Sustaining long-term access to open research resources – a university library perspective

In the third in a series of three blog posts, Dave Gerrard, a Technical Specialist Fellow from the Polonsky-Foundation-funded Digital Preservation at Oxford and Cambridge project, describes how he thinks university libraries might contribute to ensuring access to Open Research for the longer-term. The series began with Open Resources, who should pay, and continued with Sustaining open research resources – a funder perspective.

Blog post in a nutshell

This blog post works from the position that **the user-bases for Open Research repositories in specific scientific domains are often very different to those of institutional repositories managed by university libraries.**

It discusses how in the digital era we could deal with the differences between those user-bases more effectively. The upshot might be an approach to the management of Open Research that **requires both types of repository to work alongside each other, with differing responsibilities**, at least while the Open Research in question is still active.

And, while this proposed method of working together wouldn't clarify 'who is going to pay' entirely, it at least **clarifies who might be responsible for *finding* funding** for each aspect of the task of maintaining access in the long-term.

Designating a repository's user community for the long-term

Let's start with some definitions. One of the core models in Digital Preservation, the International Standard Open Archival Information System Reference Model (or OAIS) defines 'the long term' as:

"A period of time long enough for there to be concern about the impacts of changing technologies, including support for new media and data formats, and of a changing Designated Community, on the information being held in an OAIS. This period extends into the indefinite future."

This leads us to two further important concepts defined by the OAIS:

"**Designated Communities**" are *an identified group of potential Consumers who should be able to understand a particular set of information"*, i.e. the set of information collected by the 'archival information system'.

A "**Representation Information Network**" is the tool that allows the communities to explore the metadata which describes the core information collected. This metadata will consist of:

- descriptions of the data contained in the repository

- metadata about the software used to work with that data,
- the formats in which the data are stored and related to each other, and so forth.

In the example of the Virtual Fly Brain Platform repository discussed in the first post in this series, the Designated Community appears to be: "… neurobiologists [who want] to explore the detailed neuroanatomy, neuron connectivity and gene expression of *Drosophila melanogaster.*" And one of the key pieces of Representation Information, namely "how everything in the repository relates to everything else", is based upon a complex ontology of fly anatomy.

It is easy to conclude, therefore, that you really *do* need to be a neurobiologist to use the repository: it is fundamentally, deeply and unashamedly confusing to anyone else that might try to use it.

Tending towards a general audience

The concept of Designated Communities is one that, in my opinion, the OAIS Reference Model never adequately gets to grips with. For instance, the OAIS Model suggests including explanatory information in specialist repositories to make the content understandable to the general community.

Long term access within this definition thus implies designing repositories for Designated Communities consisting of what my co-Polonsky-Fellow Lee Pretlove describes as: "all of humanity, plus robots". The deluge of additional information that would need to be added to support this totally general resource would render it unusable; to aim at everybody is effectively aiming at nobody. And, crucially, "nobody" is precisely who is most likely to fund a "specialist repository for everyone", too.

History provides a solution

One way out of this impasse is to think about currently existing repositories of scientific information from more than 100 years ago. We maintain a fine example at Cambridge: The Darwin Correspondence Project, though it can't be compared directly to Virtual Fly Brain. The former doesn't contain specialist scientific information like that held by the latter – it holds letters, notebooks, diary entries etc – 'personal papers' in other words. These types of materials are what university archives tend to collect.

Repositories like Darwin Correspondence don't have "all of humanity, plus robots" Designated Communities, either. They're aimed at historians of science, and those researching the time period when the science was conducted. Such communities *tend* more towards the general than 'neurobiologists', but are still specialised enough to enable production and management of workable, usable, logical archives.

We don't have to wait for the professor to die any more

So we have two quite different types of repository. There's the 'ultra-specialised' Open Research repository for the Designated Community of researchers in the related domain,

and then there's the more general institutional 'special collection' repository containing materials that provide context to the science, such as correspondence between scientists, notebooks (which are becoming fully electronic), and rough 'back of the envelope' ideas. Sitting somewhere between the two are publications – the specialist repository might host early drafts and work in progress, while the institutional repository contains finished, publish work. And the institutional repository might also collect enough data to support these publications, too, like our own Apollo Repository does.

The way digital disrupts this relationship is quite simple: a scientist needs access to her 'personal papers' while she's still working, so, in the old days (i.e. more than 25 years ago) the archive couldn't take these while she was still active, and would often have to wait for the professor to retire, or even die, before such items could be donated. However, now everything is digital, the prof can both keep her "papers" locally *and deposit them at the same time*. The library special collection *doesn't need to wait for the professor to die* to get their hands on the context of her work. Or indeed, wait for her to become a professor.

Key issues this disruption raises

If we accept that specialist Open Research repositories are where researchers carry out their work, that the institutional repository role is to collect contextual material to help us understand that work further down the line, then what questions does this raise about how those managing these repositories might work together?

How will the relationship between archivists and researchers change?

The move to digital methods of working will change the relationships between scientists and archivists. Institutional repository staff will become increasingly obliged to forge relationships with scientists earlier in their careers. Of course, the archivists will need to work out which current research activity is likely to resonate most in future. Collection policies might have to be more closely in step with funding trends, for instance? Perhaps the university archivist of the digital future might spend a little more time hanging round the research office?

How will scientists' behaviour have to change?

A further outcome of being able to donate digitally is that **scientists become more responsible for managing their personal digital materials well**, so that it's easier to donate them as they go along. This has been well highlighted by another of the Polonsky Fellows, Sarah Mason at the Bodleian Libraries, who has delivered personal digital archiving training to staff at Oxford, in part based on advice from the Digital Preservation Coalition. The good news here is that such behaviour actually helps people keep their ongoing work neat and tidy, too.

How can we tell when the switch between Designated Communities occurs?

Is it the case that there is a 'switch-over' between the two types of Designated Community described above? Does the 'research lifecycle' actually include a phase where the active

science in a particular domain starts to die down, but the historical interest in that domain starts to increase? I expect that this might be the case, even though it's not in any of the lifecycle models I've seen, which mostly seem to model research as either continuing on a level perpetually, or stopping instantly. But such a phase is likely to vary greatly even between quite closely-related scientific domains. Variables such as the methods and technologies used to conduct the science, what impact the particular scientific domain has upon the public, to what degree theories within the domain conflict, indeed a plethora of factors, are likely to influence the answer.

How might two archives working side-by-side help manage digital obsolescence?

Not having access to the kit needed to work with scientific data in future is one of the biggest threats to genuine 'long-term' access to Open Research, but one that I think it really does fall to the university to mitigate. Active scientists using a dedicated, domain specific repository are by default going to be able to deal with the material in that repository: if one team deposits some material that others don't have the technology to use, then they will as a matter of course sort that out amongst themselves at the time, and they shouldn't have to concern themselves with what people will do 100 years later.

However, university repositories *do* have more of a responsibility to history, and a daunting responsibility it is. There is some good news here, though… For a start, universities have a good deal of purchasing power they can bring to bear upon equipment vendors, in order to insist, for example, that they produce hardware and software that creates data in formats that can be preserved easily, and to grant software licenses in perpetuity for preservation purposes.

What's more fundamental, though, is that **the very contextual materials I've argued that university special collections should be collecting from scientists 'as they go along' are the precise materials science historians of the future will use to work out how to use such "ancient" technology.**

Who pays?

The final, but perhaps most pressing question, is 'who pays for all this'? Well – I believe that managing long-term access to Open Research in two active repositories working together, with two distinct Designated Communities, at least might makes things a little clearer. Funding specialist Open Research repositories should be the responsibility of funders in that domain, but they shouldn't have to worry about long-term access to those resources. As long as the science is active enough that it's getting funded, then a proportion of that funding should go to the repositories that science needs to support it. The exact proportion should depend upon the *value* the repository brings – might be calculated using factors such as how much the repository is used, how much time using it saves, what researchers' time is worth, how many Research Excellence Framework brownie points (or similar) come about as a result of collaborations enabled by that repository, etc etc.

On the other hand, I believe that university / institutional repositories need to find quite separate funding for their archivists to start building relationships with those same

scientists, and working with them to both collect the context surrounding their science as they go along, and prepare for the time when the specialist repository needs to be mothballed. With such contextual materials in place, there don't seem to be too many insurmountable technical reasons why, when it's acknowledged that the "switch from one Designated Community to another" has reached the requisite tipping point, the university / institutional repository couldn't archive the whole of the specialist research repository, describe it sensibly using the contextual material they have collected from the relevant scientists as they've gone along, and then store it cheaply on a low-energy medium (i.e. tape, currently). It would then be "available" to those science historians that really wanted to have a go at understanding it in future, based on what they could piece together about it from all the contextual information held by the university in a more immediately accessible state.

Hence **the earlier the institutional repository can start forging relationships with researchers, the better.** But it's something for the institutional archive to worry about, and get the funding for, not the researcher.

Published 11 September 2017 via https://unlockingresearch-blog.lib.cam.ac.uk/?p=1596
Written by Dave Gerrard