Applying Open Storage to Institutional Repositories

Dave Tarrant, Ben O'Steen[1], Steve Hitchcock, Neil Jefferies[1] and Leslie Carr

IAM Group, School of Electronics and Computer Science, University of Southampton,
SO17 1BJ, UK
{D.Tarrant, S.Hitchcock, L.Carr}@ecs.soton.ac.uk
[1] Oxford University Library Services, Systems and Electronic Resources Service, Osney
One Building, Osney Mead, Oxford OX2 0EW, UK
{Benjamin.Osteen, neil.jefferies}@sers.ox.ac.uk

Repository interoperability and the capability to support preservation can be enhanced by introducing a storage layer that is independent of repository software.

Institutional Repositories (IRs) are largely characterized by 'openness', that is, most are based on open source software, conform with the Open Archives Initiative (OAI) and aim to provide open access to content and data. We introduce a new 'open' approach to repositories: open storage combines open source software with standard hardware storage architectures. Examples include platforms provided by Sun Microsystems, which we use in this work.

The paper will describe how the open storage approach has been allied to the OAI framework for Object Reuse and Exchange (ORE) to enable repositories managed with different softwares to share and copy data more easily and to be provided with extra services such as preservation services

To date, repository interoperability has been founded on the OAI protocol for metadata harvesting. This protocol supports harvesting and aggregation by service providers of metadata descriptions of content stored in repositories. The effect of this is to make the content of repositories collectively visible. In the context of Web search, OAI effectively builds an ivory tower around registered content providers.

The launch of OAI in 2001, because it made IRs feasible on a wide scale, was accompanied by the release of the first software to build IRs, EPrints. Other popular open source IR software, including DSpace and Fedora, emerged in later years.

To facilitate the exchange of contents between repositories, not just metadata but digital objects, OAI recently introduced a specification for ORE. This specification includes 'approaches for representing digital objects' and facilitates 'access and ingest' of these representations 'beyond the borders of hosting repositories', enabling 'a new generation of cross-repository services'. In this way ORE standardises the description of the relationship between digital objects. This relation could be between versions of an object, such as might be found in a repository record, or aggregations of objects - such as a Web page with images, or a collection of chapters for a book -

to form a new object. Such aggregations are sometimes referred to as complex, or compound, objects.

In combination with open storage, ORE can be used to facilitate a low level (storage based) representation of complex digital objects. A high level (repository based) ORE approach has been demonstrated as a means by which objects can be copied between repository softwares in a lossless manner. This was achieved through the extension of the EPrints and Fedora repository softwares – previously thought to have quite different, perhaps incompatible, metadata and storage architectures – to include built-in services to allow importing and exporting of ORE representations. By implementing additional, necessary conventions – conventions that are extensions to the OAI-ORE specifications – the repository software is able use the resource maps to create a new, local instantiation of an object that previously existed elsewhere

Using open storage averts the need for a repository layer to access first-class objects – these are objects that can be addressed directly – where first-class objects include metadata files which point to other first-class objects (such as an ORE representation). We can now begin to realize situations where an institution can exploit the resulting flexibility of repository services and storage: multiple repository softwares can run over a single set of digital objects; in turn these digital objects can be distributed and/or replicated over many open storage platforms.

Early adopters of open storage include Sun Microsystems, which is developing a large scale open source storage platform (codenamed Honeycomb). By focusing on object storage rather than file storage the Honeycomb server (STK5800) provides a resilient storage mechanism with a built-in metadata layer. The metadata layer provides a key component in open storage where objects are given an identifier (URI), often by the repository, which the storage layer translates to a physical location when a user requests that URI. This same layer can also be used to store information about each first-class object, or in some cases the object itself (most likely in a differing representation).

Digital preservation can benefit from storage layer services. Although digital preservation involves a range of processes and techniques, one fundamental approach is to create copies of content on systems independent of the host repository, e.g. simple backup, bitstream storage, or software-supported approaches such as Lots Of Copies Keep Stuff Safe (LOCKSS).

In the UK-based Preserv 2 project we are implementing this ORE-based approach to replicate content from selected repositories to a Honeycomb machine, initially providing reliable bitstream storage. In turn our open storage server facilitates preservation services for IRs, for example file format identification and risk analysis for digital objects. By allowing services to run directly on the storage platform, preservation services can be run without affecting the performance of the repository. The results can be reported to the repository manager who can then sanction services to act on at-risk files.