This paper was presented at the conference:

Oxford eResearch 2008

11-13 September 2008

University of Oxford

Conference website:
http://www.oii.ox.ac.uk/microsites/eresearch08/index.cfm

Conference papers collection:
http://ora.ouls.ox.ac.uk/objects/uuid%3A64aa6f39-7e81-4d42-a008-ee2d7524bd67

Conference organisers:

Oxford Internet Institute
University of Oxford
1 St Giles, Oxford OX1 3JS
http://www.oii.ox.ac.uk/
email: enquiries@oii.ox.ac.uk

Oxford e-Research Centre
University of Oxford
7 Keble Road, Oxford OX1 3QG
http://www.oerc.ox.ac.uk/

"We have just entered the Wright Brothers phase of the digital era": data management and practice in the eResearch era

Belinda Weaver
Manager, eResearch Infrastructure
The University of Queensland Library
b.weaver@library.uq.edu.au

Margaret Henty
National Services Program Coordinator,
Australian Partnership for Sustainable Repositories
margaret.henty@apsr.edu.au

While supercomputing facilities and high speed internet connections are crucial components for successful eResearch, also important are academic culture, institutional capabilities and individual skills if we are to optimise our potential.

In the second half of 2007, three Australian universities conducted surveys of data management practices among their researchers and followed these up with more detailed investigations.

The consistency of results suggests not only that researchers share many common practices and attitudes, but that the hurdles they encounter are also common. Overall, the surveys showed that researchers are more focussed on research outcomes than on good research management practices, especially when it comes to the long term stewardship of data.

Introduction

Universities around the world are looking to improve the infrastructure that supports their eResearch. However, infrastructure means more than just high quality ICT. While supercomputing facilities and high speed internet connections are crucial components for successful eResearch, also important is the availability of trained technical staff to support researchers in both complex applications and networking. Sound data management practices are also vital to ensure that all research project data is both readily accessible to researchers on demand and safeguarded for the long term, so that the data can be revisited by the original research team, or eventually made available to other researchers.

Academic culture, institutional capabilities and individual skills are critical components if we are to move towards a more effective research system where the potential for eResearch can be optimised.

The Data Management Survey

In the second half of 2007, three Australian universities undertook a survey of Data Management Practice. The University of Queensland (UQ) took the initiative in designing and implementing the survey among its own researchers and postgraduate students. The University of Melbourne (UM) and the Queensland University of Technology (QUT) followed on with surveys of their own based very largely on the UQ survey, but each with two extra questions of particular local interest. The Data Management Survey questionnaire was completed by a total of 879 academics,

postgraduate students and adjunct appointees who ticked boxes and offered suggestions, comments and, occasionally, criticism.

Each of the three universities had become actively involved in improving support for eResearch within their institutions. Improving data management practice was seen as integral to this, as was the need to provide training, and to engage more effectively with the research community on data management issues.

The use of common questions and survey-gathering software across all three institutions enabled the aggregation of results, which have now been fed into national eResearch planning processes. A full report of the survey is now available, as are the data tables on which the report is based.

A note on terminology: in this survey, and in Australia generally, the term eResearch is used in the same way as eScience is used in Britain.

So what did we discover?

A significant finding when all the results were aggregated was the remarkable consistency in the results across all three universities. This suggests not only that researchers share many common practices and attitudes to their research, but that the problems or hurdles they encounter are also common, regardless of the institutional setting.

Overall, the surveys showed that researchers are more focussed on research outcomes than on good research management practices, especially when it comes to the long term stewardship of data. This is not unexpected, but it confirms the need for doing the survey and putting measures in place to support eResearch.

Digital data

It came as no surprise that over 90% of respondents from all the universities taking part said that they had data in digital form. When asked to nominate what forms their digital data took, the list was long and is shown in Table 1.

Table 1: 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

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."

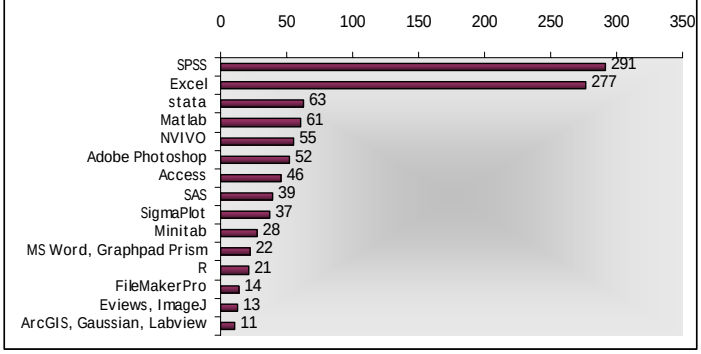
The University of Melbourne asked additional questions about research practice. Only 6% said that they use grid or high performance computing and 28% said that they regarded themselves as practising eResearch. Definitions of eResearch (or eScience) are many and varied, but tend to focus on the use of high capacity computing. Clearly there are many who regard their own research as eResearch without it. While the non-response rate to this question was not high (less than 2%), there were many comments indicating some confusion about the term eResearch and what it actually means.

Perhaps more surprising in these digital times was the 10% who said that they had no digital data. They did, however, have a wide variety of data in paper or analogue formats (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) which raise questions about the need for later digitisation, and in some instances the need for improved data collecting techniques.

Software

Our 879 respondents mentioned a total of 514 different proprietary softwares which they use in the course of their research. Only 19 were mentioned by more than 10 respondents, which shows what a remarkable concentration there is on some popular software. The 372 softwares mentioned as being used by only one respondent shows the extraordinary range in use. The most popular were mainly associated with database creation and statistical analysis – perhaps not unexpected. From the point of view of persons responsible for data, one curation (or stewardship) issue is this wide range of proprietary formats they will need to manage and sustain.

Figure 1: Software in use by more than 10 respondents



Research data management planning

There has been an increasing emphasis on the provision of data and datasets to back up research findings, with some disciplines requiring that data be made available alongside publications and/or for subsequent verification or re-use. This has in turn

led to requirements for data management plans to be developed as part of the research planning process. The survey results show, however, that we have some way to go before all research data planning is put on a more organised footing.

"Do you have a research data management plan?" was one of the questions in our survey and the answer was a resounding "no" from 82% of respondents. Preparation of a research data management plan is currently not mandatory in any Australian university, although there are suggestions that funding authorities may in future require a structured plan as part of the funding application process.

It might be expected that there would be disciplinary differences in data management planning. Current opinion would suggest that those in science disciplines would be more attuned to the need for good data management than those in the humanities and creative arts. In general, our results did not show this to be the case, as the largest proportion of those with data management plans were in the Social Sciences (26%), and Medicine and Health (21%). Business and economics came next on the list with 15%, IT and Engineering and Architecture were on 13%, Science on 11% with the Humanities and Creative Arts only marginally below that. A number of people in Law were excluded here as they were so few.

Data Storage and Backup

We wanted to know how researchers are currently backing up their data, and the results revealed some cause for concern. There was a heavy reliance on media such as USB drives or portable hard drives which, while convenient, are not reliable enough for archival storage. Over 65% used USB/Flash drives, but this may have been in association with other media as many respondents indicated that they used more than one means of backup. Other popular forms of backup included CDs (56%), DVDs (39%), or Storage Area Networks (39%). More reassuring was the finding that fewer than 1% had no means of backup at all, although a worrying 3% replied "don't know".

The many comments accompanying the responses contained grumbles about the inadequacy of the current local 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 occasional scathing calls for more up-to-date practices within the university concerned. It was often apparent that there was confusion about where to go for advice.

Responsibility for data management

The survey asked the simple question "Who is currently responsible for managing the data?" The answers were complex, sometimes confused, and revealed a wide range of data management practices, from the highly sophisticated and professional through to the ad hoc and possibly risky. Most people indicated (78%) that they are responsible for their own research data. This proportion might at first seem high, but the comments provided show that this was not seen as being other than expected. Research students, in particular, manage their own data. 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 that data on a day to day basis.

Other groups reported that responsibility was held by the Research project manager (20%), IT staff within the school, centre or research institute (20%) or a designated person on the project (19%). 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 often 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%).

On the positive side, only 1% said “nobody” and 1% “don’t know”.

Looked at by discipline, Medicine and Health seemed to be the best organised when it came to having designated responsibilities for data management, to having the lowest proportion managing their own data, and the largest proportion with a designated data manager. Standing out among the disciplines were the Humanities and Creative Arts. They had 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”.

One of the universities asked an additional question about responsibility for the data once the project was complete. Here the answers showed much more variation, with a much higher proportion either not responding or indicating that they had never thought about it.

Sharing and valuing data

The survey revealed that many researchers would be willing to share their research data, whether raw or derived, with other researchers provided they were provided with an easy mechanism to enable this. Their current practice shows that 44% make their data available to others via negotiated access and nearly 10% make it openly available to everyone. Less than 20% don’t make it available at all, most often for reasons of privacy and confidentiality. Some make it available only after a specific time period has elapsed. Another group, of 13%, said that they don’t make it available at present but would be willing to do so if an easy mechanism to do was available. Most disappointing was the fact that fewer than 1% make their data available via a recognised data sharing archive such as the Australian Social Science Data Archive. Some respondents pointed out that, in some cases, it is necessary for data to be made available together with journal publication, a requirement we are likely to see more.

The permutations and combinations of what might or might not be made available, and when, were many. Some commented that their data would be meaningless to outsiders and others said 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. Some researchers provide access on a project by project basis, where some data might have to be destroyed. 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. Some mentioned the need for a central data repository, apparently unaware that these are often already available within the universities for which they worked.

Allied to this issue was that of the lasting value of the data because planners would like to know how long they will be expected to keep data. Most were prepared to make an estimate, with 28% suggesting that their data would be of value for up to five years, 27% suggesting up to ten years and 30% suggesting over ten years. Fewer than one-fifth said they did not know.

Data ownership

There was some confusion shown about who owns the data generated in research.

The Queensland University of Technology included two questions related to copyright and data ownership to get some idea of the level of understanding researchers have about what can be a puzzling issue. The results showed a range of views about copyright ownership, with 45% saying that the university owns copyright, 43% saying that the researchers own it and 27% saying that ownership rests with the research project team. Over one quarter (26%) acknowledged that they did not know, and a further 13% ticked "other". Among the other copyright owners 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 [Australian Bureau of Statistics] and ATO [Australian Taxation Office] with our layers of value add".

A second question asked respondents how they know who owns the data. Responses to this fell roughly into two groups. One group could confidently assert that there are contracts and policies which cover such matters (14%), or who thought that this is covered in their employment contract (27%). One respondent even went so far as to quote the relevant subsection of the university's IP policy on intellectual property created by students. The other group of responses came from people who replied "I was told" (79%) or that "it is just understood" (36%). This group seems to lack the certainty of the first group. Comments such as "I don't know 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.

Training

Researchers were eager to find better ways to manage their research data, with many expressing a desire for more training and direction in data management practices. 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%) or after a project has finished (22%). Another large group (33%) wanted a data "exit" plan, a topic designed for researchers who might be retiring or leaving the university or completing a postgraduate degree.

Help with digitisation was also keenly sought, with nearly one-third of respondents seeking this (31%). 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, either because they run on obsolete equipment or because the data needs to be migrated from old media such as 5.25 inch floppy disks or magnetic tape, or from proprietary formats (23%).

Many other training topics were mentioned, although no one topic stood out. Suggested topics included collaboration in developing international research datasets, intellectual property, managing data as records and integrating data with other research records and information, archiving data from a longitudinal data set, organising databases for medical imaging, MySQL for managers, data manipulation of video action research to protect anonymity, sub-version control software, enhanced data sharing, and networking of data storage and analysis applications.

Many of the comments were very supportive, e.g.

- *Any form of training for [department] members would be useful, I think.*
- *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.*

The demand for training was high, but countered by reluctance on the part of some respondents to be engaged in what was seen to be a further imposition on research time. This indicates that the cultural and organisational change required to improve data management practices will encounter some opposition along the way.

- *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.*

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 their university. Just under one-half of the respondents indicated their willingness to be part of this initiative (45%). Some commented that only a lack of time would prevent them from taking part, or gave other valid reasons.

Follow up activities at the University of Queensland

Responses to The University of Queensland's survey were designed to feed into future planning for eResearch infrastructure provision via an eResearch Working Party which collaboratively drafted the original survey instrument. The University of Queensland followed the survey with a series of focus groups. Groups were made up of individuals who had indicated in the survey their willingness to participate in an ongoing eResearch reference group. Survey responses were probed in much greater detail, and attendees were asked for ideas on formulating a University-wide policy on data management.

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.

1. Open access for all
2. Open access to research colleagues/researchers in the same field
3. Mediated access (possibly available on request to selected researchers)
4. Access only after de-identification/long time passing/to select groups
5. 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.

Conclusion

“We have just entered the Wright Brothers phase of the digital era” was a comment from one of our survey respondents. It is an interesting observation and points to an intriguing comparison. As everyone knows, Orville and Wilbur Wright demonstrated the world’s first operational aeroplane, heavier than air and with controls, in 1903. Less well known is the systematic methods they had adopted to get to this point. Prior to 1903, they concentrated on solving the underlying problems which had prevented everyone else from flying. They undertook wind tunnel tests to provide data which helped them to improve such devices as wings and propellers. They experimented with other mechanical devices, notably bicycles.

The analogy of the Wright brothers provides us with a model for improving data management. Knowing more about the data life cycle, current management practices, and the environmental and technological obstacles that currently exist will help us to improve data management practice and infrastructure to support future eResearch.

References

Henty, M. (2008) *Data Tables from the Data Management Practices Survey*. [online] Available at http://dspace.anu.edu.au/handle/1885/47108 (July 15, 2008)
Henty, M., Weaver, B., Bradbury, S., Porter, S. (2008) *Investigating Data Management Practices in Australian Universities.* [online] Available at http://www.apsr.edu.au/investigating_data_management (July 21, 2008)