Implementing an open jurisdicctional digital repository - the STORS project

Lloyd Sokvitne
State Library of Tasmania
<lloyd.sokvitne@education.tas.gov.au>

Jan Lavelle
State Library of Tasmania
<jan.lavelle@education.tas.gov.au>

Abstract

This article discusses the development and implementation of an open jurisdicctional repository for published electronic material by the State Library of Tasmania, Australia. This repository is called the Stable Tasmanian Open Repository Service (STORS). It operates within the legal deposit provisions of the Tasmanian legislation and includes electronic documents published in Tasmania by government, commercial publishers and individuals. STORS is based on a publisher self-contribution model for content acquisition. The project has focused on the need to ensure easy and reliable contribution of content by publishers, with the provision of additional benefits to encourage use. STORS has been developed in a modular and extensible way, and will be supplemented over time to improve discovery, interoperability and preservation outcomes.

Introduction

The State Library of Tasmania\(^1\) is the legal deposit library for all books published in the state of Tasmania, Australia. This legal requirement, enshrined most recently in the Libraries Act 1984 (Tasmania. Department of Premier and Cabinet 2003), defines books in a broad and format-neutral way, so as to include material in electronic and digital formats.

Following their introduction, electronic publications in physical formats (floppy disks, CDs, etc.) have been accepted and treated the same as print publications. Since 1998 the State Library has operated a web publishing archive, whereby State Library staff have used the powers of the Libraries Act 1984 to select and download Tasmanian websites, modify them to operate independently on a State Library server, and provide access to those preserved web sites through a public web service entitled Our Digital Island\(^2\). This experience led the State Library to realise that a special repository for electronic documents would eventually be needed.

In 2000 the State Library of Tasmania was charged with the responsibility to provide the state government services portal, Service Tasmania Online. This meant that the State Library had increased responsibilities to ensure access to online government information for the Tasmanian community, including long-term access to government electronic publications.

The State Library subsequently undertook a project to define and establish a document repository and archive service. Special project funds were provided by Service Tasmania in late 2002 to develop this service which was in due course titled the Stable Tasmanian Open Repository Service, or STORS.

**Current status of digital repositories**

A review of the literature indicated to the State Library that the need for a digital archive for published electronic documents has been widely acknowledged, and the number of such archives has begun to grow (DSpace, ePrints, etc.). There are two basic reasons for this growth.

Firstly, there is the recognition that as electronic documents disappear from their original file systems or web locations, specific archives are needed to ensure retention and access. Secondly, a web archive or repository can make documents placed in them openly available that might otherwise be locked up by copyright and other publishing restrictions.

For these reasons, various types of institutions have established archives that provide enduring storage, either for the needs of that institution, or for sectors (e.g., education), or target groups (e.g., academic authors). Such archives can work in isolation or network and overlap in coverage.

The variety of repository implementations (including those for learning objects in the education sector) suggested to the State Library that most of the groundwork existed for the technical implementation of a basic repository. Following on from this, the State Library set out to develop the business rules and processes for governing a unique repository that would be both jurisdictional (covering an entire Australian state) and open (accepting contributions over the web from all types of publishers) in scope.

Throughout this process, the State Library has been guided by the Open Archival Information System (OAIS) reference model (Consultative Committee for Space Data Systems 2002) and continues to use this model to ensure sufficient data and process granularity to enable developmental growth, particularly to ensure future functional interoperability and archival extensibility.

---

The STORS jurisdictional model - self-contribution

From the beginning, the State Library saw STORS as a dynamic and developmental service that would continue to grow in functionality over time. The Library's first objective, however, was to develop a service that enabled the easy ingestion of document-like objects, in commonly used electronic formats, from the widest possible range of publishers in Tasmania.

To meet this objective the State Library sought to develop a model that relied on publishers as the primary agents in the contribution process. Experience from Our Digital Island had demonstrated that the Library could not afford to provide the internal processes or personnel to identify and capture appropriate Tasmanian content for the repository.

The self-contribution repository model meant that STORS would have to cope with contributors from a variety of backgrounds, and who would have varying levels of skill and commitment. In other words, the submission process would have to be as simple as possible.

Content scope

The scope established for STORS was that it should accept all Tasmanian content produced or otherwise published in Tasmania, including content from government, community, commercial, and personal sectors. Published is here taken to mean that the document has been made publicly available, without access restrictions, over the Internet from a server located in Tasmania.

Format scope

STORS was developed to focus initially on the acquisition of document-like objects with inherent web-browser compatibility. Basically, STORS content should be accessible using a normal web browser and common, freely available plug-ins.

As the phrase document-like objects is difficult to define succinctly, STORS attempts to focus on discrete items that have a distinct conceptual and physical boundary, rather than open-ended and dynamic entities such as web sites and databases.

The State Library decided to defer the development of STORS as a repository for entire web sites, for complex multi-part programmes, or for data sets that require specialist underlying software or applications to operate. Such content is so inherently difficult to manage that the State Library will need to rely on the wider preservation community to develop solutions that it can in turn adopt.
Encouraging publisher participation

The State Library felt that it was important to develop STORS in such a way as to provide a number of positive business benefits that would encourage publisher participation (and ensure wider content acquisition).

Immediate access to an enduring URL

The State Library decided to provide publishers with the ability to use STORS as a location for current documents as well as archived material. In this way STORS would acquire documents early in the publishing cycle. To do this STORS needed to provide an immediate and enduring URL that allowed publishers to use and access their items from the time of contribution. In OAIS terms, STORS aimed to encourage Tasmanian publishers to play both Producer and Consumer roles.

Such a service would remove the need for publishers to store and manage publications on their own file systems, and the enduring nature of the URL would also eliminate the need for publishers to maintain web links or references. This URL is in the format STORS domain/STORS unique ID (e.g., http://www.stors.tas.gov.au/au-7-0010-00001).

If STORS was to assume a role in the publishing cycle, it also had to facilitate the external discovery of deposited documents. For this reason STORS selectively allows contributed items to be indexed by web harvesters and search engines.

Document contextual relationships

It is extremely important that the end user understand the context, accuracy and currency of a document when it is delivered externally to the repository application. The presentation of the document within a normal web browser is neutral and uninformative, and the user will not know whether the document is still current and reliable, or inaccurate and out of date.

STORS solves this problem through the use of contextual metadata. The contribution process allows the publisher to indicate whether or not there is a date by which the resource will become invalid or outdated, or whether there are earlier or subsequent versions.

STORS uses this contextual metadata and a resolver service to provide an intermediary screen whenever a resource is requested. This screen informs the user of the context of the requested document, informs the user if the document has become outdated, and provides links to superseding or superseded versions within the repository. Documents are always accessed via this intermediary screen to avoid confusion over context. Any external requests for direct access to the document are intercepted by the resolver service and sent back to the intermediary page.
Format conversion

The State Library wished to ensure broad access to the general community to the content contained in the STORS repository. Much web publishing (especially in government) utilises proprietary formats such as Word or PDF. However, users from the general community do not all possess the software to access these formats.

The State Library decided that the STORS service should provide an on-demand file conversion service for documents in such proprietary formats.
The objective was for these documents to be converted into HTML4, a format that is inherently accessible with current browser technology. The State Library originally hoped to automate this process within STORS, but file conversion has remained a manual process. This is largely because the State Library has so far been unable to resolve all of the problems inherent in accurately converting some formats, such as PDF, into HTML.

Notwithstanding these problems the State Library has accepted the responsibility to continue to develop or acquire functionality to convert STORS documents into more accessible formats. This will become especially important over time as current formats become inaccessible.

Document verification

STORS provides publishers with a document MD5 checksum\(^5\) upon contribution. This checksum can be used in the future to verify the authenticity or legitimacy of electronic documents that purport to be true copies or versions of the original. It was felt that this feature would encourage contributions from a variety of government publishers.

Commercial risk mitigation

Although most material on STORS is to be openly accessible, the Library felt that it was important not to harm the viability of commercial publications by making them freely available over the Internet through STORS. For this reason, STORS allows publishers to indicate when there should be restricted access to a document, with open access only being allowed from a computer within State Library premises.

Metadata structure

The metadata fields provided in STORS are minimal and constitute only the basic data required to provide core outcomes whilst ensuring easy submission by a wide range of publishers with varying training and skill levels.

The metadata fields in STORS are:

1. title,
2. copyright owner authorisation confirmation,
3. dates of validity of the item,
4. other STORS documents that supersede or are superseded by this resource,
5. restrictions on use,

---

\(^5\) Message Digest Algorithm #5 by RSA Data Security: a one-way operation that transforms a data string into a unique highly secure value.
6. contributor,

7. date of contribution,

8. MD5 checksum,

9. unique identifier, and

10. persistent URL.

Only the first five of these are provided by registered contributors; the rest are provided by the system.

Additional metadata fields will be added over time to support enhanced discovery and preservation functionality.

**Risk management**

**Content and contributor management**

Because the STORS contributions process can be accessed by anyone on the web, the State Library had to prevent the accidental or intentional submission of inappropriate content. This is achieved by developing two types of contributor logon.

The first process is username and password based. Any Tasmanian publisher can register as a STORS contributor and receive a username and password. This is a free service and items from these contributors go directly into the repository without review, and are immediately accessible via the STORS enduring URL.

The second type of contribution is based on an anonymous logon without password. This is very easy for occasional publishers to use and operates in a similar fashion, but the contributions received are then reviewed by State Library staff before acceptance into the repository.

**Copyright**

All submissions require that the contributor confirm that they either own the copyright of the submitted item or are acting on behalf of the copyright owner. Contributors must also confirm that they agree that STORS is permitted to copy or translate that item into other formats for preservation purposes.

**Malicious software**

Material contributed to STORS is checked for viruses and other malicious software upon contribution.
STORS implementation

Software development

The State Library examined several software options during the STORS project, including DSpace (Smith, Mackenzie, et al 2003), EPrints (EPrints.Org 2003) and a learning objects repository developed locally for the Tasmanian Department of Education, of which the State Library is a part. The Library chose the learning objects repository for STORS because it already provided core repository functionality, was highly configurable, and was supported by local developmental expertise.

The learning objects repository stored metadata in an XML database and the objects themselves in a separate file system. Relatively few enhancements were needed for STORS and this work was completed by the developer, Dytech Solutions, in June 2003.

The major enhancement required was the provision of an immediately available and enduring URL. This is achieved through a local resolver service. Upon contribution, a unique ID is created and assigned to that item. That ID is appended to the high level STORS URL to create the public enduring URL. When this public URL is requested, the STORS server takes that address and interprets it as a query into the resolver service. The document is then located by its unique ID and delivered to the end user. Its true location in an internal directory or file system is irrelevant in so far as its access on the web is concerned. The resolver also ensures that the document is always accessed via the STORS intermediary page.
Usability testing

The State Library employed usability testing to help determine the specific business process underlying publisher contribution. Given the reliance within STORS on self-contribution, anything that slowed down or complicated the contribution process could jeopardise the entire service. The Library conducted usability tests with a variety of government employees including administrative/clerical staff and librarians. The results showed that the success rate for contribution was lowered to as little as 40% if even a moderate amount of discovery metadata was required from the contributors.

These results clearly dictated that only a minimal amount of information (consistent with the basic functioning of the repository) could realistically be required from contributors during the submission process.

Launch

The STORS service and web site became available on 1 July 2003 for government publishers. It was officially launched and made available to the public on 2 December 2003.
Since its inception, a number of government agencies have become contributors. Several agencies have instituted projects to enter their electronic documents into STORS retrospectively, and in one case, an agency is planning a project to digitise paper copies of its documents going back to the early 1900s and to then contribute them to STORS.

As well as meeting Tasmanian legal deposit requirements, STORS has been recognised as a place of safe deposit in the current Disposal Schedule for Common Administrative Functions issued by the Archives Office of Tasmania (Archives Office of Tasmania 2003).

**Promotion**

Given that STORS relies on publishers to contribute content themselves, the State Library is keen to ensure that STORS is well publicised and promoted.

The State Library's initial promotion focus has been on Tasmanian government agencies. During 2004, personnel from the State Library will provide an information session for each agency, followed by specific demonstrations, best-practice guides, and personal assistance. A key objective is to simplify the development of formalised contribution processes within each agency.

Particular efforts are made to engage with libraries within the agencies, as the libraries are likely to have a strong commitment to the collection and preservation of their agency's publications.

Promotion to other types of publishers has been integrated with the normal legal deposit processes already in place within the State Library. This will include contacting publishers to inform them of their legal deposit obligations and, if necessary, reminding them that STORS is a simple and easy way to meet those obligations.

**Future developments**

STORS has been developed in a modular and extensible way to allow for future enhancement and additions.

Development work has recently begun on two new STORS applications. These include a discovery module for general users and an integrity checking service. The discovery module will provide a simple web interface to STORS that allows searching and browsing, without requiring users to log in to the system. The integrity checking service will verify the authenticity of stored objects in the file system on a scheduled basis (e.g., nightly) and warn of unauthorised or malicious changes before those changes supplant valid backups.

In the medium term the goal is to add additional preservation and object representational data, better file conversion applications and business processes, and metadata interchange capabilities using the Open Archives...
Initiative Protocol for Metadata Harvesting (OAI-PMH) ([Open Archives Initiative](http://www.openarchives.org) 2002).

**Conclusion**

STORS is a starting point for the State Library of Tasmania in the digital repository area. It has allowed the State Library to gain internal expertise, to communicate and inform publishers about digital preservation issues, and to position itself to expand and develop repository functionality in the future. But most importantly, it has already allowed the acquisition of significant Tasmanian digital content that would otherwise have been lost.

**References**


