Investigating Data Management Practices in Australian Universities

Margaret Henty, The Australian National University

Belinda Weaver, The University of Queensland

Stephanie Bradbury, Queensland University of Technology

Simon Porter, The University of Melbourne

APSR
AUSTRALIAN PARTNERSHIP FOR
SUSTAINABLE REPOSITORIES

http://www.apsr.edu.au/investigating_data_management

July, 2008

Table of Contents

This survey was one of the activities of the APSR ORCA Support Network in 2007.

Documentation of the survey and completion of the report have taken place during the first half of 2008, with the support of APSR, the ANDS Establishment Project, The University of Queensland, The University of Melbourne and the Queensland University of Technology.

The Australian Partnership for Sustainable Repositories (APSR), was established in 2003 as a broad coalition of institutions promoting excellence in digital research collections for the higher education sector. APSR was supported by the Systemic Infrastructure Initiative as part of the Commonwealth Government's Backing Australia's Ability - An Innovative Action Plan for the Future.

The ANDS Establishment Project has been funded by the Department of Innovation, Industry, Science and Research (DIISR) through the National Collaborative Research Infrastructure Strategy (NCRIS).

An digital version of this document is located at:
http://www.apsr.edu.au/investigating_data_management

For more information about the report please contact Margaret Henty
margaret.henty@anu.edu.au

July, 2008

Introduction

Universities around the world are experiencing an increasing emphasis on the need for effective data management and stewardship to underpin the changing research environment, as research becomes more dependent on data in digital form and computers and networks proliferate. Data is valuable from the moment of creation, not to mention expensive to collect, so there is no point in duplicating its collection. It might also be unique, representing a snapshot in time or space and therefore impossible to replicate. Data can be re-used, sometimes for purposes not originally dreamt of, and it can be re-analysed, either to check original results or to take advantage of new analytical techniques. There is increasing pressure to ensure that data should not go to waste, and for universities to develop the infrastructure needed to care for this invaluable resource.

About the survey

It was in this environment of changing expectations for the provision of data management infrastructure that three Australian universities decided to investigate the needs of their own communities. The initiative came from The University of Queensland (UQ) and was taken up by The University of Melbourne (UM) and the Queensland University of Technology (QUT. All three are universities with an increasing focus on eResearch, and a desire to improve their support infrastructure. All three recognised that a first step towards this goal was to find out more about the current practices and training requirements of their research staff.

UQ does not have an online survey facility, so access to The Australian National University's Apollo facility was organized through the Australian Partnership for Sustainable Repositories (APSR) of which UQ is a partner. The original questionnaire was then adapted slightly to meet the individual needs of the other two Universities. A full list of questions is included in Appendix A.

All three surveys were run in late 2007. In each case, emails were sent from the office of the senior academic administrator responsible for research (or eResearch) to all academic staff and postgraduate students at the university concerned, explaining the need for the survey and seeking their cooperation. Reminder emails were sent as follow up. The response rate in each case exceeded expectations, with a total of 879 responses being received in total across the three universities. From a statistical perspective, this cannot be seen as a strictly random sample, so it is preferable to interpret the results in terms of general trends, rather than as a precise representation of practices and viewpoints in the three universities.

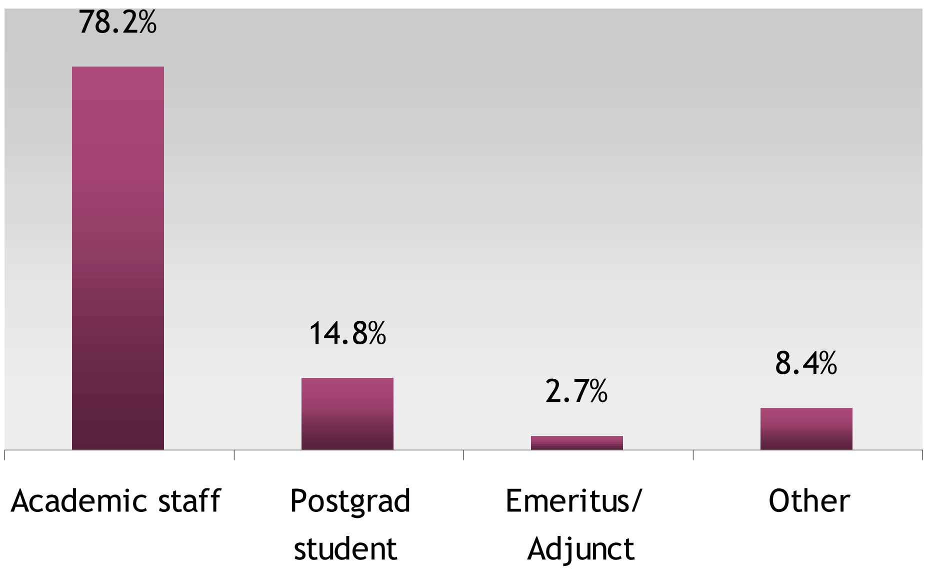
About this report

This report is presented in several parts. A full set of aggregated results is presented first, followed by three smaller sections describing the survey from the point of view of each of the three participating universities.

Comments have been edited to ensure anonymity of respondents and universities. Original spelling and grammar have been maintained. A complete set of the data files can be found at http://dspace.anu.edu.au/handle/1885/46634.

The respondents

Figure 1: Respondent status



Respondents were asked to tell us their academic status. As can be seen from Figure 1, over three-quarters described themselves as academic staff. The second largest group was postgraduate students. Emeritus/adjunct appointments were only a very small proportion, and less than 10% described themselves as “other”. The “others” were made up of research assistants, laboratory managers, project managers, post-doctoral fellows, data analysts, external contractors, specialist technologists, hospital staff and members of project teams brought in from private industry. Some respondents put themselves into more than one category.

The survey results

The single and most outstanding finding of the survey is the similarity of the pattern of responses between the three institutions. For this reason it is possible to aggregate the results for the purposes of presentation and discussion.

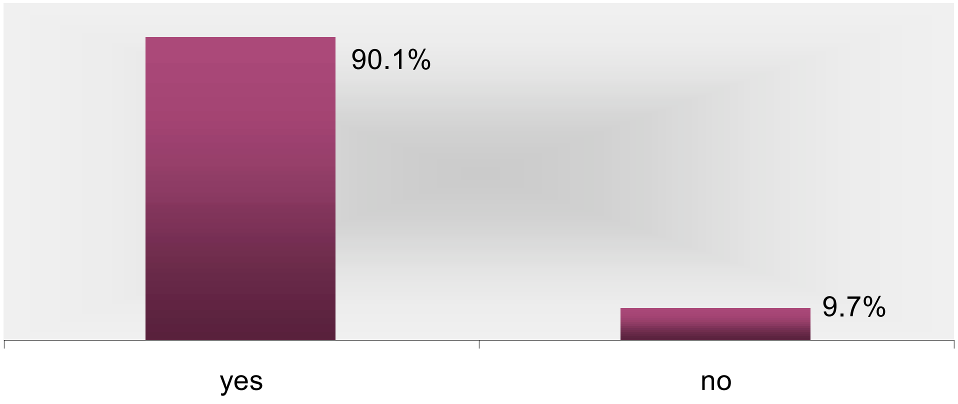
In the discussion which follows, the responses to each of the questions are shown as a table of the aggregated responses of the three institutions. Those questions asked at only one of the universities are discussed separately. The questions are not numbered, as the numbering of each varied among the three variations of the questionnaire (one for each university). Comments from respondents have been edited in such a way that their institutions cannot be recognized.

The survey results in some cases are divided by disciplinary affiliation. Respondents were not asked to identify their discipline directly, so this has been extrapolated from their departmental, faculty or other organizational affiliation. The results should therefore be read with this in mind. The disciplinary areas identified were: Social Science, Medicine & Health, Business & Economics, IT, Engineering & Architecture, Science, Humanities & Creative Arts, and Law.

Tables and Comments

Digital data

Figure 2: Has your research generated digital data?



Over 90% of respondents reported that their research generates digital data with less than 10% saying that their research does not generate digital data. It could be seen as surprising that as many as 10% say that they do not generate digital data, as it is hard to imagine in the current environment that there would be any research which does not involve at the very least the digital generation of text. Perhaps what we are seeing here is a perceived divide between data and text, with some not recognising digital text as data.

Non-digital data forms

Table 1: If no (digital data), do you maintain research-related data in non-digital forms such as paper, photographs, video or audio tapes, slides, etc?

	n	Per cent
yes	167	19.0%
no	32	3.6%
no response	680	77.4%
Total	879	

The survey asked what kinds of non-digital data was maintained, to get some estimate of what other kinds of research materials are being generated. These might at some future time need to be digitised or otherwise take care of.

The question was, however, flawed, in that it asked for a response only from those who had no digital data. The responses reflected the flaw, and included many indignant comments that research projects tend to generate both digital and non-digital data. Many more responded to this question than had responded to the previous question (that they had no digital data) in order to emphasise the point.

A wide variety of non-digital formats were mentioned in the comments: survey and evaluation forms, laboratory notes, client files, photographs, cardboard, plastic and timber models, drawings, audio tapes, radioactivity data in printed form, jewellery and clothing, rocks and shells, draft manuscripts. Some of these can potentially be digitised; some not. More importantly, some of these (such as survey forms, client files or laboratory notes) could be collected digitally to start off with, removing any need for later digitisation or storage.

Not everyone was happy with the question: “I don't 'manage' or 'maintain' my research. In fact, I find 'manage' an offensive term in this context. I write articles etc. These appear in books and journals, and in web sources. I understand this as publication and not as management or maintenance.”

Types of digital data

If universities are to develop better data management infrastructure, we need to know what kinds of digital data are being generated. Table 2 sets out the results.

Table 2: If your research generates digital data, please check all the following types that apply:

	n	Per cent
spreadsheets or databases	595	67.7%
documents and reports	558	63.5%
data automatically generated from or by computer programs	430	48.9%
experimental data	378	43.0%
email	357	40.6%
data collected from sensors or instruments	332	37.8%
images, scans or X-rays	323	36.7%
fieldwork data	286	32.5%
digital audio or video files	224	25.5%
web sites	209	23.8%
laboratory notes	185	21.0%
blogs or discussion threads	74	8.4%
other (please specify)	36	4.1%
Total respondents *	879	

* some respondents clicked more than one box so does not equal total responses

Spreadsheets and databases are the most common, with two-thirds of respondents having them. Slightly fewer have documents and reports, and just less that one half have data automatically generated from or by computer programs. About forty per cent have experimental data and email, with diminishing numbers reporting data collected from sensors or instruments, images, scans or X-rays, fieldwork data, digital audio or video files, web sites, laboratory notes, and blogs or discussion threads. Few researchers generated only one type of data.

One person seemed amused at the question: "You must be kidding - everyone has the above!"

Other responses included a wide variety of data types: "bibliographies, biographies and other textual elements," online surveys, secondary data analysis, questionnaires, bibliographic databases, mathematical models, simulations, interview transcripts, computer programs, satellite imagery, GIS data, CAD models, genotyping and sequencing data, electronic health records, music scores, podcasts, laser scanning imagery, GPS measurements, mind maps, flow cytometry data and spectral data , and "data in the form of CFD [computational fluid dynamic] codes containing specific models for turbulence, chemistry and the like."

Size of data collection

Table 3: How large (in total) is your digital research data?

	n	Per cent
less than 100MB	100	11.4%
100MB - 1GB	197	22.4%
1GB - 1TB	327	37.2%
More than 1TB	43	4.9%
Don't know	175	19.9%
No response	37	4.2%
Total	879	100.0%

Repository managers and data curators are interested in knowing how large data collections are in order to assess likely storage needs. Researchers, on the other hand, do not necessarily think in the same terms, unless the data sets are large and have known storage requirements. Table 3 shows that about one quarter of respondents either do not know how large their data is or did not respond to the question.

About one-third of respondents have less than 1GB of data and a similar proportion between 1GB and 1TB. Less than five per cent, a comparatively small proportion, reported that they have a larger amount of data, over 1TB. There were many qualifications to these figures in the comments, with estimates provided in terms of the number of CD-ROMs or DVDs held, or the number of pages of text, or the number of video films or segments. Others commented that their collections are growing, or that they have not started collecting yet.

Software used for analysis or manipulation

Table 4: Software used by more than 10 respondents

	n
SPSS	291
Excel	277
stata	63
Matlab	61
NVIVO	55
Adobe Photoshop	52
Access	46
SAS	39
SigmaPlot	37
Minitab	28
MS Word, Graphpad Prism	22
R	21
FileMakerPro	14
Eviews, ImageJ	13
ArcGIS, Gaussian, Labview	11

The use of different software for data analysis and manipulation can have an impact on data management and curation. The answers to a question about software use demonstrate what a remarkable range of software is in use. Some is proprietary and well known, some is open source and some is being developed in-house for specific purposes.

Table 4 shows the most popular 19 proprietary software applications in use, those which are being used by more than 10 respondents. This table shows two, SPSS and MS Excel, as being used by 291 and 277 respondents respectively. After these two, the numbers fall away dramatically, showing a range of software applications most commonly associated with statistical and social science analysis, spreadsheets and databases, image management and manipulation and some GIS software.

Software named between 2 and 9 times

Leximancer, Mathematica, Nudist, Origin, MS Powerpoint, Cellquest, Acrobat, Adobe Illustrator, AMOS, EndNote, Genstat, gnuplot, SPM, Coreldraw, JMP, Tecplot, Ucinet

Atlas-Ti, C++, FORTRAN/NAG, IDL, LISREL, Noldus Observer, Office, Primer, SQL, Statistica, , ArcView, Bioedit, ELAN; Epi-Info, freehand, FSL, Genespring, ImagePro Instat, Kaleidagraph, KodakImg, LeicaIM, MASCOT, MLWin, NMRPipe, PAUP, vtk , Image Quant, Praat, Root, Xepr, AIS imaging analysis, Amira, Analyst v, ANGIS, ArcInfo, Athena, Axiovision, BIAevalutaion, Bio-Rad, CARA, Cassa XPS, CERVUS, ConQuest, Delphi, Entourage, Envi, EQS, EstimateS, Finch TV, Fluent, Google Earth, Graph Pad, HLM, Image Magick, IMOD, iPhoto, IRAF, LaTeX, , MacClade, MacVector, Magpie (MEA), Maya, Medcalc, Mega, MicroFit, MySQL, N, Novel Pliance, Octave, OzQuest, Paraview, Php, Postgresqul, ProFit, RealPlayer, ScionImage, SEDFIT, Sequencher, shazam, Solidworks, Spike, SuperStar suite, Survey Manager, SurveySaid, Toolbox, Transcriber, Varian Resolutions Pro, VICON, Weasel Primer, WordPress, Xmgrace

Of the other 495 proprietary software applications in use, 122 were mentioned between 2 and 9 times and 372 were mentioned only once (see Appendix B). This demonstrates a very long tail, and has significant implications for repository and other data managers who have to deal with an extraordinary range of file formats and the sustainability issues relating to each. Many of the products come from Microsoft, and there is one school of thought that suggests that the sheer volume of files in MS formats means that the issue of their sustainability will be widely supported and a solution found should they go out of date. Others are not so sanguine. Not all of the data will need to be stored in the formats used by these softwares applications, and data can often be satisfactorily reformatted. However, this all takes time and effort.

The question was framed so that it required a free text response. As a result, analysis was difficult, because of the need to standardise names and correct spelling. Respondents used a variety of terms to describe non-proprietary softwares and it was impossible to quantify them. What can be said conclusively is that they were mentioned far less than proprietary softwares. The terms used included (and it was not always clear if these include proprietary softwares): customised, in-house, in-house visualisation, instrument specific, generic terms such as "spreadsheet", image analysis, open source tools, own scripts, freeware and xml tools.

It is questionable how much of the software named is in fact used for "analysis and manipulation", and one gets the impression that some respondents simply named all software they have to hand. Examples of these would include Digitool (a repository service), perl (a programming language) and DreamWeaver (a web editing tool).

Software storage and retention

QUT included a special question: "how do you store and retain any software used to generate your research data?" The responses, which were all in free text, on occasion showed a degree of puzzlement. Perhaps it had not occurred to some respondents that software storage might be an issue.

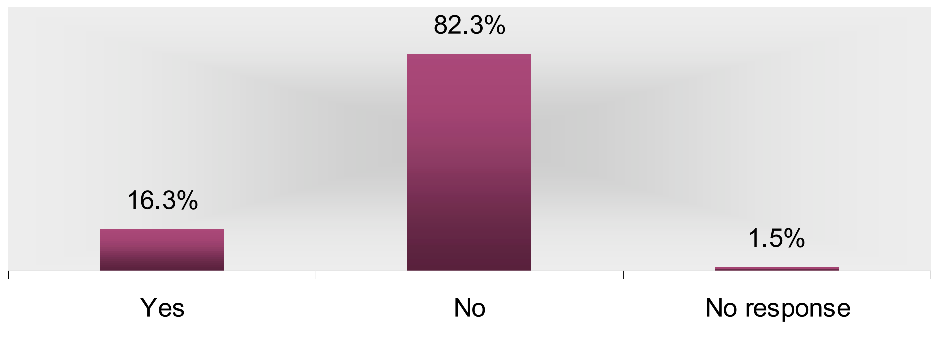
There were, at the same time, responses which indicated that the issue was well recognized and being addressed with care, especially when the software in question has been locally customized or created. Some of the responses, indicating the variety of storage solutions, are shown below:

> *For non-commercial software we archive source code for the various versions and recompile for new versions of the Java VM to ensure that the code is kept up-to-date with the software libraries*
>
> *On hard drive and cd back up*
>
> *University PC*
>
> *CD-ROMs, University virtual storage*
>
> *save on multiple discs*
>
> *corporate network*
>
> *have a license to use on my computer*
>
> *local [...] repository*
>
> *optical and magnetic media back-up.*
>
> *For non-commercial software we archive source code for the various versions and recompile for new versions of the Java VM to ensure that the code is kept up-to-date with the software libraries*
>
> *DAT tape, DVD*
>
> *password protected e-files and locked file cabinets in office*

> *Keep original media and any developed source code.*
>
> *Data is processed through databases and stored as SuperCubes. Software is licensed and updated as required. Source data is always available in open format (sql tables) so software is not the lock-in.*
>
> *Retain all install files*
>
> *Software is installed on desktop- and analysis computers, and backed up on CD (as appropriate)*
>
> *There was also some mention of license management and need to ensure that new versions are purchased and maintained.*

Research Data Management Plans

Figure 3: Do you currently have a formal research data management plan?



Research data management would be easier for all concerned if researchers, research units and research organisations all had policies and plans surrounding the creation and management of data. This survey asked whether individual researchers currently have a formal data management plan.

Over 80% of respondents acknowledged that they do not have a formal data management plan. This suggests a need for advocacy and training within the universities. There is currently no formal requirement for researchers in any of the three universities involved in the survey to have a data management plan, although this might change in the future. There is pressure from funders, especially government funders, to ensure that data, once created, is properly managed and stewarded. And there are many who would prefer that the issue of data management is raised at the beginning of the research process rather than later, when it might be too late to prevent data losses and difficulties.

An analysis by discipline shows some differences. Current opinion suggests that the science disciplines are more attuned to the need for good data management than those in the humanities and creative arts. In general, the results shown in Table 5 do not show this to be the case, as the largest proportion of those with data management plans are in the Social Sciences (25.5%), and Medicine & Health (21.2%). While the Humanities & Creative Arts appear nearly at the bottom of the list, on only 10.6%, they are only just below those in Science (11.1%), Engineering & Architecture (12.8), IT (13.0%) and Business & Economics (15.3%). The Humanities & Creative Arts do come in above Law, with only 8.3%, but the figure for Law should be interpreted with caution as the sample is so small.

Table 5: Do you currently have a formal Research Data Management Plan in place?

Discipline	Yes	%	No	%	Total
Business & Economics	9	15.3%	50	84.7%	59
Engineering & Architecture	11	12.8%	75	87.2%	86
Humanities & Creative Arts	11	10.6%	93	89.4%	104
IT	3	13.0%	20	87.0%	23
Law	1	8.3%	11	91.7%	12
Medicine & Health	64	21.1%	239	78.9%	303
Science	24	11.1%	192	88.9%	216
Social Science	13	25.5%	38	74.5%	51
Total*	136	15.9%	718	84.1%	854

* Non-responses and no disciplinary affiliation excluded

Data storage and backup:

There was a wide variety of responses to a question about data storage and backup, with most respondents indicating that they use more than one system, and less than 1% saying that they have no system at all in place. The small number who reported that they don't know how their data is backed up (2.6%) at least know that this most basic of data housekeeping is taken care of, even if they don't know who has responsibility. Presumably their backup is provided by their department or the university more broadly, and it is likely that such a central service would be effective.

Whether the other data storage and backup systems mentioned are effective is not apparent. The most frequently mentioned storage and backup systems such as USB/Flash drives (65.2%), CD-Roms (55.7%) and DVDs (38.8%) may be useful in the short term but are unlikely to have any value over long periods as they deteriorate, can no longer be read or get lost.

The Storage Area Network (38.5%), Offsite Storage (22.1%) and Tape Storage (15.2%) would seem to be more reliable, although Offsite Storage can, and did, mean a variety of things. Some of the comments provided more detail about offsite storage, which could mean at home, or emailed to a gmail account or other cyberstore facility, or held by a research partner in another institution. The 11.6% who ticked the "other" box did not provide a lot of detail about what this might have meant, with one person reporting "Some simply in boxes in research rooms".

Table 6: What data storage and backup system do you currently have in place?		
	n	Per cent
USB/Flash drives	573	65.2%
CDs	490	55.7%
DVDs	341	38.8%
Storage area network	338	38.5%
Offsite storage	194	22.1%
Tape storage	134	15.2%
Third party (incl. commercial data storage)	37	4.2%
None	5	0.6%
Don't know	23	2.6%
other (please specify)	102	11.6%
Central storage *	37	4.2%
Total respondents **	879	

* This option was available for one university only

** some respondents clicked more than one box

The many comments accompanying the responses contained grumbles about the inadequacy of the current situation and about the difficulty of keeping track of large and diverse collections. There were calls for more offsite storage, calls for more recognition of the importance of good data management and some scathing calls for more up to date institutional practices.

Responsibility for data management?

Table 7: Who is currently responsible for managing the data?		
	n	Per cent
Yourself	684	77.8%
Research project manager	178	20.3%
IT staff within your school, centre or research institute	171	19.5%
Designated person on project	169	19.2%
Research assistant	121	13.8%
ITS (or equivalent in each)	34	3.9%
External project partners	27	3.1%
Nobody	10	1.1%
Don't know	11	1.3%
other (please specify)	24	2.7%
no response	10	1.1%
Total respondents*	879	

* some respondents clicked more than one box so does not equal total responses

The overwhelming majority of respondents said that they manage their own data (77.8%). "Nobody therefore myself" was one sad comment. This proportion might at first seem alarmingly high, but the comments provided show that this was not seen as being other than expected. Research students, in particular, manage their own data for their theses in all of the universities surveyed. One Director of Research was quick to point out that the responsibility lay with him (or her) as head of the team, even where there are designated individuals who manage the data. Someone else pointed out that the designated person with responsibility varies from project to project, depending on the principal investigator or supervisor.

Smaller groups reported that responsibility was held by the Research project manager (20.3%), IT staff within the school, centre or research institute (19.5%) or a designated person on the project (19.2%). Keeping in mind that many respondents selected more than one category, this suggests that responsibility for data management is considered carefully in many cases and that there are assigned responsibilities, especially where research is conducted in teams. The same might be said where responsibility for data is held by a research assistant (13.8%).

Two very small groups responded that nobody has responsibility for their data management (1.1%) or that they didn't know (1.3%). Where "other" was nominated, this most often reflected that the researcher was part of a larger group working across institutions.

The previous question asked about a data management plan, and to some extent the responses to this question reflect the same disciplinary differences. Those in Law and Medicine and Health seem to be the best organized when it comes to having designated responsibilities for data management, to having the lowest proportion managing their own data and largest proportion with a designated data manager. The figures for Law, however, are very low and the sample cannot be regarded as representative.

The group which stands out among the disciplines is the Humanities and Creative Arts. They have the highest proportion managing their own data, the lowest proportion with local IT staff support, the highest proportion with nobody managing data, and the highest "don't know".

Other comments included:

> *I have some audio data stored at AIATSIS in Canberra and they manage access to it under pre-arranged conditions.*
>
> *I don't manage my data. I read, think and write. This is not management.*
>
> *Project data is the responsibility of Project Leaders. The Centre collects data related to our Key Performance Indicators (as required by the ARC) centrally, in a Centre-developed web-based Content Management System; note that because the Centre spans different institutions (including Schools in different faculties in [university], plus other universities), the CMS is publicly available with password protection.*
>
> *[Data] is managed by [faculty] IT section. Archiving is managed by researchers in conjunction with archivist. Reports and publications handled through [institutional repository] where possible.*
>
> *Data is generated by students and staff. They are responsible for committing it to backed-up location.*
>
> *Data collection and management conducted by the user. This includes students. The software for data collection is managed by the facility manager.*

Table 8: Who is currently responsible for managing the data?

	Business & Economics		Engineering & Architecture		Humanities & Creative Arts		IT		Law		Medicine & Health		Science		Social Science	
	n	%	n	%	n	%	n	%	n	%	n	%	n	%	n	%
Yourself	47	79.7%	70	81.4%	87	83.7%	17	73.9%	7	58.3%	220	72.6%	177	81.9%	39	76.5%
Research project manager	4	6.8%	18	20.9%	10	9.6%	8	34.8%	3	25.0%	89	29.4%	24	11.1%	17	33.3%
IT staff within school / centre / research institute	13	22.0%	17	19.8%	6	5.8%	5	21.7%	2	16.7%	65	21.5%	52	24.1%	4	7.8%
Designated person on project	6	10.2%	15	17.4%	11	10.6%	3	13.0%	4	33.3%	72	23.8%	43	19.9%	9	17.6%
Research assistant	3	5.1%	14	16.3%	8	7.7%	2	8.7%	0	0.0%	64	21.1%	20	9.3%	9	17.6%
ITS	1	1.7%	2	2.3%	3	2.9%	6	26.1%	1	8.3%	6	2.0%	11	5.1%	2	3.9%
External project partners	4	6.8%	1	1.2%	0	0.0%	0	0.0%	0	0.0%	12	4.0%	3	1.4%	3	5.9%
Nobody	0	0.0%	1	1.2%	2	1.9%	0	0.0%	0	0.0%	4	1.3%	4	1.9%	0	0.0%
Don't know	1	1.7%	2	2.3%	3	2.9%	0	0.0%	0	0.0%	4	1.3%	2	0.9%	0	0.0%
Other (please specify)	1	1.7%	4	4.7%	3	2.9%	1	4.3%	0	0.0%	9	3.0%	3	1.4%	3	5.9%
Number of respondents	59		86		104		23		12		303		216		51	

Note: non-responses and no disciplinary affiliation excluded

Some respondents answered in more than one category

Table 9: Who will be responsible for looking after the research data after the research project has concluded?

	n	Per cent
Yourself	64	49.6%
Research centre	10	7.8%
Supervisor	6	4.7%
Project Manager	1	0.8%
Nobody	1	0.8%
Don't know	16	12.4%
other (please specify)	6	4.7%
no response	25	19.4%
Total respondents*	129	

The QUT survey asked “Who will be responsible for looking after the research data after the research project has concluded?” and required text responses only. For the most part the answers reflected the responses to the previous question, with the largest proportion, one-half, nominating themselves as having long term responsibility for their data. Research centres and supervisors were the next most frequently nominated. More responded “don’t know” than did to the previous question.

Among the comments were references to external bodies taking on responsibility, the need for some data to be destroyed in accordance with confidentiality agreements, the possibility of future publication potential and the future of the research unit, expressed as “Ah, very good question. Succession planning is problematic. Only people with a stake in the data care about maintaining databases!! Maybe the CRC for […] Headquarters - but the CRC may terminate soon - so then maybe the Faculty - but I doubt it!!”

Data sharing

At a time when researchers are being encouraged to make their data available to others, it is pleasing to see that over three-fifths of respondents are willing to share their data, whether “openly” (8.6%), “via negotiated access” (44.0%), “only after the formal end of a project” (6.4%) or “only some years after the end of a project” (2.3%). In addition to these, a small proportion (0.8%) provides access through the Australian Social Science Data Archive, IATSIS or some other data archive. Some respondents pointed out that, in some cases, it is necessary for data to be made available together with journal publication, and it is likely that this is a trend which will grow.

About two-fifths of the respondents say that their data is never made available, for unexplained reasons (19.0%) or because of privacy or confidentiality issues (17.6%). About one-quarter of this group indicated that they would be willing to make their data available if “an easy mechanism” was available to do so.

Table 10: Do you allow researchers outside your team to access your research data?

	n	Per cent
Openly	76	8.6%
Via negotiated access	387	44.0%
Only after the formal end of a project	56	6.4%
Only some years after the end of a project	20	2.3%
Not at all	167	19.0%
Never, because of privacy and confidentiality issues	155	17.6%
Not at present, but I would be willing to make some or all of it available if an easy meachanism to do so were offered at [inst]	117	13.3%
Access is provided through the Australian Social Science Data Archive (or similar) after data is deposited there *	7	0.8%
respondents	879	

* alternate wording in one of the three surveys: Access is provided through a national, international or disciplinary data archive such as the Social Science Data Archive or the RCSB Protein Data Bank.

The comments provided further insights into the issues around making data available to others. Some commented that their data would be meaningless to others ("they'd have no idea what they were looking at") and others that there has never been any interest in their data. There were issues of trust in some cases, and the need for de-identification and the obtaining of consent which can be time-consuming. The following two comments are illustrative:

> *A significant problem in the work I do is recruiting informants, and an important concern to be met is that the purposes of the research be clear, the methods be clear, and boundaries set accordingly. Should the data I collect be made available to other unknown researchers, for other unknown purposes, using other unknown methods, I expect that potential informants will be justifiably less likely to participate.*
>
> *My data is qualitative and while I would be happy for it to be shared, there are some ethics hurdles that would need to be overcome.*

Some data has been made available to the researcher only under license from providers, so it cannot be passed on, and some researchers provide access on a project by project basis, where some data might have to be destroyed and other data not. There were mentions of data being made available once all the analysis is complete and after Intellectual Property has been established. There were comments to the effect that the data is not of interest, but the simulations and modelling around it "have to [be] made available for scientific scrutiny." Some saw models and algorithms as not being data, and "If any lazy sod should ask for it, I would tell them to write their own."

The possibility of an easier mechanism to allow data deposit and access was welcomed by some, as in the following comments

> *I readily share data with colleagues or students working on same or related project on an informal case by case basis. I would like to have access to an area where I could put data files for access and download by colleagues. Sending large files via email is not really possible, and sending data on CDs is also very time-consuming, especially with large files. I would very much welcome a solution to this problem that doesn't cost an arm and a leg to the researcher or school.*
>
> *Currently this is achieved through project www sites and some formal international data repositories but having an easy to use infrastructure to deal with this would be excellent and I believe would represent a high value intellectual asset.*
>
> *I have done this but there are no easy ways of doing it. I would very much like the Uni to offer such a service*
>
> *A key aspect to [Research Centre] operations are electronic links to other organisations to facilitate customer access, manage data and integrate equipment. A standard framework for such access would facilitate such links.*
>
> *Access is rather ad-hoc and depends on the instruments used. Would be preferable to have a central repository.*

It is not clear from the responses just how much of the data collected would be appropriately deposited with the Australian Social Science Data Archive (ASSDA). However, if less than one per cent of data is available through ASSDA or other agencies, this suggests a disappointingly low rate of deposit.

Data access and use

Table 11: How is your data accessed or used? Please check all that apply:

	n	Per cent
As raw data	463	52.7%
Datasets as a whole	353	40.2%
In small chunks	308	35.0%
Only after filtering, manipulation and analysis	305	34.7%
In original or print form	292	33.2%
Locally	283	32.2%
Online via a website or service	134	15.2%
Online via Grid, Storage Resource Broker, etc	22	2.5%
Other (please specify)	31	3.5%
Total respondents	879	

Note: Some respondents answered in more than one category

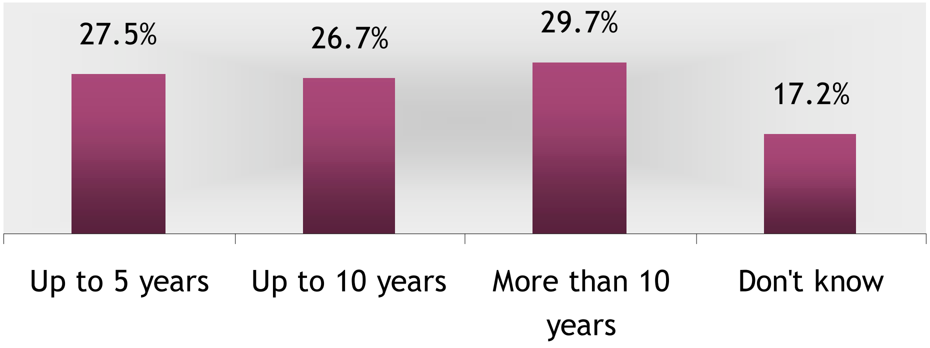
If the university is to improve data management facilities, it needs to know how researchers access and use their data. This question provides some of the answers. Table 11 shows the responses to the question, and it is immediately apparent that there were over 2,000 responses to the question from 879 respondents. The implication to be drawn from this is that researchers use a variety of means, depending on the nature of the data or the type of project, although one grumpy respondent commented “My work--whether on the web or in print sources--is not accessed or used. People read it and, I hope, think about it. Sometimes they even quote my words, which they have read."

The majority of respondents access their data as raw data (52.7%), whether using datasets as a whole (40.2%) or in small chunks (35.0%). About one-third access it only after filtering, manipulation and access, explained neatly by one respondent in the comment "Usually my data requires significant interpretation and the generating code is more important than the data". About one-third access it locally (32.2%), a smaller group (15.2%) access it online via a website or service and only a small proportion access it online via the grid, Storage Resource Broker or some comparable facility (2.5%). "What the heck is a 'Grid'?" asked one respondent, a question echoed by others whose work does not require an understanding of high capacity computing.

More unexpected is the finding that just on one-third (33.2%) are accessing their data in original or print form. The message to be learnt from this is that print is not dead, and that those original sources, whether survey questionnaires, artworks, artifacts, forms, or documents, continue to play an important role. One respondent mentioned the importance of an EndNote database, presumably supporting a text-based research project.

Data value

Figure 4: How long do you think your research data will have value?



Planners of data facilities need to know how long they will be expected to keep data. The questionnaire therefore asked people to provide an estimate of how long they think their data will be of value. Most were prepared to make an estimate, with 27.5% suggesting up to five years, 26.7% suggesting up to ten years and 29.7% suggesting over ten years. Less than one-fifth said they did not know (17.2%)

There were no comments sought to this question. The fact that a very small number of people answered in two categories indicates that some people may not be sure about their answer (or that the second option available also includes the first).

Training needs

Table 12: Would you be interested in training or advice on any of the following? Please check all that apply.

	Yes	Per cent
Digitisation advice, tools and services	269	30.6%
Creating a research data management plan at the beginning of a project	457	52.0%
Creating a research data management plan after a project has finished	197	22.4%
A data 'exit' plan (for retiring academics or departing academics and postgraduate students)	289	32.9%
Data 'rescue' for older digital materials, such as data on older media or migration of data from legacy systems	198	22.5%
Other (please specify)	24	2.7%
Total respondents	879	

Note: Some respondents answered in more than one category

One of the purposes of the survey was to find out if there was any demand for training in different aspects of data management. The results showed an overwhelming demand, but also a reluctance by some to be engaged in what was seen to be a further imposition on research time.

Three-quarters of respondents wanted training related to data management planning, either creating a research data management plan at the beginning of a project (52.0%) or after a project has finished (22.4%). Another large group (32.9%) wanted a data “exit” plan, a topic designed for researchers who might be retiring or leaving the university or completing a postgraduate degree and moving on.

Help with digitisation was also keenly sought by nearly one-third of respondents (30.6%). This could be for different kinds of material – text, images or audio. Help was also sought with older digital materials which by now have become unusable as they require older media or data migration (22.5%).

Other types of training were suggested, some of which were not concerned with data management as such (such as NVivo): collaboration in developing international research datasets, intellectual property, managing data as records and integrating data with other research records and information, “I am especially interested in archiving data from a longitudinal data set involving both quantitative and qualitative data”, organising databases for medical imaging, MySQL for managers, data manipulation of video action research to protect anonymity, subversion control software and enhanced data sharing, and networking of data storage and analysis applications were just some of the topics mentioned.

There were many supportive comments on the need for additional training as can be seen by the following:

> *As a member of the [committee for this department], we have discussed the issue of data storage and access taking into account participant confidentiality and risk. Any form of training for [department] members would be useful, I think.*
>
> *Researching new media (VOIP, Video conferencing etc), need guidance on management, manipulation and storage to protect the data collected as well as commercial-in-confidence information.*
>
> *I believe more education, and ideally coupled with easily accessible backup/storage facilities, is required for long term storage for large datasets. Colleagues of mine had an annoying incident of data loss several years ago, due to storage on recordable DVDs, which are not an adequate medium for long term archival purposes. This was in part due to ignorance of good procedure and in part due to lack of funds or facilities to maintain proper backups of large (>100 Gb) datasets.*
>
> *I feel I have had no training in how to set up or manage data and am concerned about longevity of the data storage and platforms that I am using. I anticipate that I and others will still be using the data I am now collecting well into the future (as it is historical data, so does not date as much as other disciplines) and am unsure how to make sure it is still accessible.*
>
> *I feel that it is very important to start training students in issues relating to data management including ethical issues along with technological.*
>
> *Please can training be online, as trying to find time for workshops is impossible and I would like all my staff to do it but need to manage within various commercial programs and teaching responsibilities*
>
> *I realise I don't really know what I'm doing because this survey has raised a lot of questions that I hadn't thought about. I would certainly get a lot out of training. I'm guessing the situation would be similar for most students.*

There were also comments showing that for some people, the need for training and additional time to be given to data management was not welcome. This indicates that the cultural and organisation change required to improve data management practices will encounter some opposition along the way.

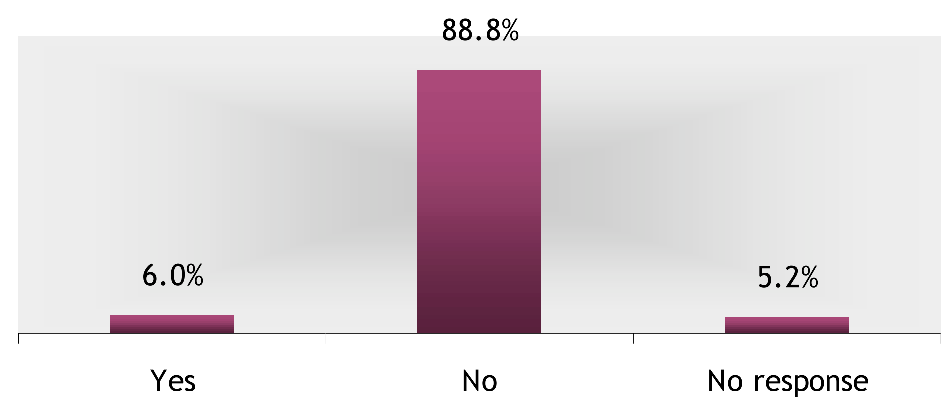
> *What you think I might be interested in, as above, makes me feel sick, frankly! A question--maybe what you are asking isn't really relevant for humanities disciplines, or at least some of them?*
>
> *There is a real problem with time for any training. I hardly have time to do the work coming across my desk at the moment. With faculty and university re-structuring, it is possible that staff workloads have increased significantly.*
>
> *No, I do not want "training”, I want somebody to do the work for me.*
>
> *NO MORE TRAINING! NO MORE USELESS UNIVERSITY BUREAUCRATS! PLEASE JUST LET ME DO THE RESEARCH.*

As an adjunct to the question about training, the survey asked if respondents would be willing to participate in, or provide information to an eResearch reference group to be established at the university concerned. Each of the universities has recognized that infrastructure support for eResearch has to be integrated into the research process and that the participation of active researchers is essential to the design and development of support mechanisms.

Just under one-half of the respondents indicated their willingness to be part of this initiative (44.6%). Some commented that a lack of time would prevent them from taking part, or gave other valid reasons.

eResearch

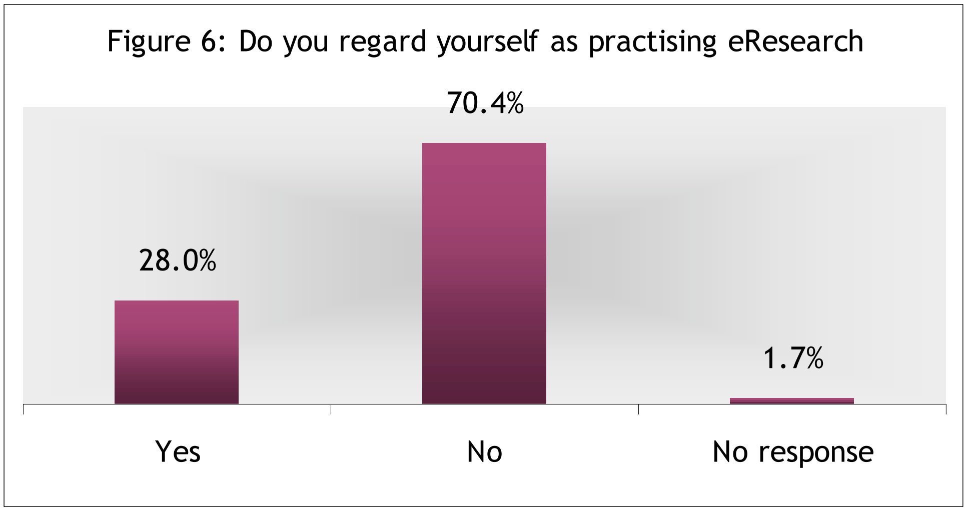
Figure 5: Do you use Grid/high performance computing?



The University of Melbourne included two questions on eResearch, one about use of the Grid/high performance computing and the other about the practice of eResearch itself. The responses to the question about use of the Grid/high performance computing indicate that very few, only 6.0%, use these facilities. This group might be small, but it is the group which tends to be heard most often in debates about data management. The large proportion who might be classified as low end users, nevertheless have a strong interest in data management issues and are important to institutional infrastructure support planning. In some instances, they still see themselves as being engaged in eResearch.

The second question asked whether researchers see themselves as practicing eResearch, and about one-quarter replied “yes” (28.0%). Among those who don’t (70.4%) are many who nevertheless are conducting research which is heavily dependent on data collection and analysis. Many acknowledged that they do not understand what is meant by eResearch, or that the barriers between eResearch and any other research, are not clear. This can be seen in the following comments:

> *I don't actually know what is meant by eResearch. My work uses digital data storage for recordings, images, video files, [and] derived data. We also undertake computer simulations using realistic models and this generates very large data sets. This would not be possible without computer based digital storage, but the key is biology, the computer is merely the tool for organising data.*



I wouldn't be practicising eResearch at the moment but as a research centre we have for about 2 years now explored what is possible re developing a database for the large amounts of qualitative data we have and that is generated by partners. We see this as necessary and important for a number of reasons. We have a small grant to explore the ethical and management issues of this at the moment but in the past we have been frustrated overall about the lack of IT expertise and guidance we have received in investigating this.

Table 13: Do you regard yourself as practising eResearch? *

Discipline	Yes	%	No	%	Total
Business & Economics	9	31.0%	20	69.0%	29
Engineering & Architecture	7	15.9%	37	84.1%	44
Humanities & Creative Arts	22	28.6%	55	71.4%	77
IT	2	66.7%	1	33.3%	3
Law	2	22.2%	7	77.8%	9
Medicine & Health	54	28.9%	133	71.1%	187
Science	24	23.1%	80	76.9%	104
Social Science	8	36.4%	14	63.6%	22
Total	128	26.9%	347	73.1%	475

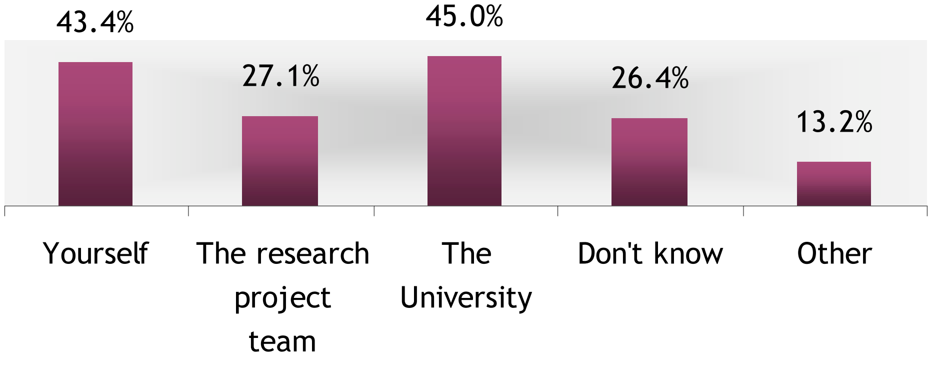
* This question was asked at only one university

Note: respondents with no disciplinary affiliation excluded

There are interesting differences in the responses when analysed by discipline. The disciplines with the largest proportions considering that they are engaged in eResearch are IT (66.7%), Social Science (36.4%) and Business and Economics (31.0%). Of note here is the fact that Social Science and Business and Economics are not usually associated with eResearch: more often seen as the preserve of Science (23.1%) and Medicine and Health (28.9%). Even the Humanities and Creative Arts (28.6%) have a higher proportion professing to practice eResearch than Science, and Engineering and Architecture have the lowest proportion on 15.9%.

Data ownership

Figure 7: Who owns the data generated in your research?



The Queensland University of Technology survey included two questions related to copyright and data ownership in order to get some idea of the level of understanding researchers have about who owns the copyright in their data and how they know it. Figure 7 shows that there is a range of views about copyright ownership, with 45.0% saying that the university owns copyright, 43.4% saying that the researchers owns it and 27.1% saying that ownership rests with the research project team. Over one quarter (26.4%) acknowledged that they did not know, and a further 13.2% ticked “other”. Among the others nominated were: data suppliers, supervisors, copyright held in association with research collaborators elsewhere, participants in the study, “my company”, another university, government funders, the project sponsor, the external body commissioning the research, and “Original contributors such as ABS and ATO with our layers of value add”.

The comments indicated an element of uncertainty about the issue.

> *I think!*
>
> *I 'own' the IP generated in the thesis and papers published. Not sure about the actual data.*
>
> *I think that as a PhD student, I own my programs and their output, as long as they're not worth anything (if you get my drift). If I did anything nifty, the University would get a large share.*
>
> *owners of research are me and supervisors.*
>
> *Still to be fully investigated.*
>
> *I own some data [Ph D] and other data are owned by the university [Research Projects].*
>
> *But not 100% certain.*
>
> *The uni and whomever it is in bed with.*

I expect the University in most cases but hope that my contribution would be rewarded if ever IP I helped to create was ever commercialised; in the case of RHD projects, IP in data usually remains with my students.

The external body that commissioned the research, own the results of the research, however from an ethical perspective I'm not sure the data would be passed over to the external body.

Not clear if the university owns this also.

Table 14: How do you know who owns the data?		
	Yes	%
I was told	102	79.1%
There are contracts or policies	19	14.7%
It is just "understood"	46	35.7%
It is covered in my employment contract	35	27.1%
Other	12	9.3%
No response	13	10.1%
Total respondents	129	
* Respondents could tick more than one response		

Responses about how people know who owns the data fell into two broad categories. The first group included those who could confidently assert that there are contracts and policies which cover such matters (14.7%), and those who think that it is covered in their employment contract (27.1%). One respondent even went so far as to quote the relevant subsection of the university's IP policy on intellectual property created by students. Another cited the instance of an agreement reached with the university that his company would maintain the IP. Where respondents said "other" (9.3%), this most often referred to IP held by a company for commercial purposes or IP held by another institution by agreement.

The second group of responses came from people who replied "I was told" (79.1%) or that "it is just understood" (35.7%). This group seems to lack the certainty of the first group, being less secure in their response. Comments such as "I don't really - I'm just guessing that as a staff member using funding from the University, it is both of us" and "I hadn't considered this before - will ask supervisor" seem to confirm this.

Final comments

At the end of the survey, respondents were asked to add any other comments regarding data management, long term data storage and access, digitisation, training, and so on. There were many who took the opportunity to make supportive comments about the survey and what they saw might lead to improvements in the current situation. A small number took the opportunity to make complaints about the survey, the university administration, or life in general. There were comments about the current state of data management, and some suggestions about what might be done to improve research data infrastructure in future. Comments related to training, to eResearch and to ethical issues relating to data have been included with those topics as discussed earlier in this report. The following long (and unedited) selection is included to provide an indication of the issues covered.

> *This is a great initiative*
>
> *A very important issue and we are struggling to get this up in a cost-effective way as research funds never pay for this type of activity.*
>
> *huge issue and not one i have adequately sorted in the lab environment despite years of trying - biggest issue is getting buy-in to manage data sensibly from co-workers.*
>
> *In a large, distributed and complex beast like [this university], the diversity of practice across the wide range of research disciplines means that evolving the infrastructure needed to support research in the digital era is not going to be easy. It is, however, essential. We have to be prepared to make mistakes, to try things out and experiment. We have to be very conscious of the broader framework in which we are working and constantly try to reveal the deeper principles of practice. Consequently, we have to be careful not to limit our research record management practices to what current technology offers - as this will have changed during the life of the project. On the other hand we do have to use the latest technologies to the best of our abilities to bring increased productivity and services to researchers. We have just entered the 'Wright Brothers' phase of the Digital era.*
>
> *Time consuming to insist that everyone does it, and to build efficient systems for handling data, but we must improve it for long term storage and data verification, security of info etc.*
>
> *It has relevance in my field of ecology, which generally has a low level of data sharing and integration between projects (i.e. poor data storage and longevity)*
>
> *It is critical [this university] facilitates data intensive research with the provision of analytical infrastructure. The issue of curating completed project's data becomes much simpler if well documented datasets are estabished and utilised for analysis during the project. [This university] needs to foster SIGs/forums etc for the linking and amplifying the expertise that is diffused throughout faculties and projects.*
>
> *Needed area of consideration as I have only just started to realise that I need to be more organised in storing my data as my computer can not cope. Also being off site from [university] collecting data and analysing it raises issues with storage as well.*
>
> *All of this sort of stuff is actually peripheral to doing the research - i.e. it takes time away from the actual stuff that you are meant to be doing.*
>
> *Clearly piles of CD's etc is an unsustainable strategy, but I am concerned about the tail wagging the dog here. Most journals encourage submission of datasets to their on-line repositories so the problem is beginning to go away (at least a bit).*
>
> *I think the issues regarding the storage of research data varies significantly between disciplines. The University seems to have a one-size-fits-all policy which is rather silly when applied to some fields.*
>
> *No time to think about it because of the very heavy workload*
>
> *This has the feel of yet another make-work exercise that will ultimately get in the way of actually getting any research done.*
>
> *This has the smell of yet another [university] bureaucracy that serves the group of geeks that run it but is nothing but an impediment for the researcher at the coal face*

this seems totally useless

With respect to the above question, I'm not paid enough to get involved in this (half time, level B, soft money)

I FEEL THAT SURVEYS LIKE THIS WASTE THE TIME OF RESEARCHERS. LESS BUREACRACY IS THE ANSWER.

Data management bureaucracy seems unnecessary for social science research.

Erk, is the best I can say. Sorry to seem hostile, but your questions don't really bear much relation to my work though I hope they are helpful to other people.

The data archive is so large, it is physically difficult and expensive to manage.

data management is a debacle within the university. if RQF outputs cannot be managed properly in [the system] I wouldn't be holding my breath that research data could be stored successfully

A key problem is discovering what solutions are available and who offers them for data management, digitisation, data storage/access in particular.

We have very limited access to secure network back-up space in nursing which I consider preferable and more secure than CDs etc

currently, student appear to be expected to manage data themselves. seeing other students, i think it is pretty poor where students that are funded by external parties are still expect to manage data themselves, rather than being given dedicated storage provided by the university or the project. DVD, flash drive or local hard drives is not really an acceptable storage and retrieval medium. it is also very very poor that it is now 2007, and the university as a whole still does not have proper electronic data management practices in place. i am aware of several projects over the past few years where the proverbial all data was lost because the university did not provide adequate systems to the student. the same philsophy shoudl also be used for undergraduate students. from what i have seen, providing student with a token amount of storage space just isn't really acceptable, especially when services such as google provide multi-gigabyte storage for no charge. it is also quite offensive when academic tell student to use these external services for data management as if to say to the student that data manage shoudl be in place, but the university isn't goign to provide any resources to the student to assist this. basically the universities in general are failing in their duty of care to the the students and certainly from a marketing perspective, grossly failing to provide adequate customer care their customers, being the students, internal funding bodies, and external funding bodies.

In Science there has been an almost complete refusal to acknowledge this issue. The policy (such that it is) is that the individual is responsible to archive and secure digital data. This covers research and teaching materials. I might add that server space is only made available on request and is subject to an unidentified process. I resorted to purchasing a back-up system from research funds, which transpired to be incompatible with the [university's system]. A complete waste of funds!

Just a general comment on what I see with respect to storage of data in the research environment at [university] in general. Far too much data is stored in one location on one PC with one hard drive inside it. If that drive dies, which they do often enough, the data is lost forever. There's data on PCs so old they can not be networked for the purposes of backing up, and in some cases have no USB ports for extracting data, leaving just the floppy drive! The importance of off-site backup is often neglected. For example, having data stored on a laptop and a USB stick stored in the same backpack is of no use if the backpack is stolen!

[The university] is a little behind on this front, as compared to other organisations I've worked for.

A major problem at University is the slowness of the intranet which greatly affects back-up processes to servers. It is so slow that transferring 1GB (for example) can take >2 hours. This results in staff and students in my group only backing up some files, with the risk of data loss.

As film and television production moves to file based acquisition (instead of tape and film) the data quantities to be managed are significant (1 second of uncompressed HD video is approx 178Mbytes) - backup, archive and management of petabytes of data is a task the industry is attempting to grapple with as we write.

Currently (at least in my school, […]) there is a lack of a good "archival" data system. Need access to storage other than the frequent use main server (which is full anyway) so that we can put data into a permanent (infrequent access) repository server or tape based system. Having our data on dvds alone is a bit scary - easy to drop and break, or get lost when staff change etc!... we need better systems for labs like mine that generate lots of high res images etc.

'Data management' has particular meaning in research - verification, consistency checking, cleaning, which are all post-collection issues and require their own specific protocol/plan. These are where resources are most needed by the typical researcher, and most cost-effective is IT input into design and formatting of customised systems for data collection, cleaning, and formatting in readiness for analysis. Rarely available and so researchers make do with limited understanding of the enormity of the task and longer-term ramifications of just 'making do'.

I have pointed out to [this university's] staff that their data management was non-existent before [a system] was established. However, their asset, HR, student, IT, and financial systems seemed to take priority.

My most pressing need is greater storage capacity and the ability to share data with researchers abroad. The [university] e-mail system is entirely unsuitable for me to send data to co-authors abroad. In addition, the limited storage capacity I have means I have to manually back up my data every so often to an external hard drive. Sometimes I forget especially when I am busy before a trip which is not good.

The amount of video data that we generate in our research will become increasingly problematic to store using current methods as I anticipate we will produce significantly more than 0.5TB a year of raw video through staff and postgraduate student projects. This does not include video transformed into formats suitable for analysis, reports and other materials generated after and through analysis.

The backup options provided via the Departmental servers are woefully inadequate - less than 200MB per staff member/postgrad is allowed, and there wouldn't be enough space on the disk if everyone used their quota. No training in research data management is provided, as the IT staff are too busy fixing breakdowns and installing software to pay any attention to advising people on how to back up their data. Members of our research group have bought their own (personal) USB drives to back up data onto.

The default storage (100 MBytes) for emails through [university service] is absurdly small considering I can be sent in the order of 20 to 30 Mbytes a week.

The pressures of research, teaching and administrative duties mean that there is too little time to implement and comply with any extra burden that might be generated by formal systems for management of electronic (or non-electronic) research data.

As far as I know, there is no secure area for me to store digital data for my thesis that is accessible anywhere in the world (e.g. via secure websites), has restricted access, and is backed up by the uni. This would be really useful.

Backup systems are the most important thing at the moment to understand and implement. It is easy to buy (and then fill) terabytes of storage for a workstation. To make it secure is extremely difficult.

Guaranteed integrity of all data is crucial for any research activity. Any server repository MUST maintain mandatory access profiling to prevent accidental or deliberate modification or destruction of stored data, e.g. malware (viruses, rootkits, etc). Any centralised service MUST demonstrate state-of-the-art backup and recovery facilities.

I can appreciate the logistical and practical difficulties of backing up data--one is constantly trying to manage everything on the computer alone, never mind managing what was backed up but no longer needs to be, which backed-up files need to be preserved, and which need to be updated. Hopefully those thinking about this problem more than I have some ideas, but I would like to caution against singular solutions which are designed to suit specific interests but are then forced on everybody (a la [the local system], which I can imagine is an accountant's or auditors dream, but is a shocking hindrance to far more people -- "self-service" indeed). I would argue that the primary problem at the local level in the short term is simply space and time/convenience, and in the long term is data format. The overhead required to partition all my data and work into bits to fit on CDs of DVDs means I'll rarely do it, and these formats aren't for forever anyway. Also, external hard drives in the sizes we would need to be practical are prohibitively expensive. And I have spent ages in the past combing through 5 1/4 floppies to transfer to 3" floppies, only to spend ages again transferring to IOMEGA and Bournulli (sp?) disks, then Zip disks, then CDs and DVDs, etc. I certainly don't have the solution, but having a keen understanding of the problem makes me cautious of suggestions that there is an obvious solution.

PhD orientation program should cover this issue.

[One part of this university] has LIEF funding to develop a prototype data archive to house digitised textual data. This will be a new node in the Australian Social Science Data Archives. The major challenge for the Australian Social Science Data Archives (all nodes) is how to secure recurrent funding. At the moment ASSDA funding has been primarily through local support and LIEF, but LIEF is not an appropriate mechanism for ongoing support for the archive.

All of these should be subject to an enterprise-level information architecture, supported by enterprise-grade IT architecture

chemoffice offers enotebook. i havent implemented it, because of lack of training/time, but believe a web-based chemoffice environment would provide enormous improvement in communication with collaborators, nationally and internationally

Could have a "tree like" depository system where users could store them after the project ended.

Data must be openly accessible, anything else is a waste of time.

How will this particular project integrate with existing national databases, and with personal lab webpages?

I have worked in a multinational pharmaceutical company in the USA where we had a greater choice of ways to keep our lab books. The method which I was particularly keen on (and was very popular within the company) was to have access to individually numbered A4 sheets on which we were then able to print our electronically kept lab *books for signing. Since the climate is such that signed lab books are very important for IP, I think that the university should seriously think about implementing a similar method. I think that it is recognised that it is quicker to write up electronic notes, rather than laboriously hand write lab books. As researchers we are under enormous pressure to maintain a high output. This would be an important strategy to allow researchers to maintain a high output as well a well kept lab notebooks.*

I think the best place to store and manage data is a database, but it needs to integrate well into scientific analysis software. I would prefer the database itself to have the functions, graphing,and statistical tools built in, but I'm not sure that it is available - MS Access comes close.

I would like to be able to make all my printed articles available for other researchers to use but not abuse but this is not currently on offer in my Faculty. As an older honorary working in Australian music history my articles will endure but need to be easily accessed by the public and all from the same site.

I would like to have a version management fro my papers - we used to have in Germany a server run by the uni where I could upload my files and got them from there every time and I could access with login from different location without the need to create a VPN.

If any prescribed practices are set in place, please make them simple and painless

It is too hard for individual researchers to run large digital data sets. We already carry large workloads and there is no use having data sets when you get no time to publish. What might work better is to look beyond fixed data sets and link data collection in the social sciences and humanities to the work the library does. A researcher could go to a librarian and specify a data set (e.g. video, pictures, electronic texts, statistical data, etc.) that the librarian would collect for the researcher. It would also be useful to get easy access to people who can create the online interfaces (e.g. online surveys) that would draw data into the university. In short, the library could see itself as a mediator between data and researchers in the social sciences and humanities. Finally, there is a lack of understanding of the potential of digital storage and access technology on the social science and humanities side of the university and that makes getting involved in such projects too hard. Much of the work I have done (and now am keen to let go of) has been regarded as irrelevant and has been met with apathy (and often worse) by the senior colleagues I report to. My chief collaborator in [this part of the university] has found it more amenable to go and work [for another university]. Seen from and social science and humanities perspective, it is clear that [another university] is years ahead of [this university] in its approach to these issues. The largest data set I have involvement in is the [department's] data set and will shortly be closed down. It has more than 1 TB but once it is gone I will mostly only have textual data stored electronically.

It would help if there was assistance from [university] centrally in data management policy, guidance, infrastructure etc. It probably exists, but is seemingly inaccessible to individual researchers.

Our requirements will include longer-term data storage/backup as well as IT support for data management and access. These have to be integrated with specialized software to ensure functionality of our entire, multi-instrument genomics platform. That is our needs go beyond data storage and management.

Perhaps it would be useful for any eResearch reference group at [this university] to collaborate with teams of [this university's] researchers who are setting up eResearch facilities. For instance, we are currently getting ready to survey identified qualitative researchers.

Qualitative research management is important too

> *The University can create a website data base and allocate space for all the schools. Every school should allocate space for each group. Every group should have allocated space open for everybody (official website) and space allocated only for people with special permission. In this manner everybody will have access to his data from everywhere (non-official website) and will have contribution with his official results to the official website of his group. Every group should have a manager of its website who will put in order the information and make it attractive to external visitors. Advertisements of the group can be made and investors can be attracted in this manner.*
>
> *The university is considering a central data repository, but this is only useful if the data placed there is well organised and is accompanied by metadata which includes the (instrumental) conditions under which it was obtained. For example, spectral or diffraction data tends to be specific to the instrument it was measured on.*
>
> *University guidelines on these topics would be useful*
>
> *We need a central site what gathers and sorts a range of data on a geographical scale. the number of times people must download the same data from ABS, etc and then construct their own datasets is too numerous to mention.*

The final comment, while irrelevant to the substance of the survey, was a vote of thanks to the Apollo system used to conduct the survey.

> *It would be great if [this university] could offer its own online survey system. This would overcome concerns about an outside body having access to the data and would be useful to have good backup in case of problems.*

The University of Queensland

Working with the Australian Partnership of Sustainable Repositories (APSR)[1] alerted me to the importance of sound data management practices. Over the life of that project, and in my role as a repository manager, I talked to many academics about their data, and was aware that, in many cases, data was threatened. For example, one painstakingly assembled dataset was housed on a single, ageing computer with no backup or networking facilities; other important data was not yet digital. Looming retirement is often a time when these issues become important for staff as the question of long-term stewardship must be addressed.

Many staff expressed frustration to me over the years about the lack of facilities for backing up their data. Difficulties with sharing data or making it freely available were also raised. Anecdotal evidence showed that staff used a variety of measures to back up their work, but there was no clear policy on how this should be done, how rigorously, or how often. Very few academics seemed to have a data management plan.

When the Online Research Collections Australia (ORCA)[2] project of APSR got underway in 2007, UQ needed to provide information about datasets held by UQ to the ORCA Collections Registry. The difficulty of identifying suitable datasets for inclusion led me to believe that we needed to get a better idea of what data we had and how it was being managed. I believed a survey of existing data management practices would help us understand our current practices and highlight those areas in need of improvement. The emerging eResearch agenda, and increasing requests for the UQ eSpace repository to house research data as well as publications, were also factors. The release of the OAKLAW report, *Building the Infrastructure for Data Access and Reuse In Collaborative Research: An Analysis of the Legal Context*[3] also influenced us to treat this issue with some urgency.

Discussions about the content of the survey were held with stakeholders such as staff of the Research and Research Training Division, and of the Information Technology Service, which manages large datasets on behalf of UQ researchers. The survey format was signed off by those bodies and an email requesting people participate was sent to all academic staff and postgraduate students by the Deputy Vice-Chancellor (Research). A news item was posted in the weekly *UQ Update* newsletter and a link to the survey was made from UQ eSpace news so as to maximise outreach.

The response was immediate and response numbers were high [approximately 8% of academic staff responded, many at a senior level]. Many respondents were generous with their comments and suggestions, so we were left with not just ticked boxes, but a wealth of anecdotal evidence about the current state of data management across UQ.

[1] See http://www.apsr.edu.au

[2] See http://www.apsr.edu.au/orca

[3] Anne Fitzgerald and Kylie Pappalardo, *Building the Infrastructure for Data Access and Reuse In Collaborative Research: An Analysis of the Legal Context*, OAK Law Project, Queensland University of Technology, 2007. http://eprints.qut.edu.au/archive/00008865/01/8865.pdf

Many also expressed willingness to be interviewed as a follow up, or volunteered to be part of an eResearch reference group.

What we got from the survey

The survey helped us identify the scale of UQ's data management issues. The survey provided actual evidence of failings in the current system, and identified the faculties or schools in which these failings were most common. The reliance on storage media that can fail, such as CDs/ DVDs or memory sticks, was a concern.

In many cases, the survey revealed that academics were confused about data management responsibilities. Many did not know who should be in charge of their data, or where to seek help or advice when they had issues with data management. Clarity on these issues was strongest in UQ's research-intensive institutes, probably because of the need to conform to the requirements of funding bodies regarding data management.

The survey showed that many staff were not aware of existing options for data storage and backup. Many had never used the UQ eSpace repository. Some staff had not even heard of it. Others used external repositories such as the Australian Social Science Data Archive but did not provide a pathway to data held there from any UQ system. Many academics simply assumed that their organisational unit was managing their data but many could not provide evidence to back this up.

Accordingly, the survey was useful in identifying areas where outreach needed to be strengthened. UQ eSpace publicity, fact sheets and marketing materials were all reworked, and more effort was put into doing presentations on the service at schools and centres. A university-wide Working Party on eResearch was proposed to try to develop a common approach to outreach. The need for consultancy on these issues was also highlighted. I have certainly taken on more of this kind of consulting on data management since the survey ran.

The survey identified people who wished to discuss their needs in person. This has resulted in greater awareness for many academics about the role of the UQ eSpace repository, and helped them understand who to contact with questions or problems. Certainly, some sections of the university now feel greater clarity about the different services on offer. Relationships between the UQ eSpace repository and many schools and centres have either been built afresh or strengthened. There is still some way to go on this, however.

Many respondents expressed a wish for more training. Three areas in particular were highlighted as training needs –

- Advice on data management plans
- Advice on digitisation
- Advice on data 'exit' plans, especially for retiring academics

Seminars on the first two were run successfully in 2007. Many academics attended both. The third will be run in 2008, with the first two repeated as well. The programs and selected presentations of both seminars were made available to APSR partners and to other interested bodies via a Web page.

Focus groups

The data management survey provided a lot of data about the status quo. However, it did not provide enough data to develop a clear view of the best way forward to

- inform academics about existing data management services and advice
- meet academics' needs for a more integrated data management system
- develop a more 'joined up' system of data management services at UQ
- offer ongoing training on data management and eResearch issues
- develop a UQ-wide data management/eResearch policy.

Accordingly, a series of four focus groups were run in late 2007 to get some insight into the above five areas. Groups were formed from people who had volunteered to be contacted after the survey and from other academics by invitation. A consultant was hired to run the groups and to present a summary report of the group's findings. Survey responses were probed in much greater detail, and attendees were asked for ideas on formulating a University-wide policy on data management. The questions asked of focus groups are listed in Appendix C, with the recommendations for action in Appendix D.

The focus groups reported several areas in data management that were of major concern. The vulnerability of existing data was a key issue – people saw threats to data from formats becoming obsolete, from migration to new formats, from hackers, from dropouts to service while accessing or manipulating data, as well as from poor or careless stewardship of data. The lack of training and support for proper data management, coupled with uncertainty about roles and responsibilities, meant that most projects operated independently and, in many cases, had to invent policy and procedures as they went along. This was viewed as very unsatisfactory. Many felt that the management and storage of research data should be identified as a key risk management item within the University's business plan, at the strategic level.

Many attendees complained of insufficient storage, and many were unaware if proper backup systems were in place within their organisational units. Inequity across disciplines was also mentioned, with the sciences feeling better served by the University than the humanities in the areas of storage and backup systems. All attendees felt the University should develop a stated policy on data management that addressed issues of networking, storage, control, access, release and re-use of data, and data integrity.

Attendees were keen for the University to provide templates that could be re-used, e.g. a template for a data management plan, a template for a negotiated access agreement, a template for the seeking of copyright release, and so on. A template that could assist in estimating the cost of long-term storage and maintenance of data so costs could be factored into applications was also high on the wish list.

Researchers were not keen to cede control of their data to a central body without very strong reassurance on issues of trust. These involved the central body guaranteeing that embargoes on release would be respected, that no improper access be granted, that privacy and confidentiality could be assured, and so on. Additionally, researchers wanted to feel sure that any central repository would be reliable, stable, able to manage their data without loss or corruption, able to make data discoverable and accessible while also keeping it secure and safe. Researchers also wanted, where possible, to be able to manage data retrieval in a self-service manner, without the need for IT intermediaries. Where IT help was needed, researchers wanted a service-focused support team who would proactively engage with and support them all year round. Researchers also wanted the repository to be able to be customised for different needs, as all disciplines would have varying requirements for their data.

Academics and support staff wanted training in what to do to comply with policy, should it be implemented. They wanted very clear advice on how to deposit data, including proper instructions, advice on protocols, clarity on timeframes – the whole "Who, what, where, why, when, how".

Researchers were keen for the University to manage centrally and advise on legal issues associated with research projects, such as copyright, privacy, intellectual property, ethical matters and obligations to funding bodies, industry partners and so on.

As a starting point, researchers felt the University should identify the instruments and laboratories that already generate enormous amounts of data, and identify the key groups with pressing data management issues. Their issues could then be managed first. Researchers wanted the University to create and disseminate 'best practice', discipline-specific guidelines and information about standards so that they could comply more easily with what is expected of them.

Researchers felt access to data would be best served by a 'tiers of access' system with different privileges granted at different levels, e.g.

- Open access for all
- Open access to research colleagues/researchers in the same field
- Mediated access (possibly available on request to selected researchers)
- Access only after de-identification/long time passing/to select groups
- No access

Researchers wanted access rights to be managed online, e.g. via a web interface, so that a single, centralised system could manage logins, allocation of privileges, restrictions, exclusions, validation and authentication.

Researchers also wanted means and opportunities to upgrade their knowledge and skills about internet tools and techniques for sharing information and/or collaborating with other researchers and colleagues.

Two other areas of concern were the need to protect the identity of researchers in contentious areas of research; and the need for the ARC and other funding bodies to recognise that funding needs to be provided if researchers are to be able to make their research data available for others to use and share. Currently, there is no incentive for researchers to jump this hurdle. While funding alone would not solve the issue, it would enable projects to hire suitably qualified staff to assist them in managing, depositing and making their data accessible. Many would like to see funding bodies mandate data digitisation and deposit. Most thought these mandates would come best from funding bodies, or from within universities themselves. Government was not mentioned as a major influence, though when prompted, researchers were happy for Government to supply a mandate, provided it came with funds that would allow researchers to comply.

To do all the above, all researchers were agreed on the need for strong policy and the right infrastructure to make it happen.

To stay updated, researchers wanted an eResearch tab in the regular weekly email newsletter, UQ Update, so that they could access the news if they wanted to, or ignore it. They wanted the option of RSS feeds for updates, as well as an alert or pop up on log in to advise that changes to the data management system had occurred since they last logged in. Most reported a feeling of information overload, so new blogs and newsletters were not welcomed, except in the area of new tools, technologies and techniques available to researchers for data sharing and networking.

The focus groups were a useful extension to the findings of the data survey. Focus groups went deeper into the issues and came up with useful suggestions for action.

Other comments

The running of the survey, and the subsequent focus groups and training sessions, were important outreach activities for the fledgling UQ eSpace repository. The growth in requests for presentations and consultancies has been striking in the wake of these events. The UQ eSpace repository has since become the official archival home of electronically deposited doctoral and research masters theses and theses for professional documents. It has also been the instrument in 2008 to gather citations and evidence for the annual Higher Education Research Data Collection for the Department of Innovation, Industry, Science and Research. Awareness of the repository and the services it can offer to staff has grown enormously. I am now working on another survey of staff regarding the collection and use of publications data. Recommendations from that survey will be presented to a University-wide working party on research publications and information management. Information from the data management survey will also be presented to that working party.

The survey helped UQ eSpace become a major player in the University's publications and data management systems, and has guided academics towards the advice that staff of the service can offer them. With the new working party in place, it is hoped that all the data gathered to date will help inform the work of that group, and help deliver a more integrated and coherent system for research data management at UQ.

Belinda Weaver

Queensland University of Technology

The uptake of web technologies, the application of advanced information and communication technologies and the ongoing developments in information technology and computer science have fundamentally changed the way that research is carried out. eResearch has enabled researchers to increasingly incorporate technologies, such as high-performance computing, data visualisation, computer simulations, high speed networks, distributed data storage and virtual collaborative environments into the actual investigation and discovery process. And this has meant that the skill set of a researcher has changed. Increasingly, researchers of all disciplines, not just the sciences, require the skills to be able to manipulate large data sets and transfer them across long distances on advanced networks.

The research environment at QUT has seen major advances in activity on a number of fronts over the last three years. As part of the focus on expanding its research capacity and performance, there has been the development of four institutes. The Institute of Health and Biomedical Innovation (IHBI) is the largest of the interdisciplinary research institutes. It comprises research staff and students from three faculties: the Faculty of Health, Faculty of Science and the Faculty of Built Environment and Engineering. As the flagship institute at QUT, IHBI seeks to provide a collaborative that is conducive to eResearch.

As Information Manager within IHBI, I was aware that researchers viewed data management as an important, yet an often unplanned element in the research process. Despite best intentions, it often happens spontaneously or on the fly, depending on resources available at the time of need.

In 2007, the National Health and Medical Research Council (NHMRC) in partnership with the Australian Research Council (ARC) released the Australian Code for the Responsible Conduct of Research.[4] The Code advocated best practices for researchers and provided advice on how to manage research data and materials, and presented separate data management responsibilities of the researchers and institutions. Of standout importance to the Library was the request (item 2.5.2) to researchers to make their data available where possible for use by other researchers (unless prevented by ethical, privacy or confidentiality matters). The expectation is that universities and research institutions will be responsible for providing robust and sustainable solutions, including the establishment of infrastructure and governance, at an institutional level for managing research data.

We were seeing that some publishers were taking steps to partner the publication record (e.g. journal article) with the relevant data by creating a persistent link from articles to the dataset/s. But we were aware that few researchers have the skills, resources and inclination to perform the tasks necessary to make their data not only available, but readily accessible and usable by others

[4] Australian Government, *Australian Code for the Responsible Conduct of Research*. Canberra, 2007. http://www.nhmrc.gov.au/publications/synopses/_files/r39.pdf

Government and funding authorities were establishing that research data must be accessible, discoverable, managed and long-lived, but before we could implement policy and plans to enable this, the current landscape of practices had to be investigated and mapped.

The Library had already been investigating using the open repository (currently ePrints) as a means of storing data alongside publications. We were aware that some researchers were already uploading their research data to websites, which presented archival and other problems. As a preliminary step in tackling the institutional and individual issues surrounding data management, the Library saw a need for an information gathering process on what practices were currently being employed by researchers to manage their data. We wanted to consider research support from the researcher's perspective.

We were aware that the University of Queensland was running a survey investigating data management practices as part of their involvement in the Online Research Collections Australia (ORCA) project with the Australian Partnership for Sustainable Repositories (APSR). We saw great benefits in running the same survey at QUT, in particular, the pooling of data from several universities. The more responses received, the greater the validity of the survey results.

Belinda Weaver, Manager University of Queensland eSpace, agreed that QUT could replicate the data management survey. Our only difference to their survey was the inclusion of three extra questions:

1. Who will be responsible for looking after the research data after the research project has concluded? Comments box for free text answer.

2. Who owns the data generated in your research?

- Yourself
- The research project team
- The University
- Don't know
- Other (please specify)

3. How do you know who owns the data?

- I was told
- There are contracts or policies
- It is just 'understood'
- It is covered in my employment contract
- Other (please specify)

An email was sent to University academics and research students. One hundred and twenty-nine valid responses were received to the Survey of Research Data Management Practices at QUT. There was a slight majority of academic staff (53.4%)

compared with postgraduate students (46.5%).

Results

The results of the survey confirmed our concerns about the lack of researchers' awareness of the importance of managing and storing research data and less than adequate data management practices.

Of particular concern were the following results:

- 84% of researchers have no data management plan in place
- 76% of researchers reported using USB/flash drives as one option for storing research data;
- 50% of researchers reported storing their data on CDs as on option; and
- 41% of researchers reported storing data on DVDs.

The survey also identified that researchers are taking ownership of the data themselves and have mixed understandings about ownership, length of time they should keep data and policy regarding data management.

The free text comments of the survey responses make a compelling case for an urgent response at an institutional level to meet the training and data storage needs of researchers.

> *The amount of video data that we generate in our research will become increasingly problematic to store using current methods as I anticipate we will produce significantly more than 0.5TB a year of raw video through staff and postgraduate student projects…*

> *The university is considering a central data repository, but this is only useful if the data placed there is well organised and is accompanied by metadata which includes the (instrumental) conditions under which it was obtained. For example, spectral or diffraction data tends to be specific to the instrument it was measured on*

> *I realise I don't really know what I'm doing because this survey has raised a lot of questions that I hadn't thought about. I would certainly get a lot out of training……*

The survey results confirm that the management of research data is not simply a matter of providing the infrastructure. Researchers have concerns about loss of control over their data, the reliability of central systems and training.

Focus Groups

Based on survey responses to willingness to participate further on data management issues, two focus groups and several interviews with individual researchers were conducted.
The major issues raised through these were:

- Ownership of the data. There is currently lack of clarity as to who owns it. There tends to be a correlation between the researchers' personal input (time and value) in collecting the data to the ownership of that data. For example, a social scientist who has spend nights accompanying police standing by the side of the road collecting drug-driving samples is more likely to feel justified in a greater claim of ownership over their data than a biomedical scientist who is focusing on knowledge discovery through data-mining of large-scale genetic, genomic and/or proteomic data.

- The sharing and publishing of research data. There are two essential reasons for making research data publicly-available: i) to make the data part of the scholarly record that can be validated and tested; ii) so that the data can be re-used by others in new research. Issues discussed by researchers on sharing data included:
- The possibility of users signing off on an agreement or licence to download data.
- The development of guidelines on the authorship of publications if data is shared.
- What incentives and motivations could be developed to encourage sharing?

Whereas some researchers are motivated to share their data by altruism, encouragement from senior peers, or new collaboration opportunities, there is currently a lack of explicit and tangible rewards to do so.

- Quality assurance in storing and describing data (including metadata). Many researchers are not aware of the term metadata, although they often use their own discipline-specific practices in describing data.
- A consistent approach to backup across the university is required. Currently, different faculties and institutes have different approaches and there is duplication of data stores.
- With the deluge of data comes the consequential need for training in data management skills and practices. As with most people, researchers are time-poor and would be grateful if such training could be offered in several ways (e.g. online plus face to face).

The Future

It is clear that future research relies on skills and capabilities that are associated with eResearch tools and techniques. As a university seeking to enhance its research profile, QUT is looking to deliberately develop mechanisms for improving the prospects of our current and future researchers.

A major outcome of the Survey of Research Data Management Practices at QUT has been the development of the eResearch Support Service Project which commenced in April 2008. The project also responds to the Division of Technology Information and Learning Support (TILS) Research Support Plan and proposes to develop a flexible and sustainable eResearch Support Service model that addresses the shortcomings of current practice and meets the short term and longer term needs of researchers for data storage, backup, expert advice in software for analysis and manipulation, visualisation, simulation and data management and access. More specifically it will involve:

- a single central service point for expert advice and service on eResearch support including data storage, management, access and HPC services
- engaging with new eResearch tools
- ongoing investigation in determining future data needs of QUT researchers
- developing strategies to address the current skills gaps in data management
- providing infrastructure and service requirements for research data management storage and services that meets those needs
- drafting a QUT Research Data Management Policy and Plan

There are significant challenges ahead for universities as they develop the infrastructure and support services necessary to facilitate researchers working in an eResearch environment. The management of research data is recognised as one of those key challenges.

The Survey of Research Data Management Practices at QUT has been the valuable first step in helping the Library and Division of TILS grab a snapshot of the current data management practices of researchers at QUT. In using the same survey as the University of Queensland and University of Melbourne, we were able to confirm QUT researchers are not alone in their concerns, practices and needs relating to data management. Other researchers are facing the same data management issues, using similar less than adequate data management practices, and are seeking guidance and training on what to do.

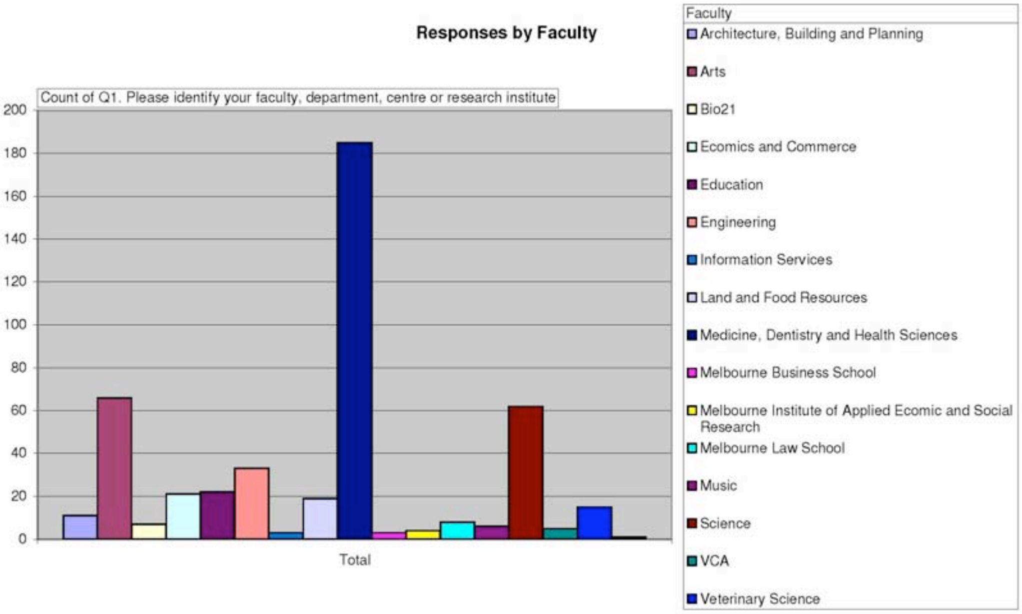
Overall, the results have been insightful and have formed the foundation of a progressive and exciting eResearch Support Service Project, whose chief objective is to provide researchers with a single point of service and expert advice on eResearch support issues including data storage, management and access.

Stephanie Bradbury

The University of Melbourne

In addition to the valuable analysis referred to in this report, the Research Data Management and eResearch Practices survey was important at the University of Melbourne because it marked a point of significant engagement with the broader University community about tangible issues concerning eResearch. Coinciding with, and internally initiated by the University's recent appointment of a Director of eResearch, the survey provided an opportunity to establish a baseline around research data management processes, and also a level of benchmarking with the other universities that conducted the survey.

The Survey was sent out to all researchers on behalf of Professor John McKenzie, Deputy Vice-Chancellor (Research), Linda O'Brien, Vice Principal, Information and CIO, and Professor Leon Sterling, Director of eResearch. Four hundred and eighty-three researchers responded, a healthy response rate given the voluntary nature of the survey. The breakdown of respondents across the university is outlined below.



Once the survey had been completed, results were broken down by faculty. Over the next 6 months, the Director of eResearch held face to face meetings with senior faculty representatives, including Associate Deans of Research, Research Managers, and senior Directors and Heads. The information was presented as 'their information', and played a pivotal role in facilitating a conversation about data management practices in each faculty. Some of the conversations helped identify new collections of significance; other conversations simply started the ball rolling. In general, the feedback from these meetings reinforced the messages we received from the survey. The thirst for information and guidance on strategies to manage research data, was greater than the currently available information.

One outcome of the survey was recognition of the need to develop training materials on Research Data Management. The first of an ongoing series of graduate training courses on this issue was presented in the first quarter of 2008.

Simon Porter

Conclusion

There can be little doubt that conducting this survey has been of immense value to the Universities concerned. Each has been able to utilise the results to progress their own internal planning and as a basis for further discussions with researchers about their needs.

As mentioned earlier in the report, there was a notable consistency across the three Universities in their results. This consistency tentatively allows us to extrapolate from these three Universities to the Australian researcher population in general. If we do this, then we can generalise the findings as follows:

- Nearly all researchers have digital data created in the course of their work
- The few researchers who claim not to have digital data either genuinely do not have any or, more likely, do not recognise what they have as digital data, probably because it is text
- There is an extraordinary range of non-digital data being collected, with implications for the need to digitise at some future time
- The size of a digital collection does not seem to be of importance to most researchers, except for those who have significant data storage requirements
- Researchers use a wide range of software, largely proprietary. The range is made up of a small number of core applications and a much larger number of specialist applications. This has implications for later data curation
- Researchers currently do not recognise the implications of their software choices and later access to their data
- Researchers use many different means of storing and backing up their data, often using storage media which are unreliable and short-lived
- Most researchers do not have research data management plans, although they do recognise the need for them
- Training is sought in areas related to data management planning, either prior to a project or after, digitisation and data rescue (for older materials)
- Most researchers are responsible for their own data management which may vary from haphazard to highly organised
- Most researchers are willing to share their data and in many cases already do so. They would like an easier means of doing so
- Most researchers see their data as having value beyond the immediate project
- Only a small proportion of researchers use the grid or high performance computing. Many more researchers consider themselves as conducting eResearch, suggesting that the two are not necessarily connected
- Most researchers are unclear about the intellectual property regime which governs their research
- There is some negativity among researchers about data management, which is seen as another bureaucratic requirement being imposed on their time. This

suggests that there will be resistance in some quarters to any change to the current regulatory environment

- Differences exist between disciplines in their approach to data management, with the Humanities & Creative Arts being least organised and the Social Sciences the best organised

These statements are very general and do not take into account the considerable variation between researchers, some of whom clearly operate at levels of best practice while others do not. Altogether, however, these finding suggest that there is much to be done to improve the training, support and technical infrastructure needed to bridge the gap between the current conduct of research and meeting the potential offered for eResearch by the emerging ICT environment. These findings also suggest that there is considerable value for research institutions taking a closer look at researcher practice to identify the gaps and take remedial action where required.

Contributing to a national data set

Any institution wishing to conduct this survey so that its findings can be added to the national data set should send an email to contact@apsr.edu.au.

Appendix 1: Questions

The questions included in the survey are listed below. They are not numbered for two reasons. Firstly, the order in which they were listed varied from university to university and secondly, because some questions were asked by only one university (these questions are marked with an asterisk). A list of tables with both statistical and text responses is provided in "Data Management Practice Survey - Data tables & responses" which is included with as a separate document.

- Please identify your school, centre or research institute

  [text response only]

- Academic status

  Member of the academic staff
  Postgraduate student
  Emeritus/Adjunct appointment
  Other (please specify)

- Has your research generated digital data?

  Yes (go to question on types of data)
  No

- If no, do you maintain research-related data in non-digital forms such as paper, photographs, video or audio tapes, slides, etc?

  Yes
  No If your research generates digital data, please check all the following types that apply:

  Data automatically generated from or by computer programs
  Data collected from sensors or instruments
  Experimental data
  Fieldwork data
  Laboratory notes
  Images, scans or x-rays
  Web sites
  Blogs or discussion threads
  Email
  Digital audio or video files
  Documents and reports
  Other (please specify)

- How large (in total) is your digital research data?

  Less than 100MB
  100MB - 1GB
  1GB – 1TB
  More than 1 TB
  Don't know

- Please list any software used for analysis or manipulation of your data, e.g. SPSS, TecPlot

  [text response only]

- Do you use Grid/high performance computing?

  Yes (go to question on types of data)
  No

- * If applicable, how do you store and retain any software used to generate your research data?

  [text response only]

- Do you currently have a formal Research Data Management Plan in place?

  Yes
  No

- What data storage and backup system do you currently have in place? Please check all that apply:

  DVDs
  CDs
  USB/Flash drives
  Tape storage
  Storage area network
  Offsite storage
  Third party (including commercial data storage)
  None
  Don't know
  Other (please specify)

- Who is currently responsible for managing the data?

  Research project manager
  Designated person on project
  External project partners
  ITS
  IT staff within your school, centre or research institute
  Research assistant
  Yourself
  Nobody
  Don't know
  Other (please specify)

- Do you allow researchers outside your team to access your research data? Please check all that apply:

  Openly
  Via negotiated access
  Only after the formal end of a project
  Only some years after the end of a project
  Not at all
  Never, because of privacy and confidentiality issues
  Not at present, but I would be willing to make some or all of it available if an easy mechanism to do so were offered at [this university]
  Access is provided through the Australian Social Science Data Archive (or similar) after data is deposited there

- * Who will be responsible for looking after the research data after the research project has concluded?

  [text response only]

- How is your data accessed or used? Please check all that apply

  In original print form
  In small chunks
  Dataset as a whole
  As raw data

Only after filtering, manipulation and access
Locally
Online via Grid, Storage Resource Broker, etc
Online via a website or service
Other (please specify)

- How long do you think your research data will have value?

  Up to 5 years
  Up to 10 years
  More than 10 years
  Don't know

- * Who owns the data generated in your research?

  [text response only]

- * How do you know who owns the data?

  [text response only]

- Would you be interested in training or advice on any of the following? Please check all that apply.

  Digitisation advice, tools and services
  Creating a research data management plan at the beginning of a project
  Creating a research data management plan after a project has finished
  A data "exit" plan (for retiring academics or departing academics and postgraduate students)
  Data "rescue" for older digital materials, such as data on older media or migration of data from legacy systems
  Other (please specify)

- Would you be willing to participate in, or provide information to, an eResearch reference group aimed at developing support for researchers at [this university]?

  Yes
  No

- Please feel free to add any other comments regarding data management, long term data storage and access, digitisation, training, etc.

  [text response only]

- * Do you regard yourself as practising eResearch?

  Yes
  No

- Your name?
- Your email address?
- Please tick here if you would like to discuss your data management and storage or training needs with staff from [this university's relevant area]

* Asked by one University only. Otherwise by all three universities taking part

Appendix B: Software used only once

@Risk
Aabel
Abaqus
ABI
Accelrys Materials Studio
Acknowledge
AFNI
After Effects
Agilent data analysis
Agilent Genespring
Agilent MassHunter
Agrobase
AmplifX
Amplify
AMWIS
Analysis Chart
AnalySys Imaging
Analyze
ANSI-C
ANUCLIM
Aquapak
Arc
ARC Map
ArchiCAD
ArpwArp
Aspentech software
ASREML
Asylum Research software
Audacity
Audiamus
authorwiz
AutoMontage
Avid
AVS
Awk
Axis
Axoscope
BASF
BEAST
Beckman-Coulter flow analysis software (CPX)
Bioanalyst
Bioconductor
Bio-D
Bionumerics
Bio-Plex Manager (Bio-Rad)
Blackboard
BLAST
Bodybuilder
Boilerhouse
Borland C++ Builder
BSCW
C
CAAT-Box
CAIC
Candid
Canon Image management tools
Canvas
CAP
CARET
CBT Data reconstruction
CEQ
CFX
Chart
ChemDraw
Chemi-Capt
chemoffice
Chenomx
chimera
Chromos sequence analysis
CLANS
ClinProTools (Bruker)
Clustal
CNS
CodonCode Aligner
Combustion
Concept Systems
Confocal
Continuous
Coot Corbett RG
C-Plan
CRC Centric (SCRP)
CricketGraph
Crimson Editor
CrystalClear
csh
CSS
Cyana
Cytel Studio
D for LSM
Deltagraph
dFdr
Digitool
DNA strider
DPlot
DreamWeaver
DTI- Studio
DVR
eCognition
EcoTect
EHRs
EMAN
EMU
epidata
ePrints version
EthoVision
Explorer Vensim
Extensis
Final Cut Pro
Finale
Firefox
Forest-DNDC
Freesurfer
FSL
GAMS
Gatan Digital Micrograph
geldoc software
GenBank
Gene Mapper
Genepix Pro
Genepop
Geomagic
geomodeller
Gimp
GMT
gocad
GOLDMINE (CCDC)
Google Analytics
Grace
Graphics
grep
GROMACS
GROMOS
Heritage Documentation Management System (HDMS)
Hg
HKL
HRV
Idiogrid
Image
Image Pro
ImaGene
IMageReady
ImageScan
ImageTool
Imagine
Imaris
imovie
Informax VNT
iNMR
Inspiration
invitrogen vector NTI
Irfanview
IRIS
Isoplot
ITracker
iTunes
J
jasco
jgraph
JK INfo Manager
JKMetAccount
JKSim**
Keynote
Kirrkirr
LCS Lite
LEGINON
LexiquePro
Lexus microscope software
limdep
Linux
Logger Nett
LSM image Browser
Lview Pro
MacLab
MacLigand
Map Manager
maple
MAPMAKER
Marxan
Materials
matplotlib
Maxima
Mayo Clinic's Analyze
MCDfit
Mediawiki

- Mendel
- metaFluo
- MetaMorph
- mGrace
- Microarray
- Microcal Origin Graphing
- Mighty EDF
- MIGRATE
- MINC
- Mindmanager
- MJ Research
- MLA DAta Analysis
- Mokey
- Molecular Sophe
- MolMol
- MPLUS
- Mr BAYES
- MR images
- MRICro
- MSPWIN
- MSR Sense
- mtvplot
- Multicalc
- MVSP
- Navicat
- NESSTAR
- NetDrawer
- Netlogo
- Netminer
- Neuron
- Neuroscan / Scan
- NIH
- NIH image
- NLKT
- NMRView
- Numpy
- Obzerver
- Oligonucleotide Calculator
- Omnic
- Online Heritage Resource Manager (OHRM)
- OpenDX
- Opticon
- Optimas
- ORTEP
- OxMetrics
- PaintShop Pro
- PAJEK
- paravision
- Pathway Studio
- paw
- PcGets
- PcGive
- PcOrd
- PDFFit
- PDFGetX
- PDQuest
- pgplot
- Pharmacokinetics
- Photo Editor
- Picasa
- PLS Graph
- PopTools
- Portfolio
- ProteinPilot
- Protools
- PSI-Plot
- putty
- Python
- QDS
- QSR
- QualBrowser
- quedm
- quest
- QuickTime Pro
- Ranges
- RAP
- RapidForm
- RapidReader
- RasMo
- RATS (Regression Analysis for Time Series)
- RCel
- REBEL
- recordpad
- Refman
- Resonanz
- revman
- Rheowin
- RotorGene
- RUM
- Runtime Revolution
- runZ
- S
- SAAM
- sage
- SaTScan
- Scientific Workplace
- screenworks
- SDS
- SEDNTERP
- SEDPHAT
- Sequence Manipulation Suite
- SerialEM
- Shake
- Sharp
- SHELX
- signal
- SIL
- Silver (CCDC)
- SimaPro
- SIMCA-P
- SimpleText
- simpson
- Slicer
- SlideWrite
- SM
- SmartPLS
- Solve/Resolve
- soundstudio
- SpaceTimeResearch
- Sparky
- SpectaSuite
- Spot
- spotlight
- Star-P
- Stat Transfer
- Statistix
- stat-transfer
- STRUCTURE
- Studio
- SUDAAN
- SuperANOVA
- SuperCross
- SuperServer
- SuperWeb
- Surfer
- surveymaker.com.au
- Swiss-PDBviewer
- Syngene GeneTools
- Tableau
- TeachText
- TeXShop
- Tiffcp
- Tilia
- Toolbox/Shoebox
- TOPSPIN VNMRJ
- TPS Dig
- tRNA-SCAN
- TRNSYS
- TSP
- Ultraedit
- VBA
- Video Virtual Earth
- VisIt
- Visual Basic
- Visual MODFLOW Pro
- Visual Studio
- VNMR
- vpmg
- Wallac manager
- WAUTER
- Wavepad
- WebBrowser
- Wiki
- WinBugs
- WinCATI
- WinCurveFit
- Windows
- Windows movie maker
- WinEPR
- WinMDI
- Winsteps
- WSXM
- www.fil.ion.ucl.ac.uk/spm
- xanim
- XCalibur
- Xcrvfit
- XEASY
- XeprView
- xfig
- xForms
- XLGRAPH
- xmgr
- XML tools
- Xplor
- XPLOR-NIH
- X-win
- XwinNMR
- ZeissMetaPhotonLaser Confocal Software

Appendix C: eResearch Focus Groups at The University of Queensland

The first two eResearch focus groups responded to the following questions:

1. Do you know about UQ eSpace [UQ's institutional, digital repository]?
2. What do you use UQ eSpace for, e.g. publications, private data collection, data deposit?
3. What improvements could be made to data management at UQ?
4. If UQ developed a stated policy on data management, what should it include?
5. What facilities and tools should be available to make information and data sharing with colleagues easier, e.g. teleconferencing, access grids, shared working spaces, shared document creating tools, private data sharing?
6. What kinds of access would you like to be able to grant researchers to your research data?
7. Who should manage this access?
8. What kinds of e-research or data management advice or training would you like to have available – e.g. grants compliance, legal requirements, copyright issues, standards?
9. If UQ developed a stated policy on e-research, what should it include?
10. How would you like to be kept up to date with developments in e-research at UQ?

The refined questions used for focus groups three and four were:

3a) In a perfect world, how would you like to use and access data?
3b) What do you perceive is the biggest threat to your data?
3c) If you had the capacity, what are 3 key things that would boost academics' confidence in a centralised data management system?
3d) In a perfect world, from an academic's perspective, how would you like data to be controlled?
4a) With regards to a stated policy on data management, what are the key behaviours that you want to see addressed?
5a) Please provide a scenario/example of how you would like to engage in information and data sharing with your colleagues
6a). Imagine that your research data has the appropriate level of security to meet all ethical and legal requirements. Who should manage others' access to your research data? For how long and why?
8a). How could grants compliance, legal requirements, copyright issues, standards, etc. become embedded into your research environment, as opposed to being "external issues" that you have to deal with?
10a). How can we minimise information overload and communicate with you at the "right time"?

Appendix D: eResearch Focus Groups at The University of Queensland

Recommendations from UQ focus groups

1. A standardised template to be created for researchers to complete when applying for funding. (This will enable an accurate estimate of space and cost associated with the storage of data and information requirements for that research project)
a. Dedicated support services to be made available to all researchers: to ensure that all hardware and software requirements for each researchproject are budgeted for in the early stages of the research project
2. All legal issues associated with research projects are centralised by the university in consultation with the Lead Researcher. This includes but is not limited to; copyright, privacy and ethical matters and funding obligations.
3. Creation of a "Google style" search engine for all data and information that is held within the UQ Intranet.
4. Implement a Change Management Program across the technology support groups – the outcome being a service focused support team who proactively engage with and support the researchers.
5. Survey all researchers within UQ to identify the type of primary data they are currently holding including; file type, size of file and other key information (this style of survey would provide the basis for developing a concept framework for the management and storage of existing and future research data)
6. Develop and manage a centralised system for all research data and information. (This system will have a "sliding scale" that allows for the different types of research data and information i.e. the rules and protocols will vary across the various fields to allow for sensitivities)
a. Explore options for on-going storage and management of primary data as software and hardware is upgraded (i.e. migration issues)
b. Explore data repositories for easy to use storage and retrieval of digital images.
7. Put in place a communication "blog" that highlights the different tools and techniques available to researchers with regards to internet style technologies.
a. Create and communicate opportunities for researchers to upgrade their knowledge and skills with regards to internet tools and techniques for sharing information and/or collaborating.
8. Lobby for the management and storage of research data to be identified as a key risk management item within the UQ business plan; at the strategic level.