USE OF OPEN SOURCE SOFTWARE AT THE NATIONAL LIBRARY OF AUSTRALIA

ABSTRACT

The National Library of Australia has been a long-term user of open source software to support generic Internet services and software development. More recently the Library has actively investigated the use of open source software to support core Library activities. This paper uses a case study approach to review the Library's experience with open source software.

BACKGROUND

Open source software refers to any program whose source code is made available for use or modification as users or other developers see fit. Open source software is usually developed as a public collaboration and made freely available. The main features that characterise open source software is the freedom that users have to:

- Use the software as they wish, for whatever they wish, on as many computers as they wish, in any technically appropriate situation.
- Have the software at their disposal to fit it to their needs. This includes improving it, fixing its bugs, augmenting its functionality, and studying its operation.
- Redistribute the software to other users, who could themselves use it according to their own needs.

Open source software has played a strong role in the development of the Internet with, for example, a majority of the world's web sites using the open source web server, Apache. The second most popular web browser, Firefox, is also open source software.

The open source model offers a number of potential advantages to the Library:

- Innovation and integration of services based on open source is easier and can be undertaken through a variety of mechanisms, whereas propriety systems generally require the vendor to both agree to and undertake the work.
- Collaborative development crossing institutional and international boundaries is common with open source software and attractive in areas such web archiving or digital preservation where no single institution or commercial player is likely to be able to develop and support a comprehensive system.
- The ownership costs associated with open source are potentially less than with propriety software, especially in the areas of acquisition (basic software is usually free), maintenance (software upgrades and bug fixes are often provided free of charge) and transition (because the systems are open it is easier/cheaper to migrate to a new system)

The Library has been a long-term user of open source software to support generic Internet services and software development. More recently the Library has actively investigated the use of open source software to support core Library activities. In 2006–07, the Library successfully completed a project to define the IT architecture that will be needed to support the management, discovery and delivery of the National Library of Australia's collections over the next three years. The project report identified the use of open source software as important tool for building and maintaining digital library services with the capability and flexibility required by the Library. This paper uses a case study approach to review the Library's experience with open source software.

SEARCH ENGINE

In 2005, the Library undertook a procurement exercise to select a search system capable of supporting both the Library's extensive metadata collections and its growing collections of full-text material. The Library adopted the open source search engine software called Lucene. This software not only met the Library's functional requirements but also offered better performance than a range of commercial software systems. It was also the only economically feasible solution for the Library as all the commercial products had per document licensing that tied the cost of the software to the number of documents that were to be searched. The Library's Web Archive contains in excess of one billion documents and none of the commercial solutions would licence this number of documents for a cost within the Library's budget constraints.

Lucene, like the majority of commercial products considered, lacked built-in support for an important library search protocol called Z39.50. As Lucene was open source software the Library had the option to either develop this capability itself or identify a commercial partner to undertake the work. In this case a software development company with experience in this area was identified and the Library was able to have the capability added to Lucene in a cost-effective and timely manner. The resultant software module was released back into the open source community for others to use and adapt as need be.

Since its introduction, Lucene has been used to provide a search service across the Library's Web Archives; it is being used as the basis of two highly successful prototyping exercises investigating next-generation search and discovery services for library materials; it is currently being used as the basis of the search and discovery system for the Australian newspapers project; and it also powers the Library's recently completed replacement of the online public access catalogue. In the future, the Library plans to replace the search engines that power its website search and federated discovery services (e.g. Libraries Australia) with Lucene.

Lucene has been easily integrated into the Library's development framework and has been a productive tool for both prototyping and production work. To date the search engine has been easy to support and has required certainly no more and possibly less maintenance than the existing proprietary search engine products used by Library. Lucene is an example where an open source software solution has proved to be superior to and more cost-effective than a commercial solution. This is not an argument that open source solutions are always superior to commercial solutions, merely that they can be in some cases and should be considered as part of the procurement of new software systems.

REPLACING THE LIBRARY'S ONLINE CATALOGUE

The Library's catalogue is the primary mechanism for providing effective user access to its collections. Unfortunately many users find the typical online public access catalogue such as that provided by the Library as an unfriendly and sometimes frustrating tool to use when compared with a modern Internet search engine such as Google.

In late 2007, a small team comprising staff from Information Technology, Collections Management, and Reader Services was formed to evaluate an open source online catalogue called VuFind. Using this software they were able to rapidly develop a replacement online catalogue for the Library. It combines the functionality of a traditional library catalogue with features found in modern web applications, and aims to make Library resources easier to discover and request. The new catalogue went into production in early 2008.

The Library believes the new catalogue is superior to the existing catalogue in all respects. Its interface is cleaner and provides relevance ranked search results that means users are likely to find what they are looking for quickly rather than having to wade through many pages of search results. Analysis of search logs indicate that some 95% of searches on the new catalogue will result in a useful item being found on the first search results page, compared with only 35% of searches on the existing catalogue. It presents the user with "clusters" of documents grouped by similarity allowing users to refine their search and easily explore the Library collections. The new catalogue provides a simplified form to request items from the Library's collections. It is integrated with the Library's digital collections and allows users to view items online where available. It supports interaction in the form of comments allowing users to see what others thought of an item and also to make their own annotation about particular collection items. The new catalogue is integrated with wikipedia, google books and LibraryThing to provide additional information not usually found a library catalogue.

The importance of open source software in this project was the ability it gave the Library to rapidly and easily modify the software to meet both the Library's and importantly our users' requirements. During much of its development the new catalogue was available as a public beta and users were invited to comment on their experiences using the catalogue. The project team, because of the light-weight and open nature of the VuFind software, was generally able to update the beta catalogue quickly (often in a matter of hours) in response to user feedback. This ability to be responsive has been fundamental in achieving improvements that the new catalogue provides.

In the spirit of open source software collaboration, the Library is contributing its software development and improvements back into the VuFind project.

WEB CURATOR TOOL

Beginning in 1996, the Library has developed a number of software tools to support PANDORA: Australia's Web archive. While these tools serve the Library's needs, they require constant enhancement in order to deal with the changing technical nature of Web publication and also to provide increased efficiency so that the Library can collect a meaningful and useful sample of Web publications. This requires a level of software development and support by the Library that is increasingly difficult and ultimately impossible to sustain. An early attempt by the Library to engage a wider community in the development and use of these tools failed as the tools had been designed specifically to meet the needs of the Library and could not easily be adapted to use by other institutions. Under the auspices of the International Internet Preservation Consortium, the Library contributed to a project to develop a next-generation set of web archiving tools. In this case the model of open source development was that the Library use its expertise in this field to develop and test specifications. Other agencies in the project team were responsible for software development. This was a successful collaboration that resulted in an open source set of tools called the Web Curator Tools. These are being deployed by a number of institutions throughout the world. The Library will replace its current web archiving system when it reaches end-of-life with a new service based on the Web Curator tools.

The Library is currently taking part in a project to develop a next-generation open source Integrated Library System to support the management and discovery of Library materials both physical and digital. This project is being funded by the Mellon Foundation and led by Duke University. Again, the Library's role in this project is to use its expertise to design and specify such a system rather than undertake the development. The Library's involvement in this project complements its existing strategic priority to develop efficient workflows for collection and processing, and if successful may influence the design of an Integrated Library System that better needs meets the needs of this library than any of product currently available either in the proprietary or open source world.

The Web Curator Tool and the Open Integrated Library System design project are examples of how the open collaborative nature of open source software can benefit an institution like the Library. They illustrate how the Library can use its expertise to influence the design of the tools that may ultimately be highly beneficial to the Library. It is generally impossible to influence a commercial off-the-shelf product in this way and historically this has required the Library to build specialised systems from scratch either through outsourcing arrangements or using its own in-house resources. In both cases this is expensive to do and, more importantly, expensive to maintain.

OPEN JOURNAL SYSTEM

In 2005/06 the Library created an ‘Open Publish’ web space using the Open Journal Systems (OJS) digital publishing open source software to manage, host and deliver an online open access journal service. The refereed publishing workflow supported by OJS requires no intervention by the Library and is managed entirely by the journal’s authors and editors. The decision to create this space was based on a collaborative trial project with the Association for the Study of Australian Literature. The Library’s prime motivator was to engage with the online community as well as to learn about online journal hosting services.

The Library was able to quickly and easily install the software using its existing IT infrastructure. The software requires minimal staff involvement to set up each journal and minimal ongoing support. This service is an example of a project that could not and would not have been undertaken if it had required the purchase of commercial software or the development of customisation of software internally.

During the course of the trial, three refereed journals began publishing on the Open Publish service, namely:

- Journal of the Association for the Study of Australian Literature
- Australasian Journal of Victorian Studies
- Reviews in Australian Studies

A successful outcome of this trial has informed the Library’s decision to include open published journals in the Library’s collections.

DIGITAL PRESERVATION

It has long been recognised by the Library and other collecting institutions that the preservation of digital materials requires immediate, active, and ongoing intervention. Moreover, the likely volume of digital material is such that for preservation to be practical it will require the development of effective automated and semi-automatic preservation tools. The Automatic Obsolescence Notification stage 2 (AONS II) project, undertaken by the National Library of Australia in conjunction with the Australian Partnership for Sustainable Repositories (APSR), aimed to refine the tool from an earlier stage of APSR, to a platform independent tool that automatically provides information from the authoritative International registries to support decisions on preservation actions required to retain access to information resources stored in repositories.

In this project the decision was made to develop the tools in an open source manner. The reason to this was twofold. Firstly, the environments and repositories these tools would be applied to developing and changing rapidly. Secondly, it is unlikely that any single institution will have both the expertise and resources to support the ongoing development of these tools. By placing the software developed in the open source space, it is hoped that future work in this area will build on the work done in this project rather than begin from scratch, and even if the tools themselves are not developed further they provide a valuable demonstration of capabilities that can be used in the design of future tools.

This project is notable because it is one of few where the Library has taken a leadership role in the development of open source software and has served as valuable learning exercise. In particular, it emphasised the need to make a decision to develop an open source manner early is essential so that the code developed is portable and not dependent on a particular institution’s IT infrastructure. It also illustrated the not inconsiderable overheads are associated with documentation and communication that is needed to support open source software development.

CONCLUSION

Open source software development offers a range of potential advantages to an institution such as the Library. Our experience shows that many of these advantages can be realised in practice. The Library has found that open source software can be more cost-effective solution than that provided by proprietary commercial software. More importantly open source software provides the capability for the Library to rapidly prototype or customise services they could not afford to do using commercial or in-house developed software.

It is interesting to note that the commercial sector is also showing increasing interest in the use of open source software. A number of companies have emerged whose business model is based around providing support for open source software or provided providing commercial value-added extra capability on top of an open source software based product. Recently, Relais, a supplier of specialised document supply software used by the Library announced that they will release their formerly proprietary software into the open source software space.