Developing the Capability and Skills to Support eResearch

Margaret Henty provides an Australian perspective on improving the environment in which eResearch is conducted through developing institutional capability and providing appropriate skills training.

Introduction

The growing capacity of ICT to contribute to research of all kinds has excited researchers the world over as they invent new ways of conducting research and enjoy the benefits of bigger and more sophisticated computers and communications systems to support measurement, analysis, collaboration and publishing.

The expanding rate of ICT development is matched by the numbers of people wanting to join in this funfest, by growth in the amount of data being generated, and by demands for new and improved hardware, software, networks, and data storage. Governments and research funders, too, are keen to exploit the potential for new discoveries which may bring societal benefits and a return on their financial investment.

There is some way to go to bridge the gap between the potential on offer and the realities with which we are living. Of particular concern to this article is the need for improved levels of data stewardship to enable good data management for long-term sustainability, both at national and institutional levels. Overall there is a need for skilled personnel to be employed in an organisational environment which enables and facilitates the research agenda. The concern is being voiced worldwide. In late 2006, in Australia, a report issued by the Prime Minister's Science and Engineering Innovation Council (PMSEIC) stated that: 'The Committee found that adoption of e-research methodologies in Australia is constrained by a shortage of skills in e-research methodologies, including information management and curation skills.' [1]

In the United States, the National Science Foundation’s report, *Cyberinfrastructure Vision for 21st Century Discovery*, outlines one goal of the National Science Foundation as being 'To support the development of new skills and professions needed for full realisation of CI-enabled opportunities' [2]. In the United Kingdom, this sentiment is further echoed in a recent consultancy report by Liz Lyon, Director of UKOLN, which offers the restrained comment that: 'The awareness and skills levels of researchers regarding data management is variable.' [3]

Terminology

It needs to be noted at this point that there are terminological differences used in various countries to describe these new methods of conducting research and the digital environment in which research is taking place. In the UK, we see use of the term eScience, in Australia, eResearch, and in the USA, the underpinning environment is described as Cyberinfrastructure.

Background to the Study

The issues of eResearch and data stewardship have become increasingly prominent in the past couple of years. In Australia, there has been considerable investment in developments to encourage eResearch, improve infrastructure at national and institutional levels, encourage open access to both data and publications and improve the take up of institutional repositories [4].
The Australian Partnership for Sustainable Repositories (APSR) was one of the projects supported by the Australian Department of Education, Science and Training through its Strategic Infrastructure Initiative. The APSR brief included training and outreach activities, which found a large and appreciative audience. APSR also conducted a number of surveys which have added to our understanding of researcher practice and repository management [5].

My own background is in social science research and librarianship and I have published extensively in library research for over twenty years. I was prompted to undertake this study by a perceived need to understand better where the gaps lay between the ideal and current practice, with the expectation that this information could be fed into planning for the new Australian National Data Service (ANDS). My intention was to concentrate on the need for skills development and it was only as the study progressed that it became clear that skills development is only part of the answer.

I therefore took on the role of investigating the issue of skills and capabilities through the Skills for eResearch Project. This had as its aims: to identify the range and types of skills required to undertake and support eResearch; to provide information on which to base proposals for future events, training and other means of skills development; and to provide information for institutions and government for consideration in relation to formal Higher Education and vocational training options.

**Groups of Interest**

Four groups were identified as likely to be engaged in forwarding the eResearch agenda. These were researchers, particularly those engaged in data-intensive research; systems developers, data scientists and other technical staff; data managers of institutional repositories, data archives and discipline-based data centres and their support staff, with those who liaise between depositors and the repository as being of particular interest; and those who are engaged in high-level policy formulation, either in government or research institutions.

**The Study**

The study was undertaken in the latter half of 2007 and the results of four separate investigations have been combined for this report. A further part of this study is planned for 2008, to take the form of an in-depth study of a research unit. This will relate to the needs of researchers and those who provide technical support within the disciplinary context.

Interviews were conducted with twelve key established researchers in six Australian institutions, with a focus on academics engaged in data-intensive research. Interviews were conducted also with the manager of a large data centre, and a repository administrator. The institutions concerned were the Australian National University, the University of Melbourne, the University of Tasmania, the University of Queensland, the University of Sydney and one area of the Commonwealth Scientific and Industrial Research Organisation (CSIRO).

APSR conducted a questionnaire survey of those who attended eResearch Australasia 2007, held in June. About 320 people attended and 76 responded to the post-conference questionnaire which asked, ‘what skills would you identify as necessary in your workplace to support the more rapid uptake of eResearch’. There was also a question asking people to identify themselves as: an academic engaged in eResearch; the manager or support staff of an institutional repository, data archive or discipline-based collection; a system developer or software technician; an administrator engaged in policy or governance issues; or some other category. The results of this question showed that there were few who identified as only one of these, with most people ticking more than one box and the largest single response being ‘other’. Among those ‘other’ were ‘all of the above, plus direct support, plus awareness raising’, an ICT strategy manager, an ‘advocate for the humanities’, a librarian working with academics on e-projects, an ICT solutions/technology vendor, ‘eResearch support staff, a research technology communicator’, an ‘education program leader’ and more, indicating the broad range of people intersecting with the eResearch agenda. Overall, this group identified 140 separate ‘skills’.

In June 2007, a workshop was held in association with eResearch Australasia 2007. ‘The Researcher/Librarian Nexus: The challenges of research data management in institutional repositories’ brought together those with an interest in discussing the role of the librarian in the long-term management and sustainability of research data. Some of the people attending were librarians who already have, or will have, a designated role in data stewardship in their universities. They identified the skills which they saw as being needed as they take on a role in repository management and data stewardship.

The South Australian Partnership for Advanced Computing (SAPAC) conducted a small survey early in 2007 targeted at South Australian-based initiatives related to the National Collaborative Research Infrastructure Strategy (NCRIS). South Australian research groups likely to be involved in national NCRIS initiatives (i.e. not based in South Australia) and the broad South Australian research community. They received 27 replies. The survey was a broad one relating to generic shared infrastructure requirements.
The Results

Two themes emerged from the surveys. One related to skills and the need for training and staff development. The other was the identification of barriers to eResearch, which together contribute to the notion of capability and the need for cultural and organisational change.

There was wide agreement that there are three types of skills required for practitioners of eResearch, their support staff and repository staff. Not surprisingly, there was a strong need for technical skills. Perhaps not as obvious was the identification of a wide range of non-technical skills. Less obvious again was mention of an assortment of personal qualities, which, while not skills in the formal sense of the term, were singled out as being important. In terms of capability, many responses identified the need for communication between and among the different groups associated with eResearch to increase understanding of what each has to offer and what each needs to function effectively. There were also calls for organisational and cultural change to cover policies, practices and structures.

Quotations from interviews are included in blockquotes below.

Technical Skills

The surveys indicated that not everyone needs the same level of technical skills to conduct or support eResearch. However, there do seem to be minimum requirements for all the groups identified above as being of interest, either to practise, or, as in the case of the policy maker or the administrator, to understand fully the import of what is being done.

One senior scientist described the researcher requirement in the following terms, while pointing to some of the deficiencies in support services:

‘In order to get to the more useful parts of eResearch, you have to have a definite skill base to work from, and that skill base is really largely anchored in a Linux /Unix background and to some extent in Windows and less so in the Macintosh field. […] So you need a basic literacy level to look after your computers where you’re storing your data, and then in order to access, like a remote repository, you need to know something about how to connect to that remote repository, what the format of the data should be to go in it, how to convert your data to that required format.’

The list of technical skills identified for researchers and the people supporting them was very long. It included skills related to high-performance computing (HPC) and the access grid, data (and database) management, data curation, information engineering, information modelling, software development, remote communications, distributed processing, informatics, portal design, computational fluid dynamics, database integration, visualisation and programming of all kinds.

Some of these skills are tightly connected to specific disciplines, especially informatics. The field of informatics deals with the storage, retrieval, sharing and use of scientific information, especially as it relates to modern computing and communication technologies. It has had a recognised place in many disciplines for some time, but informaticians are increasingly being sought after to be part of the research team. The SAPAC survey referred to above contained calls for more people in bio-informatics and chemo-informatics.

The need for technical skills is allied to the ability to understand end-to-end workflows, especially for repository managers and developers who need to be able to think like the researcher and to apply that understanding to developing the repository. By workflows, I mean the many software applications, processing operations and interactions required for research tasks to be carried through to completion. For example, the workflow of writing a paper through to publishing it online and archiving it in a repository for open access.

The group of librarians at ‘The Researcher Librarian Nexus’ workshop identified a need for further development of their technical skills, mentioning in particular metadata, something which did not feature among any of the other responses, other than by implication.

Non-technical Skills and Personal Qualities

There was widespread agreement that computer-based technical skills cannot exist in isolation and that non-technical skills are equally critical. These vary from skills in data analysis (including the use of statistical packages and other techniques such as data mining) through information seeking to a broader range of general skills. Project management, business analysis, communications, negotiation, intellectual property, team building and train the trainer were mentioned specifically. Another was generic problem solving, because, as one researcher aptly put it, the kinds of problems which arise when undertaking
eResearch mean that ‘There’s never going to be someone who has done it before.’

On the subject of project management, one senior researcher commented:

‘[...] it’s related to the time urgency associated in the higher education field in general. The management of projects is now devolved back down to the researchers much more than it was a couple of decades ago, and it really depends on the management skill of those people, by definition academic staff. I mean, I know myself, I didn’t go into academia because I wanted to do management. I went into it because I wanted to do research. And I’m not a good manager in terms of projects. You know, not as good as someone who has a skill base in that. So it’s an effort for me to learn that sort of thing. And it’s not so much because it’s hard to learn. It’s because I don’t really have that much of a desire to learn it.’

Others did not see project management as an issue only for academic staff. The librarians involved with the Researcher/Librarian Nexus workshop also identified it as being of high priority for repository managers, along with marketing, advocacy, copyright, metadata, educational outreach and grant submission writing. They also singled out the intriguing skill of ‘researcher management’ while not specifying precisely what this might entail.

A good grasp of copyright and intellectual property issues was seen as essential, with many comments to the effect that this does not currently exist. As one person put it: ‘There’s pretty much a misunderstanding of copyright issues across the campus. Not many people are aware of their ownership, their rights as a generator of intellectual property, where their rights end and the rights of the university take up’.

One repository administrator pointed out the need for all parties to have a good understanding of the broader policy environment, not just at the discipline level, but at the ‘institutional level, at the national level, the funding level and at the international level which again might be disciplinary.’

There was one call for researchers to improve their skills in ‘mathematics; abstraction/inductive reasoning; process-oriented introspection’ on the basis that Towards 2020 Science argues that ‘researchers will need to increase their abilities in the above areas’.[6]

While the various surveys conducted asked about skills, there were many responses concerning the importance of having people with the particular personal qualities. These were listed as: open-mindedness, patience and an ‘ability to cooperate and collaborate rather than compete’.

**Bridging the Discipline/Technical Divide**

While not a skills issue as such, one aspect of the eResearch environment is the difficulty associated with bridging the divide between those with a high level of disciplinary expertise and those who have technical expertise. This may occur within research teams, or where repository managers and developers are communicating with researchers in order to encourage deposit or to provide other technical support. Within research groups, it can be a major issue. For example, one researcher, in the field of finance told me of his need for programmers who have a high level of expertise in economics, econometrics, statistics, maths and programming; ‘otherwise all the programming expertise doesn’t really help because then they make strange assumptions in their coding that just result in nonsense output.’ For others, this is less of a problem and each side seems to be able to learn sufficient of what the other does to perform effectively as a team. One data centre manager was firmly of the view that it is easier to overlay a disciplinary knowledge over a strong ICT background than the other way around. One senior repository administrator was equally firm that the opposite was the case. There does not seem to be an absolute answer here.

One solution to the need to bridge the disciplinary gap is to use graduate students to help with the technical aspects, where those students have an interest and aptitude for this kind of work. In some cases this might be done by providing scholarships, the students then graduating with a PhD on the basis their contribution to the research project has been of sufficient originality to warrant the degree. One example of this is the University of Queensland’s project to create an ‘e-atlas’ as part of the Historical Atlas of Queensland Project[7] being funded through an ARC Linkage Grant. Three PhD scholars will be involved: one with a background in IT, one with GIS and the third in cultural landscapes. Alternately, PhD graduates might be employed to undertake post-doctoral work. Sometimes researchers find themselves taking on all roles: ‘So basically I’m the chief software architect, programmer, plus I have a research agenda’.

**Developing Capability**

There is a range of people associated with research in the digital environment and the provision of support for appropriate data management and long-term sustainability. At one end of the spectrum there are researchers familiar with their disciplines, and whose technical understanding varies from outstanding to basic. At the other end of the spectrum are those responsible for technical support whose disciplinary
skills vary from outstanding to basic. There are also non-technical skill sets which need to be brought to bear: relating to ethics, privacy, intellectual property and so on. Research institutions need to cover this range, but the value of skills can only be optimised in an organisational and cultural setting where they are visible, appreciated and available.

Many of the people surveyed and interviewed commented on frustrating institutional barriers to their capacity to conduct their work effectively. Some of them relate directly to the availability of skilled personnel, while others are more general.

The barrier to research most often mentioned was the difficulty in assembling all the skills required to conduct a project, particularly in relation to data management and stewardship. In some cases the gap is organisational, as happens for example when the researcher is either unaware of or unable to tap into the skills of a central IT unit. More often the gap was in a lack of understanding of what each group needs, what each has to offer and where responsibilities lie. Examples of this can be seen in comments like the following:

‘There is a gap in the methodologies of these two groups that needs to be acknowledged and worked with.’

‘For instance if you’ve got data in say NetCDF file formats and the repository wants it in TIFF format, well you need to know something about the technicality of getting your data from NetCDF format into TIFF format. And that’s actually a technology that’s not widely known. It’s known by people like myself who deal with images but the general people in applied areas don’t know that sort of information. And even if you go to say the IT departments in the university they don’t know that either, because it’s just not considered part of data management per se. The conversion of data is not seen as part of data management. That’s seen as the problem of the person that either holds the data or generated the data.’

‘Few have an appreciation of the appraisal and transition process from the operational environment to the preservation and re-use environment.’

There was wide-spread agreement that everyone would benefit from well-developed and well-propagated policies and practices at the national, international, disciplinary and institutional levels. Some researchers reported that they find it difficult to find solutions to the various issues associated with the complexities of conducting their research. As one person put it: ‘Knowing what we can do, and having clearer guidelines as to who/what/where provides the infrastructure (at the department, school, university, state, or federal level). It all seems to be “somebody else” who will provide the ground level support.’ All research projects are unique: ‘... so you feel like you’re starting from scratch all the time’.

The lack of good tools to integrate the scholarly communication process was also seen as a barrier: ‘more seamless paths from field or library research through writing into publication and archiving’, as were tools developed specifically to meet research needs rather than using those developed primarily for business needs.

Advocacy

There is a need for better communications between all the different groups involved in the eResearch agenda. This was often referred to as advocacy, and it is discussed here because it was mentioned so often by survey respondents and by those interviewed. In this context, however, the need for advocacy means different things to different groups.

Among researchers, advocacy tends to refer to the need for attracting the engagement of academics through ‘a greater understanding of the opportunities available.’ One researcher put it this way:

‘For most discipline areas, the skills are generally fairly basic, relating to databases, internet, etc. For researchers, it is not so much a matter of skills as such, but more an appreciation of what e-research can do to facilitate research in different areas. This will be the primary driver of the uptake of e-research.’

The humanities and social sciences are notable areas where the take-up rate of eResearch has been slower than, for example, in the hard sciences, and where there have been calls for exemplars to be publicised. Many practitioners in the humanities and social sciences find it difficult to envisage where their work might fit into the concept of eResearch. The importance of skills development is one of the issues currently under investigation by the Australian Academy of the Humanities which is undertaking a study of ‘Humanities technologies: Research methods and ICT use by humanities researchers’ [9].

Researchers, data managers and repository staff recognise the need for advocacy when it comes to
increasing understanding of what they do for the purposes for attracting political and funding support. This applies within the institution as researchers try to explain the importance of this new kind of research to vice-chancellors and the like to achieve ‘cultural buy-in – across the organisation and at all levels’. Data managers and repository staff at the same time are putting their case for more resources to provide the infrastructure to support research. Outside the institution, researchers are competing for funds from the different funding agencies, and data managers and repository staff are seeking recognition of the importance of their contribution to data sustainability and reuse.

Data centre managers and repository managers must also persuade researchers that they can contribute to the research agenda. Few researchers are aware that there are such things as repositories, so it is important that the repository is seen as (and indeed is) ‘a good repository – that it’s good in the sense of its high quality but also good in that it adds value for [the researcher].’ The repository manager must also advocate for open access, of both data and publications. The ways in which data can be reused are many: ‘we have valuable data here. And we can re-use that. We can use it in scholarly portfolios, we can use it in government reporting, we can use it in research collaboration which is ultimately the main goal anyway.’ The difficulty here is for researchers to know what they do not know. Many fail to develop a proper data management plan at the outset of their projects, so there is a great need for advocacy of what can and should be done. Liz Lyon in Dealing with Data suggests that ‘Advocacy messages need to be clear and consistent, and ideally will be harmonized across funding bodies. Surveys suggest researchers particularly need advice concerning technical standards for data curation.’ [3]

**Solutions and Suggestions**

While the surveys and interviews that went into this study did not go out seeking advice about what could and should be done, it was perhaps inevitable that there were would be a lot of advice and opinions expressed as to what was needed. These fell into several different categories: suggesting what might happen even if nothing is done or pointing out opportunities to be developed. This section will look at some of them, and also discuss some examples of programmes already in operation.

**Generational Change**

There were strong suggestions that the problem will take care of itself as the researcher population ages and a younger generation of digital natives takes over. Digital natives are those who ‘have spent their entire lives surrounded by and using computers, videogames, digital music players, video cams, cell phones, and all the other toys and tools of the digital age’. [9]

‘... it’s a generational thing. I tend to encourage technology uptake but I recognise that each generation that’s coming through is becoming more and more comfortable with the technology. So you can advocate, but whether the population you’re advocating to has the ability to pick up the technology and run with it is really an open debate. I think if we are advocating more connectivity, which is what eResearch is about, it’s about connectivity between groups and individuals. [...] I don’t want to say age but - because it involves not just an age thing - but also in some ways some technologies you can’t just pick up at a young age because you don’t have the background for them. You haven’t learnt enough to pick them up.’

An increasing supply of digital natives will not be sufficient though if the academic curriculum does not incorporate data management. If you look at the range of skills which are usually identified as comprising information literacy, i.e. to ‘be able to recognize when information is needed and have the ability to locate, evaluate, and use effectively the needed information’ [10], the inclusion of data management is a logical extension, for both undergraduates and graduates.

‘We’ve talked with a few academics about data management being something you actually taught. Even if it’s just a one-off lecture. [...] In my undergraduate days when I was doing labs, I was taught how to do error analysis, how to do measurement, how to do significant figures reporting and so on. [...] So why not explain to people at the same time, oh, and if you want to store that image from that instrument over there, store it in this format. Why do I store it in that format? Oh, because it’s preservable. What do you mean by preservable? So we can keep it - it’s a documented format. That’s it. It’s a one minute exercise. But you just plant those seeds.’

‘It’s significant that a large take-up of high performance computing is coming out of places like physics where the training that the early career researchers is getting is really about data manipulation, and they use the Linux and Unix world, whereas those parts of the education spectrum where you’ve got people who are using computer technology but not developing it don’t get that sort of exposure.’
On-the-job Training

When asked how training could best be offered to researchers and their staff, responses varied. ‘One of the great maxims is that users hate training. Users love support’ was one response, another was ‘I'd love to be able to say one-on-one training’.

The list of issues relating to on-the-job training was large: when to do it, how to do it, what to include, how to cover the cost, how to provide the right trainers. Training is most effective when given at the point of need, making timing a particular issue.

On-the-job support also means assisting with the design of research projects to influence the selection of file formats and other data management procedures. It means trying to influence the researcher at the beginning of the project, rather than picking up the pieces (from a data management perspective) some time further into the process, or indeed at the end of it.

One possibility for example, would be to create small learning objects to support data management, along the lines of the icons which pop up in some word processing programs which interpret what is going on and offer help.

Certification

One of the researchers interviewed, a senior epidemiologist, had found it very hard to locate appropriately qualified staff and asked plaintively:

‘...is there a Masters of Data Management? Are there people out there that actually have an expert [qualification] – like you might find someone who’s done a librarian course? Well, is there a Masters of Data Management so that there’s these people that actually know how to do all this stuff that we want to do. ‘The answer to this question is no, in Australia at least.

There are some courses appearing in the USA, the Digital Library Curriculum Project [11] which builds upon a collaboration between Virginia Tech and the University of North Carolina, Chapel Hill, and the Digital Information Management Certificate offered by the University of Arizona School of Information Resources and Library Science [12] to name but two.

In Australia it is possible that a formal training mechanism will be developed through the proposed Australian National Data Service (ANDS). ANDS will take a strong role in providing advice and assistance as ‘education, training and consultancy programs are required to build the capability of data creators, data stewards, and data consumers to participate in the commons’. [13]

The ORCA Model

The wide-ranging list of needs outlined by survey respondents means that there will not be one delivery mechanism for skills development. Some of these needs will be met through the formal education system, some through training offered within the institution and some by training bodies set up for the purpose.

The Online Research Collections Australia (ORCA) Support Network was established to support the ORCA Collections Registry [14]. Its training programme was designed to provide some of the professional development required by researchers, technologists, and collection managers and was based on the assumption that training at a local level can best meet local needs.

The ORCA Support Network members are situated in major Australian research universities and employed within central areas responsible for repository management and development. During 2007, they offered consultancies and advice to researchers and shared their locally developed training programmes with each other. Local training programmes offered, subsidised by APSR, included: developing a data management plan, XML for online collections, data rescue, data exit planning for retiring researchers, cultural informatics, archival practice and the records of eResearch, and digitisation for beginners. All training materials have been shared through the ORCA wiki.

Instruction on developing a data management plan was given a high priority among the ORCA members for two reasons. Firstly, to improve local data management practice and secondly, because it is expected that the Australian Research Council and the National Health & Medical Research Council will require such a plan in future in research funding applications. The Australian National University has elected to develop a course which can be offered through its Graduate Information Literacy Program, both online and in the classroom.

Conclusion
If research institutions are to minimise the gap between the ideals and realities of eResearch, there is some way to go in providing both institutional capacity and appropriately qualified individuals. While eResearch is dependent on good ICT infrastructure, this is not sufficient in itself. The results of the survey outlined here show that capacity in information technology skills is important but must be accompanied by a range of non-technical skills in such areas as project management. Equally important is the creation of research environments which are covered by well-propagated and understood policies, which are appropriately organised into structures with clearly delineated roles and responsibilities and which minimise the current barriers experienced by many researchers.

Acknowledgements

My thanks are due to Adrian Burton and Chris Blackall of APSR, Belind Weaver, Rowan Brownlee, Gavan McCarthy and Simon Porter from the ORCA Support Network, and all those who kindly gave me their valuable time and insights into the world of eResearch.

References

5. A list of these can be found at http://www.apsr.edu.au/publications/discussion.htm
12. University of Arizona Graduate Certificate in Digital Information Management http://sir.arizona.edu/digin/
14. For further information about ORCA, see http://www.apsr.edu.au/orca/

Author Details

Margaret Henty
National Services Program Coordinator
Australian Partnership for Sustainable Repositories
Australian National University
Canberra
ACT
0200
Australia

Email: margaret.henty@apsr.edu.au