Digital Preservation: the need for an open source digital archival and preservation system for small to medium sized collections,

Kevin Bradley

ABSTRACT:

Though the solution to all of the problems of digital preservation of all types of digital objects is a convoluted and difficult problem, the sustainable preservation of simple digital materials, such as sound or images, is quite well understood. However, though the processes are understood for this subset of the digital heritage, simple, complete and affordable tools to undertake all the preservation tasks are not available. Thus small collections that exist to preserve such content do not have the tools to undertake that process. As a first step in addressing this issue the MoW Sub Committee on Technology released a document called "Towards an Open Source Repository and Preservation System: Recommendations on the Implementation of an Open Source Digital Archival and Preservation System and on Related Software Development" which argues that many of the components of such a possible system exist, but that there is a need to develop a software package that draws together the components, and to build a support community around it. The document lays out a broad specification for such a tool and suggests where development is needed. This talk will discuss the findings described in the document and describe developments and plans to achieve its aims.

Introduction

Some of the most technically advanced Libraries in the world, have been working in developing standards, making tools and technologies available, and encouraging all sorts of digital services with impressive results. The problems that these complex objects, and the changing standards of the infrastructure that supports them, presents to the digital archivist is convoluted, and while apparently not totally intractable, certainly difficult. The tools, systems and software that they will develop will take account of these many variables, technological changes and shifting standards. In effect the digital archival community broadly agrees on what to do, but is locked into resolving the details around complex digital objects.

The problems that these complex objects, and the changing standards of the infrastructure that supports them, presents to the digital archivist is convoluted, and while apparently not totally intractable, certainly difficult. The tools, systems and software that they will develop will take account of these many variables, technological changes and shifting standards.

The direction digital preservation must take is known, but the details and complex relationships which must be resolved are still being pursued. And it is a matter of record that the array of resources and expertise are making inroads into the solution of all this

digital complexity.

But what would need to be done if we were dealing with digital simplicity rather than digital complexity? Is then, the problem already solved? Let me make a second point before I explain what I mean.

Interestingly for our discussion, almost all of the digital repositories exist within established institutions and have dedicated staff with technological expertise up to the task of wrestling with aspects of these issues. These include, for example, National Libraries, National Archives, Universities and other places of learning and research, and media and cultural museums and archives.

The IT backup regime and data security is integrated into the core business of those institutions and is almost invariably invisible to the repositories' managers: It is part of the infrastructure you need to manage a large digital resource. In fact, the existence of the backroom technical aspects and data security are so taken for granted that the development of most of the repository software tends to leave these issues to one side. In other words, most of the repository solutions assume the existence of a technological infrastructure without which their systems would not survive.

So the point of this paper, and UNESCO MoW report, "Towards an Open Source Repository and Preservation System: Recommendations on the Implementation of an Open Source Digital Archival and Preservation System and on Related Software Development" http://portal.unesco.org/ci/en/ev.php-URL_ID=24700&URL_DO=DO_TOPIC&URL_SECTION=201.html on which it is based, is to imagine a scenario where there is a need to preserve a collection of simple digital objects, but where a digital preservation infrastructure has not yet been developed. In other words; to develop a sustainable preservation standard digital management and storage system for a collecting institution that doesn't happen to be one of the world leaders in digital preservation.

Digital Simplicity and the Smaller Collections

Digital simplicity involves archival responsibility for only a small number of files and formats. Such an approach is not in any way new. The DSpace consortium defines different levels and a three level support scheme, while stipulating that it is the responsibility of the host institution to "determine the exact meaning of each support level, after careful consideration of costs and requirements" (DSpace 2004). The three levels of the scheme are *Supported*, in which the format is recognised and future access and usability is guaranteed, *Known*, in which the format is recognised, can be retrieved and it is hoped will be accessible in the future, and *Unsupported* in which usability is not guaranteed and only the retrieval of the bit-stream is possible.

Supporting particular formats and relegating the others to known and unsupported impacts on both the depositor and the repository differently according to the type of resource. Some aspects of sustainability are easily implemented, while other are less well defined. The difference between a sustainable format, and a distributable format is often

very marked, and frequently both are required to maintain a sustainable repository. The types of standard formats may be grouped under the following categories simple digital objects: images, sound recordings, textual content, and video, which are the basic components for more complex objects such as reports, web sites and multimedia works.

If an archive has control over the content of the material they are creating, then digital simplicity is achievable, it can define the standard and select only full bandwidth robust formats. Thus the solution I'm suggesting is mostly about a digitisation process, such as the digitisation of images or the preservation conversion of sound recordings, or video tapes, or the creation in the digital domain of original items for long term reasons, such as audio, video, images or test.

The size of the collection is another aspect of digital simplicity. Large scale collections need large scale solutions. As the number of items in a collection and the size of the storage environment increase, the way that data is managed changes fundamentally. A small scale archive requires less sophisticated management technology. For the purposes of this report we identified the break point at less than 20 terabyte.
The issue is to recognise that the socio-technical aspects of a smaller scaled digital collection are fundamentally different to the requirements of a large and complex collection. The solution for the latter does not scale down to suit the former.

Digital simplicity and the OAIS

The parts of digital preservation (6 bits)
Ingest
Access
Administration
Preservation Planning
Archival Storage
Data Management

So what do the small scale collections do now if they want to store and preserve their digital collections? Frequently, and regrettably, they use optical carriers such as CD or DVD because there appears to the novice to be no technical barrier to creating CDs.

The Risks of Optical Carriers

The UNESCO Memory of the World (MoW) Sub Committee on Technology (SCoT) published in September 2006 its combined knowledge on the way CDs and DVDs should be managed in a document entitled "Risks Associated with the Use of Recordable CDs and DVDs as Reliable Storage Media in Archival Collections - Strategies and Alternatives" (http://unesdoc.unesco.org/images/0014/001477/147782E.pdf).

The conclusion of the document is that these carriers are risky in terms of their reliability and that the cost of mitigating that risk makes using CDs economically much less viable. Besides which, an approach which tries to preserve content on un-networked, discrete

carriers cannot take advantage of the many developing tools and technologies which will aid in long term preservation. So the main impediments to using a reasonable open source system is probably not cost, or at least not just cost, but complexity.

The need for a low cost, open source digital preservation repository and storage system which uses more sustainable technologies than optical disc is clear; and the report recommends ways that one might be constructed.

What did the Report Find?

The initial part of the report sets out to see what open source tools are available and how they fulfil the requirements of an OAIS planned sustainable preservation repository.

The technologies which have practical implementation for these principles exist, however they are generally expensive and complex, or such low cost options as are available are fragmented and dispersed. The open source community has developed systems and support for systems which meet the needs of many users world wide, some of which are directly relevant to the aims of this project.

These include Repository software, and other tools and services such as some Preservation planning tools, Ingest tools, Data sharing (metadata harvest), Migration tools, Obsolescence notification tools and services.

A number of open source initiatives address one or more of these functions, however, not all necessary functions have been fully addressed and no single open source or low cost system addresses all, or even most, of these issues. The consequence of this is that only commercial systems can currently meet all the requirements of a sustainable digital storage system, though many of these commercial systems take advantage of the open source developments.

To expand, repository software exists which fulfils most of the requirements of three or four of the functional categories of OAIS, that is; Ingest, Access and Administration and data management. It is important to note that repository software addresses only some of the aspects of digital preservation, and generally assumes that the technical aspects are taken care of “elsewhere”. Henry Gladney observes this when he states that most approaches to digital preservation “fail to distinguish between digital repository and digital preservation. The former topic is well developed, with software offerings that have been refined for about a decade.” (Gladney 2004).

Though there are various individual tools and projects, none of the major repository software implements preservation planning to any extent. Preservation planning for a small scale institution which is solely interested in the preservation of its own simple digital objects is a readily identifiable task. While a well funded national institution must be involved in research and development of the major digital preservation problems, a smaller institution requires much less. It need only have a system with an architecture that will allow it to take advantage of the solutions that the major initiatives are

developing, and to this end the system must have the ability to acquire the necessary information for that purpose.

Though some tools exist, the functions of archival storage tend to be components of the commercial systems which sell data storage.

The end result of this is that only well funded or technically proficient archival institutions can afford to undertake this approach. It also means that, somewhat ironically, only well funded or technically proficient archival institutions can take advantage of the free and open development work being undertaken.

This report recommends that UNESCO supports the aggregation and development of an open source archival system, building on, and drawing together existing open source programs. This report also recommends that UNESCO supports and facilitates the development of an open source distributor who can provide support along the lines of existing providers for other desk top services.

Approaches for small collections:

Though a particular institution may be responsible for the management of a collection or set of audio items, it does not necessarily follow that that institution will undertake the responsibility for maintaining the digital storage system. They may establish a relationship with a third party provider. That provider may be another archive which will take the collection and store its content, or may be a commercial provider who will provide and manage the storage and content for a fee. However, they should still be able to manage and exchange data and will still have a requirement for many, if not all, of the functionality described in the report.

So the specification of the system in the report expects that the archive can manage data on their own site, and create and exchange data packages and standard exchange formats between them and other suppliers. Standardised approaches are critical to preservation, and also protect the archive from being locked into proprietary or commercial systems.

What the report suggests

1. UNESCO establish a steering committee based in the MoW Sub Committee on Technology to support the development of a single package open source digital preservation and access repository
2. Support and resource a pilot project with a number of communities or institutions who can articulate their requirements and act as beta testers of such a system
3. Through that and other committees and projects, influence and support the development of specific software, as discussed in this report
4. Investigate the development of solutions to the system gaps noted in this report, particularly in the area of preservation planning and archival storage systems
5. Support the integration of a number of open source tools to develop a single package open source repository system based on existing open source platforms as described in this report

6. Encourage the development of federated and cooperative approaches through the adoption of standard data packages
7. Ensure that, low cost notwithstanding, the solution is based in international standards and best practice.
8. Support and expand existing training and education to include technical training in the envisaged system in parallel with work on intellectual property and cultural rights.
9. Liaise with existing open source distributors such as Ubuntu, or with development communities, such as the Australian Partnership for Sustainable Repositories (or other suitable) to support these aims.

Future Plans

To build a successful system the project need partners, collection partners, development partners, and distribution partners.

In 2007 UNESCO, through IFAP, the Information for all Programme, agreed to provide $80,000 as a partial contribution towards the project in order to attract private sector inputs and hopefully obtain the full amount required for the development work.

Currently discussions are underway with potential partners, including the Australian National University, who we hope will host the project, and with MEMNON, a Brussels based archiving company, who are prepared to work with the project to develop audio ingest tools. We hope to develop a steering committee from out of the Memory of the World Sub Committee on Technology, while drawing in partners with appropriate collection to test the ideas and help develop the functional specification.

The vision

The intention is to build a system that meets the needs of the managers of small to medium sized collections. The system will be designed on the assumption that the system is being implemented in an environment with low, but not non-existent, capital investment. This report examines how it is possible to build a low cost technology system, but cannot escape the conclusion that there must be some level of technical knowledge and recurrent resources, albeit at a low level, to make it sustainable. Regardless of the design complexity and robustness of the system, it will need to be replaced at some time or risk losing the content it manages.

The role of the project implementers is to build the system: the sustainability of the system is dependent on the uptake of the system, and the only way that a system will be used is if the user community are provided with appropriate functionality. To this end it is vital that the control of the project, its future design and development be handed over to a true open source community of users. In the initial stages, the role of the collection partners will be to test and provide advice on the approaches envisaged.

The system must be simple to install and initiate, and must provide user support. To maintain this, the whole approach must eventually become community and grant funded, modelling on the business models of other open source software aggregators.

Most importantly, the system must be standards based, for even though we are proposing a simple digital system, the simplicity must not be at the expense of compliance with standards. Standards will allow the system to be integrated with international digital preservation developments and so take advantage of the many tools which are under development.

The aim for this archival repository project should be to build an easy to use, low cost system that the community of users supports, and to which it contributes. The initial project should be guided by best practice, but informed by a pragmatic approach. The hardware should be available as a set of options selected from affordable solutions, and the software written and distributed in an open, supported way. Eventually the community of users should guide its direction and manage its development so that it becomes a truly open and responsive system.