Collaborating to develop research data management services and collections

Martin Borchert
*Queensland University of Technology*, m.borchert@qut.edu.au

Stephanie Bradbury
*Queensland University of Technology*, s.bradbury@qut.edu.au

Philippa Broadley
*Queensland University of Technology*, p.broadley@qut.edu.au
Collaborating to develop research data management services and collections

Martin Borchert  
*Associate Director, Library Services (Information Resources and Research Support)*  
Queensland University of Technology Library  
m.borchert@qut.edu.au

Stephanie Bradbury  
*Research Support Librarian*  
Queensland University of Technology Library  
s.bradbury@qut.edu.au

Philippa Broadley  
*Project Manager*  
Queensland University of Technology  
p.broadley@qut.edu.au

**Acknowledgements**  
The Australian National Data Service (ANDS) has provided funding to the Queensland University of Technology to complete a number of research data management projects. The Queensland University of Technology (QUT) acknowledges and thanks ANDS for its support.

**Introduction**

Research data is rapidly becoming accepted as a valuable output of the research process, rather than as just an intermediate output consumed in the process of producing scholarly journal articles, conference papers, and books. The good management, sharing and reuse of research data can be used to verify research outputs, speed the research cycle, and increase research efficiency.

The Australian government has been funding the development of research data capability and capacity within the higher education and research sector funded through the National Collaborative Research Infrastructure Strategy (NCRIS) [http://ncris.innovation.gov.au/Pages/default.aspx](http://ncris.innovation.gov.au/Pages/default.aspx) which has in turn been distributed to institutions via competitive bidding processes since 2009.

The focus on research data management from eResearch groups, government agencies and library consortia has intensified in recent years and there is now much information and support available for libraries that wish to focus on developing competencies and services. The Australian National Data Service (ANDS) [http://www.ands.org.au/](http://www.ands.org.au/) has produced a growing compendium of information on good practice, standards and exemplars. The Association for Research Libraries (ARL) has the Transforming Research Libraries / Evolving E-Research / eScience program [http://www.arl.org/rtl/eresearch/escien/index.shtml](http://www.arl.org/rtl/eresearch/escien/index.shtml) and in Europe/the United Kingdom, The Joint Information Steering Committee (JISC) has the Digital Curation Research Centre [http://www.dcc.ac.uk/](http://www.dcc.ac.uk/).

The purpose of this paper is to document QUT’s story in developing and implementing a range of data management systems and services to date.
Identifying and managing QUT’s research data assets

Advocacy

Executive advocacy for the development of research data management services needs to be strong and comes from a range of senior positions and forums. It is important to keep data management on the research agenda by submitting project, service, infrastructure proposals, plans and reports to promote discussion. This is to ensure academic stakeholder engagement and to ensure that services meet researcher needs. Appendix 1 includes stakeholder roles and groups in research data management advocacy work at QUT.

Policy

The QUT Manual of Policies and Procedures includes a policy on the Management of Research Data http://www.mopp.qut.edu.au/D/D_02_08.jsp since 2009, which incorporates sections on definitions, roles and responsibilities and the management of data including reference to the Australian Code for the Responsible Conduct of Research (The Code), data planning, intellectual property, record keeping, storage, retention and disposal, privacy and confidentiality, and access and re-use. Reflecting The Code, the policy requires that research data is managed and also that data description is centralised while data storage is decentralised.

Staff development

Research data management is complex. It combines a wide range of information management, open access, infrastructure, research and research support process and research discipline specific aspects into a service. To cover all aspects equally, it helps to be an academic researcher, librarian, an IT software specialist, infrastructure specialist, a statistician, a researcher and a manager. It is impossible to imbue staff across a broad range of positions, levels, professions, and experience with a common perspective, understanding and level of confidence. Data management infrastructure has taken a long time to deliver and will continue to develop. It is especially difficult to train staff to use infrastructure before it is delivered and available for “getting your hands dirty”. Most library staff have a broader range of responsibilities, and even if a position is relatively specialised, like Research Support Librarian, there are a myriad of competing demands on time, capacity and capability.

The approach used at QUT has been discussed in some detail in Borchert and Callan (2011), but was basically to use the Research Support Team to develop the service(s) and training modules and resources, and have Liaison Librarians progress through a process of familiarity until they had gained the knowledge, skills and confidence in understanding the full complexity of research data management and were comfortable enough to be the trainers.

In 2009, a range of sessions was provided to Liaison Librarians starting with awareness raising sessions through to more specialised instructional supports including sessions on more specific and complex data management issues and practices. A variety of expert speakers from across the University presented different aspects of the research lifecycle and research data and the importance of and practices of its management. Through 2010, alongside research students and researchers, Liaison Librarians attended ‘Managing Research Data’ workshops presented by the Research Support Team. Through purposeful observation and their prior learning, the Liaison Librarians built on their knowledge and skills and were able to transition to co-presenting their own Managing Research Data sessions to researchers. They co-presented with a Research Support Team member or a more experienced Liaison Librarian, and then graduated to becoming the primary presenters. The result has been that most data management workshops are now presented by Liaison Librarians and no longer the research support specialist librarians.
It has been found that IT Helpdesk staff need to be familiar with services and infrastructure in order to provide first tier resolution or to refer appropriately to second tier. These staff should receive training about the application of research support infrastructure within a within a research support services framework.

Staff within the Division of Technology, Information and Learning Support have attended a number of eResearch Australasia Conferences in which data management features strongly, as well as participating in various development projects provided by ANDS, including workshops, webinars and meetings for the purpose of sharing information.

**Promotion, information and training**


Research data management workshops have been provided to researchers and Higher Degree Research (HDR) students on a monthly basis since 2010 and focus on good data management practice, data planning, storage options and the benefits of sharing. Eighty researchers attended workshops in 2010, 202 in 2011 and 204 in 2012. In addition, there has been an annual ‘Managing Your Research Data’ seminar featuring academics and HDR students speaking about data issues and successes. These have attracted 70, 50 and 30 people in 2010, 2011 and 2012 respectively. The IFN001: Advanced Information Research Skills (AIRS) unit is a 4 credit point, formal unit which is mandatory for all PhD and Masters Research students to complete. It is optional for Masters Coursework and Honours students. The AIRS curriculum includes a module on research data management. 1,596 students complete the unit in 2012. Additionally, Liaison Librarians and HPC support staff provide a research support consultation service which covers the full spectrum of services, and for which data related queries are very common. The Library recorded 750 research consultations in 2012. The IT Helpdesk also fields support queries about all aspects of the use of IT for research purposes. Common questions with answers have been entered into AskQUT, a customer relations management knowledgebase, so researchers can access help information at their point of need.

**Facilitating open access to institutional research data (data “out”)**

**Leveraging QUT ePrints**

Launched in 2003, QUT ePrints is an open access institutional repository of research outputs and has grown into a highly successful open access collection of 41,000 records and 25,000 fulltext deposits...
averaging over 200,000 fulltext downloads per month. A major contributing factor to the success of QUT ePrints has been the continued effort the QUT Library has invested into promoting and training academic staff of the benefits of putting their final author manuscript version on open access. This has resulted in increased accessibility and increased citation counts. With ten years of active participation, support for open access is strong and widely spread at QUT and this can be leveraged for the next wave of change in scholarly communication – open data. The benefits of open data to the author are not as easily defined as for open access to research publications (journal articles, conference papers, book chapters), and the infrastructure (described below) is being developed ahead of the expected wave of demand. However, the depth of general understanding and support for open access places the institution in a good position for when the infrastructure is ready and research policy environment drives demand for the service.

Linkages between QUT ePrints and the QUT Research Data Finder (metadata repository) will help increase visibility of datasets and will be key to ensuring academic engagement.

Infrastructure

QUT Research Data Finder and the ANDS Research Data Australia

In 2012/2013, ANDS funded QUT to develop a research data metadata repository, QUT Research Data Finder http://researchdatafinder.qut.edu.au/vivo/ . The software chosen was the open source VIVO/VITRO http://vivoweb.org/ which originated in Cornell University et al (USA). Major features of the repository include: having an open interface to sharable datasets and an authenticated QUT-only interface to all records (for sharable and non-sharable datasets) so the repository performs the function of a data registry for QUT; integration with the ANDS Research Data Australia http://researchdata.ands.org.au/ national repository of sharable datasets, for automated ingest to RDA; integration with ResearchMaster, the university’s research management database (source of research activity information; integration with the university’s Academic Profiles system http://staff.qut.edu.au/ (source of human resource information); and the aforementioned integration with QUT ePrints (related publication information).

The repository holds metadata records of the datasets and does not store the datasets themselves. This is in line with the QUT policy on Management of Research Data which states descriptions will be centralised into a registry while data storage will be distributed.

In 2010, ANDS funded QUT to develop the Metadata Hub which was a precursor to QUT Research Data Finder. It implemented the basics of VIVO to allow for the passing of metadata records to Research Data Australia.

In 2011, ANDS funded QUT to implement three Data Capture projects by which software was developed to interface with public performance, groundwater, and environmental acoustics sensing devices to capture metadata on the fly and stream this information to the Metadata Hub and Research Data Australia.

Identification, description and discovery

The repository uses the Registry Interchange Format – Collections and Services (RIF-CS) http://www.ands.org.au/resource/rif-cs.html which was developed and supported by ANDS. The purpose of the integration work is so that the repository pulls as much information as possible from existing university systems in order to minimise duplication of data entry and so make the service more attractive to academics and library staff alike.
The RIF-CS metadata schema supports the development and interlinking of a network of description records for each dataset – collection record (dataset), party record (people and groups), activity record (research project) and service record (software and devices). Collection records are mandatory to keep the focus on the datasets.

In 2012, ANDS funded QUT to implement the Gold Standard Record Exemplars project by which QUT investigated information sources and procedures for implementing the RIF-CS metadata schema to a high standard to produce the most content rich and connected data descriptions possible. Procedures need to be institution-specific, and QUT published its good practice procedures (https://projects.ands.org.au/id/SC37) to make them available to the research community and also hosted a webinar (http://youtu.be/rOL2vb--9G8) to introduce the manual and provide opportunities for questions and answers.

In 2010, ANDS funded QUT to implement the Seeding the Commons project which was designed to investigate existing datasets at QUT and also academic practices and attitudes to data management and data sharing. 300 researchers were interviewed and 200 datasets were described. Of these, only 12 have been made sharable. Researchers cite numerous reasons for not sharing data, and this is the topic of a prior QUT paper (Milne, Thomson and De Vince, 2011).

QUT’s 250+ dataset records have been mostly produced under the various ANDS projects. The operational side of the data description service has been trickling along for some time with three ‘early adopter’ Liaison Librarians trained in applying RIF-CS and using the QUT Research Data Finder system to date. The Office of Research provides the Library with “notification of intent” information consisting of lists of researchers working towards submitting research grant applications in future. This information acts as a flag to Library staff to contact these researchers to offer them the range of research support services, including data management and description. The operational aspect of the business of research data management has become the recent focus as we move out of project mode into mainstream service provision.

Data storage

The data storage story is multifaceted and ongoing with some aspects being institution-specific and others being deployed at a national level.

QUT has implemented EStore, a university networked storage system with backup that is available to all staff. Staff are allocated space and have the option to request more space. Access is via authentication. Another solution, informally named RStore for now, will be tailored for research data purposes and more suited to data sharing is being investigated. It is likely to be a replication of existing technology, but with a more open architecture and more data sharing-friendly policies allowing cross-faculty and external access to datasets. Staff are discouraged from using portable devices. Dropbox is undoubtedly popular, as is email for sharing smaller datasets.

Investments in Research Data Storage Infrastructure (RDSI) http://rdsi.uq.edu.au/ and National eResearch Collaboration Tools and Resources (NeCTAR) http://www.nectar.org.au/ at the national level are seeing the implementation of state-based research storage hubs and the development of discipline-specific virtual research environments. The Queensland Cyber Infrastructure Foundation (QCIF) http://www.qcif.edu.au/ is implementing a data storage service. It is likely that QUT will employ the service for storing and sharing large scale datasets and its internal infrastructure for the long tail of smaller datasets and datasets in an incomplete or draft state. A QUT-specific implementation of such infrastructure, but perhaps on a smaller scale, could be used to meet demand for the long tail of smaller scale datasets within QUT.
Demand for data storage will no doubt increase as research policy drives management, sharing and reuse. Future data management and data sharing mandates from research funding agencies such as the Australian Research Council (ARC) and the (National Health and Medical Research Council NHMRC) would no doubt provide considerable impetus for future development.

**Computation, analysis and visualisation**

The HPC team provides a range of computation, analysis and visualisation services to support QUT researchers to generate and use data in their research. The HPC computation service provides large scale computation power to provide large scale datasets. Team members provide advice and coding for optimisation of machine time. Advice is also provided for quantitative and data qualitative analysis using mathematical modelling (MATLAB, Mathematica, Maple) and statistical analysis, data mining tools and visualisation software such as Leximancer and Enterprise Mining (SPSS, SAS, NVIVO).

QUT Library’s role is to assist HPC to promote these services to new staff, early career researchers and HDR students in particular so appropriate data collections and methods can be integrated into research design processes.

**Facilitating institutional access to external research data (data “in”)**

Information resource management (collection development) is a traditional and core function of libraries and the provision of resources, whether online or in print, is central to the definition of libraries, as opposed to study hall or some other service provider. It is a responsibility which cannot be performed by any other department of a university (or at least as professionally). Researchers (and students) look to the library to provide the current and historical, comprehensive and high quality peer reviewed academic scholarly information resources to connect them with research and researchers in their discipline. This is of course in addition to other scholarly information networks including open access via the internet repositories and peer-to-peer networks for exchanging documents. At QUT, the Library’s website [http://www.library.qut.edu.au/](http://www.library.qut.edu.au/) and Database of Databases [http://www.library.qut.edu.au/find/](http://www.library.qut.edu.au/find/) are some of the most heavily used web pages in the University.

There is an obvious opportunity for libraries to leverage their expertise in collection development and the strength of our brand with researchers looking for information, to assist researchers with awareness and ease of identifying, accessing and using research datasets in research activities. Being supportive of open access, there exists an imperative to incorporate research datasets into scholarly information discovery services. Descriptions of datasets can be used to recommend key resources to researchers, either in the disciplinary context with ebook, ejournal and database subscriptions or as a collection of dataset services based on format.

Both open access and commercially available databases should be incorporated into the collections in order to try to provide a useful, if not comprehensive service.

**Incorporating open access research data into the library collections**

QUT Library is attempting to identify, describe and incorporate key open access research dataset services into existing information management systems and collections. Examples are provided in Appendix 2. It is important to focus energy into identifying and using open access resources in this way in order to support open access initiatives and to encourage researchers to use them. The service is likely more useful for HDR students and early career researchers as it is more so designed to bring these resources to awareness rather than provide a comprehensive catalogue. Once familiar, researchers can use search engines to discover dataset collection services in their discipline. Potentially, there is a lot more work to do in this space, depending how comprehensive you might want the catalogue of resources to be.
Incorporating commercially available research data into the library collections

The other side (to open access) is to identify, procure and provide access to commercially available datasets. There are two types described here – commercially available datasets accessible under similar business models to ejournal databases and other library resources; and datasets available under confidentialised or individual license arrangements. The latter is covered separately.

There is a growing range of commercially available research datasets in the information marketplace for academic libraries to consider for procurement for use by researchers. The increase is perhaps threefold: firstly because the eresearch or escience movement (combining research with IT and the internet) is transforming the way researchers do their work and thus generating both supply and demand; secondly, traditional publishers are starting to seek alternative business models, content and technology based products to broaden the scope of their business in the light of increased competition from open access sources; and thirdly, an increased awareness of academic libraries to the availability of non-bibliographic datasets which were once of interest to and the domain of only the corporate sector (economic and finance data for banking; company information for marketing; engineering data for construction; social data for government).

Since 2010, QUT Library has been placing some focus on liaising with faculties to identify demand for large scale commercially available research datasets which, by their nature, are unaffordable for individual researchers or even schools and faculties, thereby representing an opportunity for the Library to add value. The QUT Business School (faculty) is where most unmet demand was identified in the first instance, and the School of Accounting and the School of Economics and Finance in particular.

Over time, QUT Library and schools have worked together to identify, assess and select the high priority commercially available datasets to support research goals. The selection and assessment has been complex as the content and interface benefits of each dataset is not as immediately apparent to library staff as it is for journal databases. Many dataset providers rely on customers having considerable prior knowledge of the discipline and data types available. Academics naturally tend to focus on their own immediate needs and research focus rather than take the overview the library needs to in order to determine faculty-wide priorities, ensure content duplication between products is kept to a minimum, and also ensure that the University obtains the best price, most relevant content and maximum return on investment. It can also be difficult for the library to determine interface and functional requirements without considerable subject knowledge and specific knowledge of each researcher’s projects. Also, there are often competing faculty demands to navigate. A very useful process was to request the assistance of the Heads of Schools to sponsor and “adjudicate” the prioritisation of the numerous datasets to create a list of achievable and affordable goals.

It has taken QUT three years and close to a million dollars to develop and provide the following range of research datasets for the QUT Business School. Readers of this article will be able to access the dataset descriptions http://www.library.qut.edu.au/services/datasets/, but not the dataset themselves unless your home institution provides access.

Some products have been licensed or limited in access for staff and postgraduate students (including HDR students) only. This is either because of cost or to manage the risks associated with license conditions. The datasets which are available to undergraduate and coursework students are often integrated into the curriculum for assignment-based work, and may assist undergraduate students to transition to higher degree studies. At all levels, the availability of these datasets creates a real world experience for developing professionals. Most of these products provide some usage statistics information which is used for collection development purposes.
Some researchers in the QUT Business School have reported that the availability of these resources has transformed their opportunities to conduct research. They also report that the availability of these resources is helping to attract postgraduate and HDR students and post doctorates locally, nationally and internationally. It is possible that a wave of undergraduate students will transition into HDR studies on the basis of research using these datasets.

The library as research data broker

QUT Library implemented the External Research Data Service http://www.library.qut.edu.au/services/datasets/ in mid 2010 as one of the new services to come in under the Library’s Research Support Strategy. The purpose of the External Research Data Service is to facilitate researcher and HDR student access to confidential and secure datasets to support research activities. Discussions with faculties in 2010 indicated an existing demand and distributed procurement activities for specific external datasets which require licensing and a nominal payment. Some of the more commonly required datasets were being procured by numerous individual academics within a faculty each year, resulting in duplicated payment, duplicated contracts and inefficient practices. The Library consulted with researchers on the datasets most in demand as to which datasets the Library would procure and license on behalf of the University in order to “seed” the service.

The process of procuring a new dataset commences when the Library discusses the data requirements with the researcher, then contacts the agent (usually a government department or university), obtains the contract or license and uses the Library Resource Allocation to pay any purchase or subscription fee required. The Library acts as a broker between the researcher and the agency, and will provide the service whether one or a number of researchers require access to a specific dataset (according to licencing conditions). Because of the confidential nature of the data provided in some of these datasets, many agents require that the QUT researcher sign a terms of use agreement in addition to the library signing the institutional license. In many cases, when an HDR student requires access, the principal supervisor is also required to sign the terms of use agreement. Depending on the distribution and access model, the Library either hosts the dataset on a central secure server, or distributes a copy on disc or provides access to the dataset on a secure computer housed in the Research Data Lab, which is a bookable room available for accessing these datasets.

To promote and support the service, the Library provides an External Research Data Service web site http://www.library.qut.edu.au/services/datasets/ which lists and provides information about datasets available and contacts. Datasets which have been procured for a single researcher are not listed. The service was developed by the Library’s Research Support Team, and services are also provided by one of the Business Liaison Librarians, reflecting the source of much of the demand for this service. Twenty data access requests were satisfied in 2010, 70 in 2011 and 71 in 2012. The most popular datasets are Household, Income and Labour Dynamics in Australia Survey (HILDA), Longitudinal Study of Australian Children (LSAC) and ABS Confidentialised Unit Record Files (CURFs) Microdata.

Benefits of providing the External Research Data Service include increased awareness; easier and faster access to datasets for researchers; increased confidence in contractual and compliance matters for the researcher and the library; and increased return on investment for the University. It is estimated that the University has likely entered into 100 fewer contracts and payments because of the centralised service.

While it would be preferable to be able to access all datasets via open access, the datasets provided in this service are largely restricted due to holding confidential and sensitive information. The range of datasets offered under this service is likely to grow. This is a somewhat time consuming service for
academic libraries to offer, using core library capabilities, yet provides an opportunity for “an easy win” early in the development and delivery of otherwise more complex research data management services.

Conclusion

Data management is a complex business involving many stakeholders, a range of new skills, infrastructure at various levels and above all a desire to collaborate externally to library and organisational boundaries. There are many opportunities for academic libraries to lead and be involved. The development of a data management culture, skills and systems takes years, even with the injection of substantial project funding like QUT has enjoyed from ANDS. It is hoped that QUT Library’s unfolding data management story will prove of interest to other libraries beginning or already on this road and that it might inspire or assist with some ideas for the pathway other academic libraries may wish to take. Academic demand for data management services is likely to increase as research funding agencies demand a greater extent of good data management practice, openness, sharing and reuse. The authors are most happy to be contacted and to provide more information and documentation, or to liaise directly via a visit to QUT or videoconference.
References


Joint Information Steering Committee (JISC), *Digital Curation Research Centre*. Viewed 7 January 2013. Available at: [http://www.dcc.ac.uk/](http://www.dcc.ac.uk/)


Appendices

Appendix 1 – Stakeholder roles and groups in advocacy of research data management at QUT

- Deputy Vice Chancellor (Research)
- Deputy Vice Chancellor (Technology, Information and Learning Support)
- Director, Research Collaborations (eResearch, Science and Engineering Faculty)
- Director, Library Services
- Director, Information Technology Services
- Director, Office of Research
- Associate Director, Library Services (Information Resources and Research Support)
- Associate Director, ITS (High Performance Computing and Research)
- Associate Director, ITS (Infrastructure Services)
- University Research and Innovation Committee
- Faculty Research and Innovation Committees
- eResearch Working Party
- Research Support Committee

Appendix 2 – Incorporating open access datasets into collections (examples)

- Within the database service http://www.library.qut.edu.au/find/databases/subjectguide.jsp?su=001019000
- In the QUT Library Catalogue http://libcat.library.qut.edu.au/search~S8?/X%22research+data+archives%22
  In subject guides e.g.:
  - Accounting http://libguides.library.qut.edu.au/accounting
  - Banking and finance http://libguides.library.qut.edu.au/bankingfinance
  - Company information http://libguides.library.qut.edu.au/company
  - Data mining http://libguides.library.qut.edu.au/datamining
  - Datasets for research in social science http://libguides.library.qut.edu.au/DatasetsForResearch

Appendix 3 – Incorporating commercially available datasets into collections (examples)

- Data Citation Index (Thomson Reuters) http://www.library.qut.edu.au/find/databases/fullrecord.jsp?id=7058
- EIU CountryData (Bureau Van Dijk) http://www.library.qut.edu.au/find/databases/fullrecord.jsp?id=2140
- Orbis via Bureau Van Dijk http://www.library.qut.edu.au/find/databases/fullrecord.jsp?id=5435
- Osiris via Bureau Van Dijk http://www.library.qut.edu.au/find/databases/fullrecord.jsp?id=5434
- Passport GMID http://www.library.qut.edu.au/find/databases/fullrecord.jsp?id=5336