A researcher's viewpoint

Arthur Sale
Professor of Computing (Research), University of Tasmania
Primary research fields: bioinformatics, mobile computing, and open access

The two hats

In discussing the roles that researchers take relative to Open Access, it is important to note that they approach it with two different attitudes, depending on which phase of the research they are in. The most familiar is the researcher while conducting research and looking for information about the research topic – the *searcher*. Libraries have long dealt with searchers. The other role is that of researcher as disseminator – the *author*. Libraries infrequently deal with authors, and usually as a special case.

The distinction between these two 'hats' or roles is important: the needs of the researchers are very different, and so is their behavior. Let's tease out the consequences of this classification.

Researcher as searcher

Client communities

When an Institutional Repository is proposed, one of the first questions to ask is 'Who are the intended readers?' Unfortunately, the answers are not so simple.

- Every operator of an IR would of course nominate researchers in other institutions as one of the prime client communities. Open Access is supposed to open access to local research to researchers globally. Such searchers may be in universities, research institutes, or in business operations. This is of course absolutely correct and the first priority, but this group does not comprise all searchers.
- The second obvious group is really a class of meta-clients: the institution's research management entities, the grant giving authorities, and often government. Because of the power of this group, their needs as searchers for meta-information may be given almost an equally high priority as the genuine researchers. Their influence can often be spotted in otherwise unnecessary metadata and search facilities. Other meta-clients include the researchers into repository usage and impact.
- Thirdly, there is a diverse group which I will call the general public. These may comprise school teachers, school students, and the simply interested individuals. Since much research is often of esoteric interest and may be written in highly technical language, members of this group may not be interested in it. However there are classes of research for which this is not so. Personal health, eco-systems and environmental issues, politics, history, culture and art are examples. For instance a paper I wrote 30 years ago on generating pythagorean triads (whole-numbered right-angled triangles like [3,4,5] and [5,12,13]) continues to evoke a consistent stream of enquiries from this group, mostly amateur mathematicians or school teachers (145 downloads in a year).

Knowing the target readership can affect the responses to the rest of this discussion. Let's concentrate on the researchers.

Priorities and Tolerance

Searchers have clear ideas of their priorities, which are often surprising to the operators of repositories. Their top priorities and expectations are that

(a) what they want should be easily discoverable

(b) everything provided by the institution should be available online

These two aspects are not negotiable. Searchers will simply not pursue hard-to-discover resources, and if the second expectation is not met the resource will not be discovered. A subtle variant of this occurs if the repository contains only a bibliographic record of the resource, and not the 'full-text'. Most searchers will ignore such resources since trying to acquire them does not seem worth the effort. Some improvement can be achieved by placing an email link on the metadata display page that creates and formats a request to the author for a copy thus minimizing the work involved (one or two clicks), but this is at best a palliative. Much lower on the priority list is:

(c) authoritative content

The evidence suggests that searchers would like to see authoritative content such as actual refereed research papers published in journals but their need for this is low. Their primary need is to read the paper to determine whether it is of interest to them, and even the provenance of the paper is of lesser interest. A pre-publication postprint, a preprint and even a version in plain-text with no formatting and all the diagrams removed may be quite acceptable. If the paper interests them, then they will be prepared to invest more work in finding a more authoritative source, if only to quote the page numbers in a reference of their own. The lowest priority is:

(d) visual sugar ('eye candy')

Searchers couldn't care less about visual sugar on the pages they are presented. By visual sugar I mean added headers, footers, prettiness, etc added to the basic scholarly paper they have come to read. At best it just wastes screen space or paper; at worst it irritates the reader. Plain is good.

What are the consequences for the repository operator? They are simple to enumerate: capture 100% content, and make your content discoverable by as many means as possible. Provide an email link on plain bibliographic records, and certainly always provide a link to the authoritative source. Keep the web pages simple and clean, maximizing information content.

Journals respond to this analysis too. An institutional repository is not an alternative to a journal for a researcher intending to reference a paper. *Rather it is an alternative discovery tool directing researchers to the authoritative article.* This is possibly one of the reasons why the OA movement does not seem to have any negative influence on journal subscription rates in research institutions: researchers as searchers still want to have access to as much authoritative content as they can, even if they discover the content otherwise than in paper or at the publisher's website. Although the actual accesses to the publisher's website might drop, there is no pressure from the researchers to cancel subscriptions.

Discoverability

I suggested that making content discoverable by as many means as possible is desirable, and so it is. For example, the metadata for a PhD thesis in several repositories (such as the University of Melbourne) is harvested by the *Australasian*

Digital Thesis Program and by the *ARROW Discovery Service*. ARROW also harvests from the ADT Program, so the thesis metadata appears twice in it. Google, Yahoo, Scirus, OAISter and other search engines harvest from ARROW, ADT Program and the Melbourne repository. Find it whatever the route.

However, since the Google Scholar program was announced in late 2004, it has been increasingly obvious from inspection of the logs on my repository (and from the user statistics) that a high fraction of the hits on the repository come from Google. Even more significantly, these hits were direct to the 'full-text' file. On analysis it was discovered that Google was indexing pdf files, and the searchers were choosing them preferentially over the pages that presented the metadata (title, abstract, etc). An example of a search result with both destinations is shown in Figure 1.

Figure 1 – Result of a Google search

The implications are serious. It is totally desirable that people using Google as their search strategy, perhaps their only one, found a resource on our site. However, it meant that all the metadata and all the extra information that we might put on the metadata page (such as links to the authoritative source) was simply not viewed by the searcher. This posed questions for repository management investment.

Our consequent analysis suggests that metadata generation, and especially 'perfect metadata' should take a low priority. Author- or automatically-generated metadata may well be satisfactory. The metadata may increasingly have the main role of allowing porting of content to a new repository and similar library and archival functions. Only local searchers use local repository search; few searchers use federated national gateways either since they don't know about them. Federated global gateways are the primary discovery tool.

This is not to say that federated national gateways have no use, rather that they address a different group of clients: those I nominated as meta-clients. National gateways are used by in-country librarians (who know about them) and government. They also help slightly in multiplexing the discovery routes. But their search engines and federated metadata repositories should not be seen as major contributors to searcher discovery.

Researcher as author

Let us turn our attention to the quite different behaviors exhibited by researchers when they are acting as authors.

Research Impact

Researchers that know about open access practice it for one major reason: to get their research disseminated to as many people as possible. The reward comes in knowing that the research has been used and valued by other people, and that the effort and money in producing it has not been wasted. Secondarily they may receive monetary awards in the form of prestige, tenure, promotion, or more research grants.

One measure of research impact is citations of the work. A citation means that someone, whose research it presumably influenced, thought the article significant enough to include a reference in their own publication. Research-measuring authorities are slowly realizing that journal impact factors are just a surrogate for citations, and as they do researchers will become more and more interested in citation counts. Other chapters in this book will address the increased citation rate that open access articles generate, compared to paper-only articles.

However there are other forms of research impact. There is increasing literature showing that download statistics predict citations. Some downloads are not related to citations, but may affect the behavior of non-researchers nevertheless. Examples are government policy changes, changes in teaching practices, and industrial developments. Some self-archiving open access researchers have been known to complain that they get too many emails about their articles; my riposte is always 'Would you prefer to be ignored?'

Copyright

Authors seldom have any knowledge of copyright law, and are extremely hazy about what they sign away in a journal-author agreement. They just know the university didn't care about their rights, and they have [roughly speaking] signed them away for free. Consequently when asked to self-archive, their instinctive response is to be risk averse: 'I don't know anything about this and I don't want to be sued, so let's play safe and say no.' This is a significant barrier to overcome, even if it is nonsense.

University copyright officers and librarians also raise this problem to a much higher level than is necessary. They do know or want to know copyright law, but they often mistakenly conflate music and video piracy with scholarly publishing. The two domains are worlds apart. Scholarly research output is always given away for free; indeed sometimes the author is asked to pay to have their work published, and the publishers make their money out of disseminating this material they get for free.

Their angst is totally unnecessary since over 90% of publishers have no problems with self-archiving at present. Conference organizers are similarly approving. However it is a major barrier to take-up of OA self-archiving, and strategies to attack it must be undertaken by the repository managers. One useful strategy first employed by the Queensland University of Technology is to say to the authors: 'Just deposit your article. We (the library editors) will check your copyright agreement and make it open access if possible, otherwise it will be restricted to non-campus use. Of course if you want to, you can specify the article should be open access or restricted.' This works!

An interesting feature of author response to OA is that once a researcher has deposited one or two articles in a repository and seen the value, they cease to become copyright-sensitive. They integrate the issue in their general research practices at the back of their mind, many of which have legal consequences of which they are blissfully unaware or reluctantly aware, and never look back.

Mandatory policies

My colleague Alma Swan has written about her study of research attitudes to policies that require (or imply censure for not) archiving an institution's research publications. Briefly, most authors won't self-archive voluntarily. Only a little work is required especially compared to producing the publication in the first place, but this work is avoidable and will be avoided. Aversion to getting involved with copyright may also play a part. However if required to self-archive, authors will comply willingly since the authorities obviously value the activity and will handle the copyright, and it is only a little work.

Let's look at a graph of how this works out in practice, examining all seven Australian universities that operated a repository in 2004 and are harvested by the *ARROW Discovery Service*. Figure 2 shows the number of items in each repository with a publication date of 2004 or 2005, as a percentage of the officially reported research publication count to the Australian Government. Three universities are identified as exemplars of the major factors:

- University of Tasmania (low library support, voluntary deposit),
- The University of Queensland (strong library support, voluntary deposit), and
- Queensland University of Technology (strong library support, the only one of these seven having a requirement to deposit all research output, commenced on 1 January 2004).

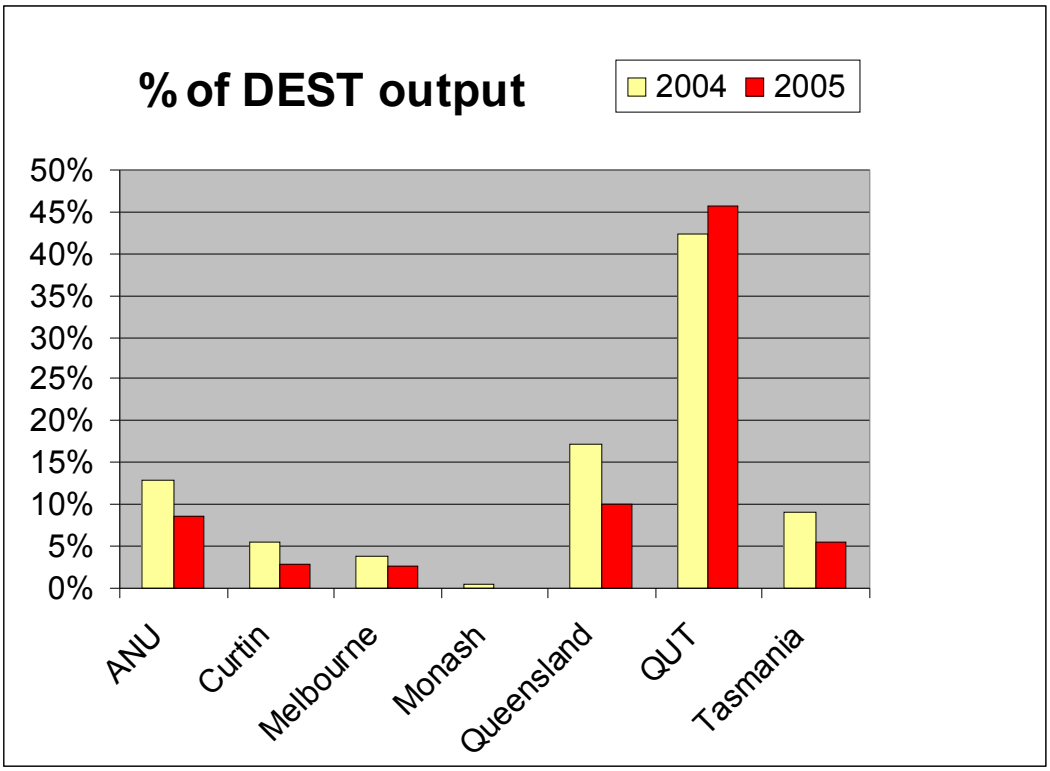

Figure 2 – Content of Australian repositories, January 2006

Clearly, voluntary policies don't work, even with above average effort. On the other hand, requiring researchers to deposit is easily accepted even in a nation of rugged individualists like Australia. Since at the time of writing all 2005 input had not been received, Queensland University of Technology content looked like reaching 60% of available research in its second year of a requirement policy. The implication is clear. Any institution that does not have a compulsory deposit in its kitbag is wasting it

money establishing a repository. 20% success is the most that can be expected otherwise and even that is optimistic.

Conversion

There is an interesting phenomenon to be observed with authors. Although they are difficult to convince to self-archive, for the reasons discussed earlier, once they have self-archived one or two articles, they don't look back. It becomes a routinized part of their research activity, and a significant number become enthusiastic.

It is almost like St Paul's conversion on the Road to Damascus: many researchers become evangelizers and start infecting their own colleagues with their enthusiasm. Email feedback from readers, citations, and the evidence of the download statistics, pays off in spades. Some of the consequences are discussed later. However, this is good news for the operators of institutional repositories: initial hard work to provide author support decreases with time, as more and more of them come on board, and fewer and fewer need support.

Research training

If OA access is a key activity for researchers of the 21st century, are we doing enough to train the researchers of the future, even if they are more Internet-savvy than their elders? Maybe not yet totally integrated, but in my university and school we are doing our best. I run a generic skill short course for PhD candidates in self-archiving. PhD candidates immediately see the benefit of self-archiving their publication (citations, exposure, comment, claim to priority) much faster than faculty, and adopt it very easily. They then have a Trojan Horse effect on their supervisors – weak, but does sometimes work.

We have also incorporated this into our Honours program (4th year) by making the First Class Honours theses available online, as well as any publications that these students achieve, thus encouraging transfer into the PhD program by exposing the students to modern practices. I view this as part of training the candidates and inducting them into the practices of 21st century researchers.

Retrospectivity

As noted, some researchers become avid OA supporters. And frequently they will scavenge their old files to find old articles that they can mount. The more enthusiastic will even bring out their paper-only articles and scan then as text images. Generally this behavior is restricted to those articles that the author feels really proud of, or thinks could stand the test of time. Articles that are somewhat dated may be passed over. To give an example, my institution's own archive has 15 articles with an original publication date prior to 1980.

Some researchers adopt a different approach, which I call the 'just in time' strategy. They don't post all their old articles, but as soon as someone asks for a copy of an old article, they arrange to have it scanned (or scan it themselves) and put it on the repository, sending the URL to the requester. This is equally effective, but driven by the readers rather than the authors. The problem with this is that the article may not be discovered, because even its metadata is not on the OA repository…

Why do authors do this? I believe that the answers are in the next two sections.

Avid dissemination followers

Some authors become interested in their dissemination success, and add this into their research strategy. The benefits are that they see where their work is cited, in broad

terms who is interested in their research, and which areas might be most productive for future work. They also learn about citations and their importance and tend to follow some of the meta-literature about research.

This can be encouraged by providing the authors with feedback from the repository in terms they can understand. Conventional web statistics are no good as authors cannot understand the ICT jargon – the statistics must be couched in meaningful language. For example, I wrote a statistics package which is used by my own and other universities. At the bottom of each document metadata page is a statistics link (alternatively available from the home page) which gives access to counts of metadata views and downloads for the last 4 weeks, month, year, or all time, broken down by country of access and month. Figures 3 and 4 show sample statistics for a document, for the year 2005.

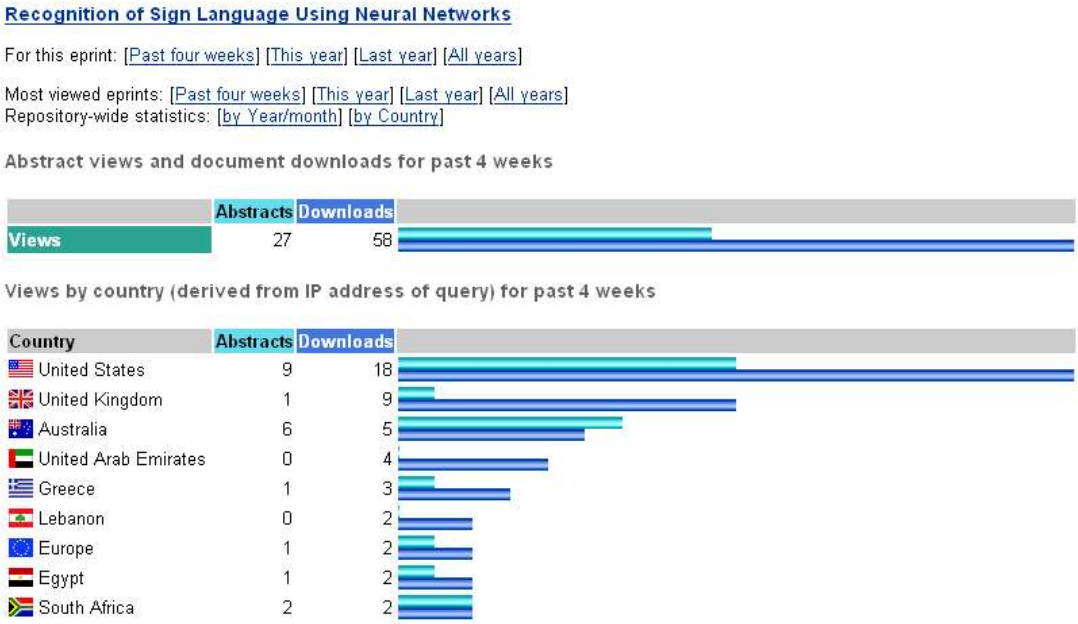

Figure 3 – Example of download statistics, first screen

There are several salient things to notice.

- As previously discussed, the number of downloads may exceed metadata views, indicating that some searchers are finding the full-text file without going through the OAI interface or the local search engine.
- This document (a PhD thesis) is downloaded from a variety of countries, but the USA, the UK and Australia predominate (there is a long tail of countries with lesser downloads).

Figure 4 shows the time series analysis of the same document. Something happened, probably in July/August 2005, to cause a surge in downloads of this document. The author's interest was piqued, and he traced the cause down to a citation of another paper of his, which itself cited this document. This resulted in him identifying the research of another person working in his field, half a world away.

History of views for this ePrint

Period	Abstracts	Downloads
2005 Dec	18	33
2005 Nov	26	60
2005 Oct	43	69
2005 Sep	28	73
2005 Aug	52	61
2005 Jul	22	8
2005 Jun	14	16
2005 May	29	16
2005 Apr	17	4
2005 Mar	14	11
2005 Feb	17	6
2005 Jan	15	1
2004 Dec	14	13
2004 Nov	17	7
2004 Oct	19	13

Figure 3 – Example of downloads by month

CVs and web sites

Besides becoming involved in the dissemination process, the OA repository can be a useful tool in reducing work for the researcher, and in making a case for promotion. Let's look at these two cases.

Some researchers realize quite early that if they self-archive their articles they do not need to mount the same articles on their personal or research website. They therefore modify their website so that instead of links to an internal copy of an article they provide a link to the persistent URL of its repository metadata page. This is an easy realization, and many make it instantly.

Another development may occur to the researcher, or as I have observed, it may spread like a meme. The researcher will delete all the links to articles, and all the papers on his or her website and instead they put a simple link which is a search on the repository for their name, of course with some text like 'Click here to see all my articles since 2003'. With one simple move they have simplified their website maintenance (the article lists never need to be updated) as long as they keep self-archiving. A similar approach may be used on websites devoted to a department's research, or to recruiting new graduate students, with even more saving in effort and better accuracy.

Promotion, improved jobs and grant success are cases dear to every researcher's heart. Citations have been estimated to be worth between $AUS100 to $AUS2000 per annum to a researcher in either direct income prospects or grant success, so the increased citation benefit of OA is obviously a plus. However, evidence from download statistics can also be quoted, especially as evidence accumulates about how they translate into citations. Some researchers have been observed to use download statistics, or download rankings, to mention in a promotion application or a grant application. The relevant committees are generally not yet sophisticated enough to fully realize what they are seeing, but they soon will be. The Internet generation is growing up into influencing decision-making at this level.

It is also possible to extract data (like a list of papers in a consistent format) for insertion in a *curriculum vitæ*, whether that be for a job application or a promotion application. This is simply using the repository as a personal database: convenient, accessible, provided by the institution, and backed-up by professionals.

Plagiarism

One feature of open access repositories that is seldom mentioned is their ability to detect plagiarism, and thereby lower the level of scientific fraud. One author was experimenting with a popular piece of plagiarism software, and tried it out on one of his own papers. He was interested to see that it turned up a substantial direct quotation from his paper by an author in another country, but less pleased to find that the quotation was unattributed. He took legal advice, and the offending author was contacted for redress.

This application worked only because the document was open access on the Internet. Conventional paper publications and toll-access journals cannot be searched for plagiarism.

→ ENDS ←

Links

ARROW Discovery Service: http://search.arrow.edu.au/apps/ArrowUI/
Australian Digital Theses Program: http://adt.caul.edu.au/
Google: http://www.google.com/
University of Tasmania repository (and statistics): http://eprints.comp.utas.edu.au:81/

References

Sale, Prof Arthur (1975) Pythagorean Triads. Technical Report R75-2, School of Computing, University of Tasmania. http://eprints.comp.utas.edu.au:81/archive/00000144/.

Sale, Prof Arthur (2005) The impact of mandatory policies on ETD acquisition. To appear in D-Lib Magazine. http://eprints.comp.utas.edu.au:81/archive/00000222/.

Sale, Prof Arthur (2005) Comparison of IR content policies in Australia. Preprint. http://eprints.comp.utas.edu.au:81/archive/00000230/.