- Toga, A. (2002). Neuroimage databases: The good, the bad, and the ugly. *Nature Reviews Neuroscience* 3(4), pp. 302-9.
- Treloar, A. & Harboe-Ree, C. (2008). Data management and the curation continuum: How the Monash experience is informing repository relationships. *Proceedings of VALA 2008, Melbourne, February 2008*.
- Ure, J. Procter, R., Martone, M., Porteous, D., Lloyd, S. et al. (2007). Data integration in ehealth: A domain/disease specific roadmap. *Proceedings of HealthGrid 2007*. Geneva: IOS Press.
- Weick, K., & Roberts, K. (1993). Collective mind in organizations: Heedful interrelating on flight decks. *Administrative Science Quarterly* 38 (3), pp. 357-382.
- Whyte, A. (in press). *Curating brain images in a psychiatric research group: Infrastructure and preservation issues*. SCARP Case Study Report, Digital Curation Centre (forthcoming).