

Ten Major Issues in Providing a Repository Service in Australian Universities: extended interview comments

This paper provides supplementary information to
"Ten Major Issues in Providing a Repository Service in Australian Universities" by
Margaret Henty, *D-Lib Magazine*, May/June 2007, Volume 13 Number 5/6.
<http://www.dlib.org/dlib/may07/henty/05henty.html>



Margaret Henty
Australian Partnership for Sustainable Repositories
Margaret.Henty@apsr.edu.au

http://www.apsr.edu.au/airs/major_issues_comments.pdf

May 2007



This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivs 2.5 License:
<http://creativecommons.org/licenses/by-nc-nd/2.5/au/>

Introduction.....	5
The repository environment	5
The issues.....	6
Roles and Responsibilities	6
Sustaining the service	7
Engaging the community	7
Guaranteeing quality of service.....	8
Defining the collection	9
Research reporting and compliance.....	10
The repository and eResearch	11
Skills and staffing	12
Technology.....	13
The regulatory environment: copyright and digital rights management	13

Introduction

The Australian Institutional Repository Survey (AIRS) was conducted during the third quarter of 2006. An article on the ten major issues identified by respondents to the survey was published in *DLib Magazine* in 2007. Because of limited space, it was not possible to include in that article all of the comments made by interviewees about particular issues, so these are made available here.

It is recommended that the complete article in *DLib Magazine* be read before or in association with these comments.

Margaret Henty, "Ten Major Issues in Providing a Repository Service in Australian Universities", *D-Lib Magazine*, May/June 2007, Volume 13 Number 5/6.

<http://www.dlib.org/dlib/may07/henty/05henty.html>

The comments came from thirty-three people in fourteen universities who were interviewed in 2006. Those involved were Deputy Vice-Chancellors (Research), Pro-Vice-Chancellors (Information) or University Librarians or any of their equivalents, with other personnel such as repository administrators, library staff or IT specialists on occasion.

The participating universities include Australia's eight largest research universities (known as the Group of Eight), partners and associates of APSR and ARROW (Australian Research Repositories Online), and a small group of others who were geographically accessible and willing to participate.

The repository environment

Clearly the issue of mandatory deposit is one that is going to come up at some point, probably more at a national level. My instinct is not to push that sort of line but I'm aware of it. It might well come possibly as a by-product of the RQF process. And clearly some of the nutters out there in the open access environment are pushing hard on the mandatory route.

I've always been unwilling to ... go around bashing and hitting everyone over the head and saying "come on, it would be good for you to submit your material into a repository". And more critically than that, our Deputy Vice Chancellor Research shares the same view... And I've always myself felt much more attracted to the concept of a research repository in the context of a research community.

I didn't want to go through all of that effort to put up something that was already available. Despite all the urgings of open access and all of those higher order things, to me it was just a question of resources. And also I would have to say I haven't been able really to engage more than a handful of academics around the place on open access.

And by now it was clear that a discipline-led approach was not going to happen, and that's what I had thought would always work - that it would be the psychologists and the physicists and the medical scientists would each find their way to a more rational and more cost-efficient way of sharing the knowledge in their disciplines, and with

the kind of massive precedent for this being what the physics people were doing in Los Alamos. But it didn't happen.

So it was an acknowledgement that we needed to have stewardship over digital assets, particularly at the research output end, and that we needed a strategy to do that, and then within that, we needed some practical solutions.

The issues

Roles and Responsibilities

... if not the library, who? There really is nobody else that I can identify in the university community who has probably the skills and necessarily the will to actually take this responsibility on. Libraries are in the business of describing and providing access to information, whether it's in a book or a journal or something like that. And at the end of the day, what we're talking about is exactly the same sort of service - but on behalf of a much more distributed community.

I think we're expected to do it. I mean, I think quite honestly people just assume that the library will do it. And organisationally it makes sense. I just see it as part and parcel of our activity. In another sense, I thought about the university archives as actually being a more logical group to take it on but I don't think they're at the centre of the organisation. So, no. I think it's just something the library should do.

I think a really obvious one is if somebody says it's an institutional job to start working on the data curation issue, then which part of the institution is going to take that on? And I think the answer should in 100 per cent of cases be the institution's library. It may end up being that in 80 per cent of cases. It depends. It depends what the state of readiness is for the library to move into it.

We've been looking down at our [Conservatorium of Music] recording a student who comes in for their audition to go into a course, and then, as they get into the course, recording every performance or whatever they do throughout their career, and everybody just sort of thinks that the library will be in there doing it.

I think the university needs to address the issue of governance, and that obviously is related to management but is different in the sense that it would need to be broadly-based. You would need to draw in different parts of the university who are making different contributions.

And with that will go a question about who's got custody of this? One of the issues for the University as a whole is the issue of data custodianship, whether we're talking about scholarly information and knowledge or the data that the University uses to do its business - like HR data and student data and so on... Wouldn't it be good if in the end we had a reasonably seamless system? So e-ness - eData, eRecords, eManagement thereof, eAccess is all sort of becoming a bit intertwined in people's minds.

Sustaining the service

At the moment here, for instance, it's all funded by the library more or less. We've got to get the university centrally to agree that the equipment is part of a central infrastructure.

The biggest issue as far as I can see it is to secure the funding to put it on a stable and ongoing basis. We're relying on project funding for this year, and [...] for next year we're relying on people giving extra time from their other responsibilities in the library. And I think that is one of the major things - to get it established as a university operation that needs to be funded centrally and the library will be recompensed for the staffing.

And the sustainability issues around funding and staffing. A number of us have managed to achieve things through external grants and external funding. When they dry up, how do we keep it going? Is it something that institutions will agree to fund and if so what benchmarks of success will they look at? Or is it something where we divert staff?

So yes, resourcing I think is still a question mark. We're repurposing some people.

Is the repository going to become our new core business? I would pose that as a semi-serious question. Perhaps not in the next five years but in the next ten years we're going to see many, many libraries wrestling with the idea that there's all of this other stuff that people are far more interested in than the sorts of traditional library services we have.

One thing that does come out is the new kind of partnerships that are emerging out of this, where in fact, while in the past the library was seen as perhaps just a place you dumped material, now there's a more active partnership where [...] not only are we seen as being able to help manage material but [...] there are possibilities of how this can be re-used and reflecting our involvement in ARC projects and things like that.

Engaging the community

The proof of the pudding will be its value to the university, so that's the key issue. Are we creating something that is perceived as being valuable to the university?

I guess from the average researcher's point of view, to use that phrase in inverted commas, the idea of digital repositories and their value et cetera is still very much something that ... if we stopped 100 of our researchers in the street and asked them what a digital repository was, what was its purpose and how would you access it, and how would you use it and is it important to you, then you'd get a very wide range of opinions. But you'd get a lot of people going "a digital what?"

I mean people look at your CV to see what you've published and see where you've published it, and whether or not it's been in a repository doesn't really cut much ice. So although I think these are valuable things, I think some of the arguments which

have been used to try and promote them haven't been very effective, and certainly in research intensive universities.

There would be people who have a genuine concern that the data they create is still around in 50 years because it might still be useful. But I think for a lot of academics, this is just too hard. They'll do what they have to do today and somebody else will have to worry about that bigger picture.

I'm hoping that the [Research Committee] will recommend to council that they will become mandatory. ... Experience shows already with other institutions that the technology is the easy bit. Getting researchers to engage with the repository is probably one of the biggest challenges, and I can't see how the library can ... the library doesn't have the authority to mandate that, and also doesn't have the authority to invoke any consequences of not following that restriction.

We also were surprised in science... They suggested it should be mandatory, and we thought, boy, if the library had said that, they'd have had us up against the wall. But when they suggest it ...

I certainly think if mandatory deposit doesn't kind of become generalized, that repositories aren't sustainable because it takes a lot of work to get people self-archiving. [But] once they're self-archiving, they'll keep on doing it.

You say, okay, we're going to talk about this, and have this conversation every time we get in front of an academic. And that will then probably drive the intake, but you have to be able to start that at the other end too, and say okay, can we provide enough administrative support to get this stuff in and actually make it manageable? Because the last thing you want to do is have someone come to the table and have such a bad experience that they then say, "I'm off. I'm never going to do this again because it's horrible" And then they tell all their colleagues and go "this is crap". Don't do it.

Guaranteeing quality of service

We have to be able to deliver what we say we can. We can't offer the world and offer every solution to everyone when we actually can't do that. And if we try to do that we're going to come unstuck. And so I think we have to deliver and be credible, as well as position ourselves to be able to extend as things grow.

If you do these badly, you've lost your chance.

We need to be able to run fast enough.

I was at a humanities workshop on Monday and Tuesday, and one of the things that I kept saying to them was that the number of Australian academics who are retiring in the next 5 to 10 years makes it absolutely imperative that we come up with some strategies. And I suppose that's what's driving us in a way, is to capture the stuff that people have been gathering over many years and they're now getting to the point where they're retiring, and it's no different to academics retiring and wanting to give the library their collection of books.

So it's the ease in facilities which is really I think so critical ...

How do we put in the tools, the standards, the processes and approaches, and the underpinning technologies to kind of make that happen seamlessly so it is easier for the researcher? And they don't know what's happening behind the scenes. They just know it's easy and we're doing what we need to do to contribute to this.

The other thing ... is to try and find the tool that will actually enable the researcher to work with a repository more easily. I don't know enough about it but the Scholars' Workstation that an academic at the ANU is doing - something that encourages and makes it easier for the researcher to work with a tool and transparently ... So the software development that allows the integration from the actual research process of an academic or a researcher, and then the automatic feeding of that information to repositories and publishing and the automatic links to whatever the research, the RQF or the government reporting bodies require. We don't have that yet.

It's part of normal business for them. Whatever they would normally do with their output, it just happens easily and they only do it once. And then the university's purpose is behind that and they are myriad - untouched by an academic hand so to speak. Because they can see the point of something like an individual research portfolio which they can use for career purposes, for grant applications and so on, and if that's the thing that drives them, then that should be the front end that drives our process. We shouldn't have a different process. It's that sort of thing. It's getting that seamlessness so it fits in with an academic activity rather than ... it's something they were going to do anyway and they're happy to do or they're keen to do.

Defining the collection

The future policy issues as I see it from the library perspective - first of all one of the most pressing issues I think, is definition of the scope of content, and that picks up that whole issue of research data. And if so, how do we define research data? How do we handle it? We're already, even at this early stage, having a little bit of difficulty with the scope of content, because on the one hand we have some people who are saying we should have no scope - we should have no boundary drawn, and basically anything anybody wants to put in should be okay. Why not? Then we have another group that are saying, "No, no". As a university repository, we have to have some control mechanism, particularly as this currently is an open repository. We need some control mechanism to ensure some kind of quality control about what goes into the [...] repository.

And then, the issue that people aren't thinking about and I guess are a little antsy when I go to meetings, is how do you get rid of stuff in the repository? Everybody is intent upon taking all this stuff in but no one's thinking about how they're going to make sure that the dross is taken out at the end. And it's all very well to say that storage costs are dropping, but it will get to the stage that we get with libraries, where you end up with a major problem in 5 or 10 years time.

I think there are some retention issues. Some of my discussions to date have been around the issues of retention and disposal of books, and that's an issue that librarians have never really got to grips with in a sense, or in a coherent way. The same thing is going to come up with repositories. It's one thing just filling them up, but there is going to come a point where we recognise there's redundant stuff in there that's doing more harm than good and needs to be archived off the live repository. How we deal with that is something that a collection policy ought to address.

I mean, the reality is that huge amounts of teaching material are generated every semester and lots of them have no value beyond the semester. We've adopted an interim strategy ... they delete the entire content of the Learning Management System unless requested to retain it. That's just to keep the storage requirement under control - just to keep it down to 140,000 objects.

Research reporting and compliance

Of course everything we're doing is because of the RQF - everything. Even if it's nothing to do with the RQF. In a way, the research repository has got absolutely nothing to do with the RQF, but you can sure as hell make it have everything to do with the RQF.

Research Master is the system we're putting in here for our research data, and the library has been having discussions with our Research Office and those discussions are at a very early stage but clearly ... it's obvious to me and to everybody I suppose, that the future development of the repository must be linked to the requirements of the Office of Research and the RQF requirements. And any interface we can develop between our repository system and Research Master is going to benefit everybody.

I think there's a creative tension between what the university needs to be doing in order to comply with the RQF and some of the principles of the preservation of scholarly information. I think ... the need to provide access to the research content in order to measure quality and impact sometimes appears to take priority over good preservation practice.

There's a standard requirement, a global requirement I guess, that in many disciplines, the data from which publications are drawn should be kept for a specified amount of time. Typically it's five years and at [this University] in areas other than those where it's required to be kept for longer, there's an expectation that the researchers will maintain datasets that represent the primary source of information from which their publications have been drawn. On a few occasions we've actually audited various parts of the University to find out whether that's being followed, and by and large, the auditors didn't find any problems with that, although they warned us about the vulnerability of the holdings because of backup problems and things like that.

So as you would know [...] the ARC has had a longstanding requirement in its guidelines about putting data into secondary repositories. Most people don't even know what they are let alone comply with it - let's be candid. And the tools that were developed some years ago when that requirement first came in were not very

friendly. The ARC never checked up on it and most of our people didn't even know what it was that was being referred to in the guidelines. So it was a bit of a perfunctory sort of condition really.

The repository and eResearch

But the idea that DEST funds people to do ... the ARC funds people to gather data is seductive but a bit pernicious. It does in fact fund certain data gathering exercises and, under those conditions, I think obliges the people to deposit their data in ACSPRI or somewhere like that. But, by and large, it's actually funding people to get research results, and along that path they generate data that they need to keep because of institutional policies, but they're not funding the data. And the cost of producing that data so that other people could use it is almost unbelievably high given that almost nobody else would want to use the data anyway. It's an interesting situation.

Just to put my cards on the table I'd go one step further and say that of all of the modern descriptions of the world, the fact that we're flooded by data is true, and anything that can be done to help us deal with that flood of data should be done. We will find it increasingly difficult to operate the communities that we live in if we can't get control of data and turn it into information and knowledge, and the best way I know for doing that is to thin it out. Either throw away stuff that isn't important, which requires a very careful judgement, or to aggregate the data so that it produces high levels of information and knowledge. And that's what you do see in research publications or in monographs. The notion that the primary data should be kept forever doesn't make any sense at all, although it shapes people's views about things

It's very easy for an ethics secretariat to say, "You will look after terabytes of data for 7 years", and it's quite justifiable for people to then turn around and say "I don't know how to do this and I don't have the resources to do it". But I'm also not going to have the library sort of pretend that it knows how to do it because it doesn't. We don't have the resources and we just don't know how to do it.

But as these things get bigger and bigger and bigger, pretty soon we're going to reach a point where manageability or the requirements you need to put on it to retain manageability outstrip your ability to fund it. And I think that's going to be a real issue for many institutions and many libraries to address.

Coming back to being early days, we're finding out more about what we don't know than what we know about some of these things. The traditional library has been fairly up to speed on classes and teaching and learning and all of those things about supporting study areas, but it's interesting how little some of our liaison librarians know about some of the research activities of the staff and I think we need to move into that area. So I think there's a lot more to discover.

And when it comes to research computing services, that brings with it its own set of issues. Number one it's incredibly hard to get people with the right skills. Number two there's a bit of a credibility issue. So we've experimented with academic

appointments as opposed to general staff appointments, but that's not easy to manage either.

We've touched already on data storage, and if you start to think about maybe some of the visual image collections around medical and veterinary science research, you get into the realms of petabytes, and storage is getting cheaper every year, but still that's a big issue.

Another key issue is for us to get right the continuum through the various repository layers. So if there's a large data store sitting at the bottom layer of the university that's also a repository, how will that be managed? ... And then, how does it interoperate with or speak to or be part of a landscape of repositories which includes the [...] one which is the one I think you're most interested in. But although it's the one I'm managing, I don't actually think it's more or less important than the mass data store ... So it's not the size of the store that's necessarily important. It's the use that needs to be made of the data. And we've always taken an information management perspective on this. The issue for us to resolve over the next five years is how to get the landscape of repositories working together.

I quite like the idea of some national support, but I'm not sure it's about infrastructure in boxes. I think it's more to do with setting standards and policies, and giving people the support through that, and also that because researchers work across institutions, the idea of some national support there is a good one. But I don't think of it in terms of a great big super computing facility with lots of stuff.

The difficulty is going to be around all those interfaces - all those borderlines between different levels of organisation and different types of activities. What data are you going to have? Who's going to manage it? It's a local versus state versus national level. And all those issues are where the problems are going to be I think.

But as far as managing the data I don't know what the answer is. I'd be surprised if it was a centralised answer but maybe some kind of centralised policy and so on would be relevant.

Skills and staffing

I think one is staff - making sure we've got a continuity of staff and their skills and developing the skills. I think we underestimate just how important staff are and I think more than the technology, the staff is probably the major one.

I've found that in libraries, if we're going to be managing repositories, the gaps that we've got in our skill set. Some elements of programming would be really good to have on hand. We don't have a library IT department here. We use the University's IT Department. It's amicable, it works well, but we need a kind of a day-to-day, hands-on skill set there.

I think you need two types of skill base to manage this. One is the technical knowledge and the quality assurance for the actual system that both accepts the content and checks its bona fides and then makes it available or manages the access. The other type of skill are the staff who'll actually provide information about the

system - how to use it, encourage the use of it, troubleshoot the researchers' interface. And some of those skills are transferable from existing library operations.

Technology

Far better I think if we accept that there are going to be many different flavours of repository or many different ways of storing the stuff. But if we are all talking the same language and using the same interfaces and the same management protocols and the same type of security infrastructure, at least we've got a chance of getting interoperability.

And I think a significant issue too is the choice of platform. We've chosen dSpace for a number of reasons, but principally because it seemed to offer the sorts of facilities that we needed immediately in a package that I could readily install and bring online very quickly. It only took all up two days for me to install dSpace and have it running. But we are aware that there are other products out there.

And I think also another issue longer term is how we determine some of the really good technical solutions for all of this. We are having to try and put an umbrella integration across disparate software developments [...]. How do we integrate that so that for the user, they do one search across and they can achieve it, but it's in a journal, it's in a repository or somewhere else?

I mean as someone who's worked in some very large libraries with some very large catalogues, I think you realise the importance of having the means of actually locating and finding useful stuff in large collections of whatever it might be - book, journal, whatever. In other words, are you going to rely on just free text searching of all this stuff and not have any means of, I think, classifying or providing key words or other metadata that will enable you to successfully and sensibly get classification of stuff of relevance and recall, to actually extract out of maybe a federated system of 30, 40 or 50 repositories - useful stuff for your search. And in fact if you're going to have some sort of either controlled vocabulary or some way which requires, I guess, professional input when you're putting the stuff in, so that you can actually get it out sensibly then you're talking about possibly major inputs of staff resource and effort.

The regulatory environment: copyright and digital rights management

I think intellectual property and copyright given our experiences so far with our adviser here is a major issue.

Data that you generate as an academic at University A and you move to University B - do you leave it all behind? Do you take it all with you? Do you create it in such a way that you can do both? Who really owns it? What if it's externally funded anyway?