

To protect and serve: making digital repositories safe and accessible for the long term

Tom Ruthven
Executive Officer
Australian partnership for Sustainable Repositories

Introduction

This paper discusses issues relating to the sustainability of repositories and the material that they hold. The issues are informed by what we consider to be a repository. Two things define a repository. Firstly, a repository is a place, a place where researchers can store objects and data safely, for the long term. Secondly, a repository provides access to the data and objects and as one researcher aptly put it to me: “a repository is not just a warehouse, it is a source of new ways of combining and representing material”.

Admittedly, this is a broad definition. An institutional-wide service using software specifically designed for a repository is covered by this definition. However, it also encompasses services in the many other guises that researchers know are repositories. Repositories may be based in a faculty, research school, library, IT area, or archive. They can be for use within a school, across the university, or shared between universities, and one university may have all of these. They can use software designed to run a repository or custom-built software that was designed for a specific service. What they have in common is providing access to material that is to be held in storage for a long while. It is this form of repositories that should be kept in mind when discussing sustainability: storing and providing access to research data and output.

For repositories fitting this broad definition, this paper examines the sustainability issues under three headings:

- the objects and data that sit in a repository,
- the experience of using the objects and data, and
- the business of running a repository.

The issues are intertwined so some issues crop up under more than one heading.

The objects and data

This is the “stuff” that researchers want to keep. Researchers produce and want to keep an enormous variety of material:

- Satellite images showing vegetation in Australia
- Health statistics on breast cancer
- PDF's of pre- and post-prints
- Sound recordings of songs from Tabar Island in Papua New Guinea
- Learning objects for a language course including flash animation and searchable scripts
- Blogs of Afghan war rugs

- Scholarly mailing lists recording 10 years of debates in nucleic acids research
- Video-recordings of conference presentations

Simple and standardised material is relatively easy to sustain. If all the images in the world were in TIFF, then there would be only one transformation to do to get to the next-big-thing. All effort could be focussed on making sure this transformation is done perfectly and then applied billions of times. However, choosing what material researchers add to a repository is not usually something that repository owners can control, nor should they. Rejecting material because it is not in an approved format will result in the loss of some important research material and dissuade some researchers from depositing material.

On the other hand, having no limits on formats becomes unsustainable because the effort required to provide access for each format through multiple channels is great and the effort to maintain access when hundreds or even thousands of formats and viewing software become obsolete is overwhelming.

A balance between having any format and very restricted formats is needed. This is most likely to be achieved through a partnership or collaboration between researchers and repository owners. Developing knowledge, agreeing on standards, and developing guidance on formats and data descriptions all form part of this partnership. Only some formats can be mandated such as the format for the lodgement of electronic versions of theses.

Once the material is in the repository, the researchers need to know it is authentic. Researchers are very happy when they find something valuable to their needs and doubly happy if there is integrity or provenance information to prove it is authentic. They do not become unhappy when there isn't any integrity or provenance information. It just means they will invest a lot of time to determine authenticity.

In the digital world, there are simple things we can do to assist researchers determine authenticity. By recording information about the source of the material and changes it has undergone we can provide a provenance trail for researchers to understand how it got to its current format. The trail can include the equipment and software used to create the material, and the processes and software used to make any changes to the material. Software tools are available to extract metadata from the material so people do not need to type much of this provenance trail technical metadata. Standards or widely used metadata tags assist here. Not only do they provide a template to use, such as ones derived from the work of PREMIS, they make the information interchangeable and more likely to be understandable in the future. Running regular checks for data consistency to ensure the data and objects have not changed, and comparing duplicate copies with "archival" back-up copies will identify problems early. Determining which is the authentic copy is assisted if the repository can produce comparison reports.

One of the nice things about a repository is that researchers know that someone else is looking at longevity and they can get on with doing their work, i.e. the responsibility for long-term preservation sits with the repository. There are a number of tools being developed to help repositories in this stewardship. This includes things like:

- The Global Format Registry which hopes to record enough information about formats and software so that when this information is long gone we can reconstruct what would be the equivalent of an arcane language, or it can be used to determine the transformation needed as a format becomes obsolete.

- Conversion to standard formats when adding material to the repository resulting in a manageable number of formats to keep watch over.
- System alerts to tell us which formats are becoming obsolete or unusable and which material in the repository is affected.
- Ensuring standard IT back-up regimes are in place in case of disasters to equipment.

The experience

Search access is just one aspect of how researchers use material in repositories. Other aspects are multiple access paths, look-and feel, and openness. All these add complexity to the sustainability of the material.

Broadening access is often cited as one of the goals of repositories. Making the output of research easily and widely available is easier with digital than printed material. Material, or its metadata, can be added to other services and metadata can be exposed to search engine crawlers. Consideration should also be given to forming alliances with other repositories to make like-information able to be found and used across multiple repositories. In the opposite direction, a repository can provide services to researchers by making it easy for them to find new material in other repositories or identify researchers who are working in similar areas by the material they are depositing around the world. Access can also be deepened by providing collaborative workspaces and allowing ways to combine material from different repositories.

Look-and-feel may be an important consideration. How data sets of astronomical observations or social statistics look on a screen may not be important as the data continues to be accurate. However, the screen design for a recording of a performance piece may be an integral part to understand the argument being propounded in a research paper. Consideration needs to be given to maintaining relationships between items and keeping the context of the material and this can be complicated when there are multiple pathways of access. In addition, the viewing and data manipulation software used by researchers can vary greatly. Standard, widely-used formats are more likely to be useable across different platforms. Researchers can be encouraged to use these formats when depositing material. Having a master version can also help. Variations based on the master can be presented to researchers, in the knowledge that they are not necessarily looking at the original presentation. Middleware tools may need to be developed to allow easy presentation and to standardise access by analysis tools.

Maintaining material into the future is only useful if the material can be used, and repositories generally start from the basis that the material is open to all for free. However, although much of the material may be open some may be closed and repositories need to provide various levels of access. A dispersed research group may want a closed workspace to add and share data whilst their analysis is underway with the intent to release the data when their findings are published. Permanent restrictions may be needed for sensitive material, with a mechanism to grant people access (and remove that access).

Sustaining these areas of “experience” adds a level of complexity to sustainability. When formats or software become obsolete, the material can be migrated or emulation tools can be devised. However, there is always a need to be faithful to the original and the original experience in order to understand the context of material.

The business

Repositories need to be in a home that itself will be there for the long term. Repositories are currently in departments and faculties, IT infrastructure groups, libraries and university archives. Understanding where the interest and energy is to keep the material safe and provide access is an important factor in deciding the right home. The energy to start up a repository may be in a different place to running a service, from an enthusiast in a faculty to benign archiving within a library. And, all of these may exist in the one university at the one time.

The material in repositories is used for research and in teaching and learning programs and perhaps the strongest foundation for longevity is the support the repository gives in these areas. It is clichéd to say a repository is valued if it is core business or an essential tool. However, behind these truisms lies the need for researchers to drive the development so the repository is designed to meet their immediate needs and their long-term aspirations.

Support by researchers can be generated in many ways. The key is to make life simpler and enabling better research. This will attract and keep users. Adding material to a repository needs to be a simple task, one that is a by-product of their work or integrated within it. It needs to provide new ways of combining and analysing data, possibly from multiple repositories. It can provide new services such as advice on which digital camera to buy; tools to extract information from objects and data which the researcher no longer has to type in; match-making services to find people depositing similar research in a repository half way around the world; a peer-reviewed area for objects, data and research results that confers prestige on the research; or a publishing framework to generate professional multi-media publications. It can be used to simplify the number of services that researchers need to deposit their material. For example, material placed in the repository could appear in course management software ready for use, or material deposited in the course management software could automatically be added to the repository. Above all, it is the researchers who should drive the priorities to develop the service and software. This is the key to them valuing the service offered by the repository.

Making the service known to researchers is a first step in getting their support. Advocating, publicising, cajoling and coaching are tools that can be used to assist researchers begin to use a repository.

Making repositories valued also implies a need to ensure that funding bodies within and outside the university see the benefits a repository bestows on the university. Researchers may lobby funding bodies to advocate the benefit of the repository service, however it will help if funding bodies are aware of the service and the benefits for the research output of the university. Providing background information and news of achievements will make it easier for funding bodies to understand the importance of the repository and make sense of comments and support from researchers. Keep funding bodies abreast of developments in the field and its relationship to other programs in the university also helps.

Another aspect is being able to influence the development of repository software. Currently, the groups developing software for repositories range from open source consortia to commercial companies. To meet researchers needs and hence provide a valued service requires this development to be informed by the priorities of

researchers. Having a voice in development priorities in this new software area is critical in the long-term business sustainability of the repository.

Lastly, there is linking with other like organisations, both locally and internationally. Repositories are relatively new things, and contributing to thinking, sharing practical implementation successes and failures, and discussing new ways of doing things will lead to a more robust business for all repository services.

Diana by Paul Anka, rendition of this song by Caetano Veloso, a Brazilian musician

I'm so young and you're so old
This my darling I've been told
I don't care just what they say
'Cause for ever I will pray
You and I will be as free
As the birds up in the trees
Oh, please, stay by me, Diana

Thrills I get when you hold me close
Oh my darling, you're the most
I love you but do you love me
Oh Diana can't you see
I love you with all my heart
And I hope we will never part
Oh, please, stay by me, Diana

Oh my darling, oh my lover
Tell me that there is no other
I love you with all my heart
Oh oh, oh oh, oh oh oh oh oh oh oh

Only you can take my heart
Only you can tear it apart
When you hold me in your loving arms
I can feel ill given all your charms
Hold me darling, hold me tight
Squeeze me babe with all your might
Oh, please, stay by me, Diana
Baby, I love you, Diana
Baby, I love you, Diana