

# Critical issues in Open Source repository software and its contribution to Open Access

By Dr Edmund Balnaves

Prosentient Systems, [ejb@prosentient.com.au](mailto:ejb@prosentient.com.au)

The phenomenal growth in electronic publishing has not been matched with equal development in the management, ownership and preservation issues surrounding the intellectual output of institutions. While many universities have established institutional repositories, there exist outside our major institutions many special libraries and research organisations that want to achieve better knowledge management and intellectual property control over the output of their organisations.

The availability of well accepted, robust, open source applications for institutional repositories is acting as an enabler for organisations to implement open access systems to support their research output. This can include not only journal publications, but also research reports, image collections and many other assets. The example of Dspace will be presented in the context of developing an enterprise model for open source in libraries and the potential this has to foster open access, and the critical issues around adoption of open source repositories, including: institutional adoption, persistent referencing, digital archiving, content ingestion and the mitigation of obsolescence.

## INTRODUCTION – WHY OPEN ACCESS?

There is a growing body of published experience in the process of building digital libraries in major institutions, including the work of JISC (<http://www.jisc.ac.uk/>) and Europeana (<http://www.europeana.eu/>). Prosentient Systems works principally with special libraries in the Asia Pacific region to deliver open source library systems. These libraries cover a diverse range of areas, from government departments through to medical research centres. Some, like the Heart Foundation in Australia have a narrow but quite specialist research output. Others, like the Queensland institute of Medical Research, have a large research output. Typically, the libraries in these services rely on either the traditional website to deliver their publications or the standard journal publication model (which entails the institutions funding research which they provide at no cost to publishers and then buy back again through journal subscriptions). Two movements have come together as enablers for special libraries to make better use of their institutional resources: open source and open access. While the term “open access” may generally imply the unrestricted external visibility of publications, open access can also be important in leveraging the research output of an organisation within the institution (even where copyright restrictions prevent external visibility).

## WHY OPEN SOURCE?

One of the synergies in modern library systems has centred on the relationship between open source systems (OSS) and open access (OA). Adoption of OSS and OA has been extraordinary over the last ten years, has progressed in parallel, and in similar timeframes. The synergies go deeper than this. First, open source can be an enabler for the adoption of open access in an institution. Equally a successful OA project can justify the ongoing improvement of the OSS implementation. Second, OSS can provide a level of certainty for an institution in their operation costs. The larger the community of adopters of open source the stronger the overall support. Third, the OSS can provide a level of security in that there is no

proprietary lock-in and the code is visible (and therefore can be corrected). The functional depth of this security will be improved by the work of those adopting the open source model.

Finally, open source systems can provide a cost-viable model for implementation of open access in smaller institutions. A common confusion is that open source means “free”. While it may be lower cost, no information technology system operation is free. The ongoing nurturing of a system, software upgrades over time, support for customisations and enhancements, server administration, network costs are just a few of the base-line elements of managing an information system. Nevertheless, the amortisation of the software support across a wide installed base makes for an effective cost model for smaller institutions. This paper examines the critical issues around open source as an enabler of open access.

## **IT CAN'T BE THAT EASY - THE CRITICAL ISSUES**

### **Building internal acceptance – institutional adoption**

The internal acceptance of open access is one of the obstacles to adoption. Reluctance can come in several forms: resistance to open source (based on lack of trust or experience with these systems) and lack of take-up among institutional clients. Some of the pathways to internal institutional adoption include:

- Ensuring submission on of content is intuitive and straightforward
- Ensuring that the process of submission guarantees the correct visibility of content according to relevant rights for access

A number of open access guides to building a digital library are available, including the IFLA publication “Designing and Building Integrated Digital Library Systems - Guidelines” (Rathje, McGrory et al. 2004). The groundwork laid by institutions through their implementations provides a well-chronicled pathway to the technological implementation, including extensive documentation on OSS implementation processes.

### **Obsolescence**

The rapidity of technological development brings long-term difficulties in the management of intellectual and creative output in digital form. Libraries and museums have a key role in the preservation of analytical and creative endeavours over the long term. However, most libraries are ill equipped to undertake research into the preservation of new media artefacts and creations. Where the preservation of printed works is well understood, issues of obsolescence of new media technologies affect all aspects of the new media artefacts. As each new technological innovation introduces new methods of creative content delivery, our long-term horizons of archive planning appear to reduce. The widespread adoption of Information Technology as an integral part of the research process, and the speciation of software vehicles for content creation, mean that on the basis both of cost and volume of content creation the meagre budgets of most libraries simple are not sufficient to sustain the role of comprehensive collection builders. Longstaff, Chittister et al (2001) argue that the risks associated with moving critical business and government infrastructure to an Information Technology framework are poorly addressed(Longstaff, Chittister et al. 2001). They identified several adverse impacts at the national level of increased reliance on Information Technology:

- “Increased complexity of our information systems because of the added interconnectedness and interdependencies between and among infrastructures
- Reduced operational buffer zone in most infrastructures, and the ever-increasing adherence to the just-in-time philosophy as a vehicle for cost reduction and efficient operation.
- Enhanced accessibility of would-be terrorists to our defence, banking and financial institutions, and to other critical infrastructures.” (p.43)

Digital library collection building has associated with it inherent risks of technological obsolescence. In addition to the **systematic** risks associated to critical information technology architecture, are the problems of software and hardware obsolescence. Most libraries are in no position financially to undertake fundamental research in areas of technological migration from different versions encoding, media storage or content delivery platform (Ekman 2000).

- The management issues go beyond simple system continuity management, including the management in the long term of all aspects of:
- Hardware architectural requirements to operate the software. In the context of dynamic websites, this might include the server hardware, the client hardware, and intermediary network architectures.
- The physical and logical storage media requirements.
- The operating system and network communication protocols used to delivery /retrieve the information
- The layered software products to interpret/present the content.

Issues of obsolescence are not inherent obstacles to the move to open access – but they are issues that need to be addressed by the institution in the planning for implementation of a digital library. Consideration of these issues should be considered part of good ethical practice in the establishment of such systems. Information systems inevitably go through a continuous series of transformations over time, as do digital objects stored in an information system. One of the core methods of mitigating obsolescence is to build in an open source framework. Here again open source can play a part: while there are inevitable operating system and application interdependencies in open source as with any software platform, the ability to support the continuous transformation required for long-term management of digital objects is less problematic where there is no proprietary component to the software and where all software components are open to scrutiny and change.

### **On the move – persistent referencing**

Your open access implementation will change over time. These changes may entail institutional name changes, website redesigns or changes to the website platform. The identification, access and archival characteristics of a traditional print publication and an electronic resource located by a Uniform Resource Identifier (URI) are very different (see Table 1). Only time will tell whether Internet delivery of content is not in fact *essentially* ephemeral. The print form of a book or a journal has the virtue of a static nature: the content is the same for all readers for a given publication. Distributed access is simple. Personalisation, on the other hand, dispenses with any degree of finality of information

delivery: the content delivery may be different for each individual. Without a fixed point of reference in which content can be thought to have reached a “final” form – that is, which is essentially dynamic, the issues of attempting to preserve content in its final generated form become problematic.

PRINT PUBLICATIONS	URL CITATION
Content is static	Content may be variable
Content can be sourced uniquely through international delivery mechanisms	Content may only be available through one source only (the website)
Content is widely distributed (multiple repositories)	The location of the content may change

**Table 1: contrasting print and electronic resources**

The URI is an apt metaphor for the difficulties of unique content identification in the digital era. Digital content design and publishing systems lend themselves to a rapid rate of content design and delivery. The corollary to this is the rapidity of the entropy effect on this content. (Lawrence, Pennock et al. 2001) examined 270,977 computer science journals, conference papers and technical reports, extracting 67,577 URI references. They demonstrated both a dramatic increase in the use of URI’s in citations and the substantial increase in broken links over time, peaking at 53% after 6 years. This highlights the ephemeral nature of URI referencing, even within academic publications: “First, personal homepages tend to disappear when researchers move. Second, many who restructure Web sites fail to maintain old links. These problems are likely to persist without improved citation practices” (Lawrence, Pennock et al. 2001) p.28. In practice the Uniform Resource Identifier as used in most websites is far from satisfactory as a location identifier for electronic resources over the long term. The longitudinal study by Bar-Ilan and Peritz (2009) confirmed this issue – with only 19% of pages remaining stable over the period of the study.

One way of supporting the portability of electronic resources through website and organisational changes is the use of Digital Object Identifiers (DOI). These generally entail the registration of objects through a central referencing agency that provides a proxy-based reference to the current web page/resource location. Dspace, for instance, includes full integration with the public DOI handle.net service (Corporation for National Research Initiatives 2010). It also incorporates functionality to host and manage your own DOI handle service.

### **Ingestion and indigestion**

One of the challenges to institutional acceptance is the efficiency of the ingestion process. The more complex the process the less likely the institutional acceptance. Here again, open source lends a hand. The open design of such systems lends itself to the addition of “plug-ins” adapted by institutions to suite their local preference for file uploads. Dspace for instance supports several paths for file uploads, including:

- An integrated, highly structured and configurable web-based workflow system
- A batch-oriented file upload system for bulk ingestion
- Use of plug-ins or internally built workflows

The broad institutional adoption of a system is one of the factors that can be a factor in the breadth of support in functionality “around” the product that supports such functions.

### **Breathing life into the OPAC – an open source vision**

Clients want digital information – and in many cases will only use the resources that are in digital format. More than that, ubiquitous discoverability is also important, and here an open, web-2.0 enabled OPAC can play a part as it may provide metadata discovery and annotation abilities that go beyond that of the digital library system. In this context Prosentient Systems is encouraging an enterprise vision of open source that takes advantage of the best aspects of both systems. One of the elements of institutional adoption of repository software in conjunction with web-2.0 enabling emerging from open source catalogues is the breathing of life and excitement into the library catalogue. The catalogue can now become a metadata hub and an interactive, web 2.0 enabled gateway to resources. This can include:

- User tagging
- User reviews
- Indexing and searching of institutional repository content, including OAI (Open access Initiative) harvesting capabilities
- Federated views of other content and resources

Figure 2 illustrates how open source can be the glue to build an enterprise library architecture that can include:

- Single sign-on architectures
- Content management
- Digital library management
- Web-2.0 enabled catalogue / federated searching

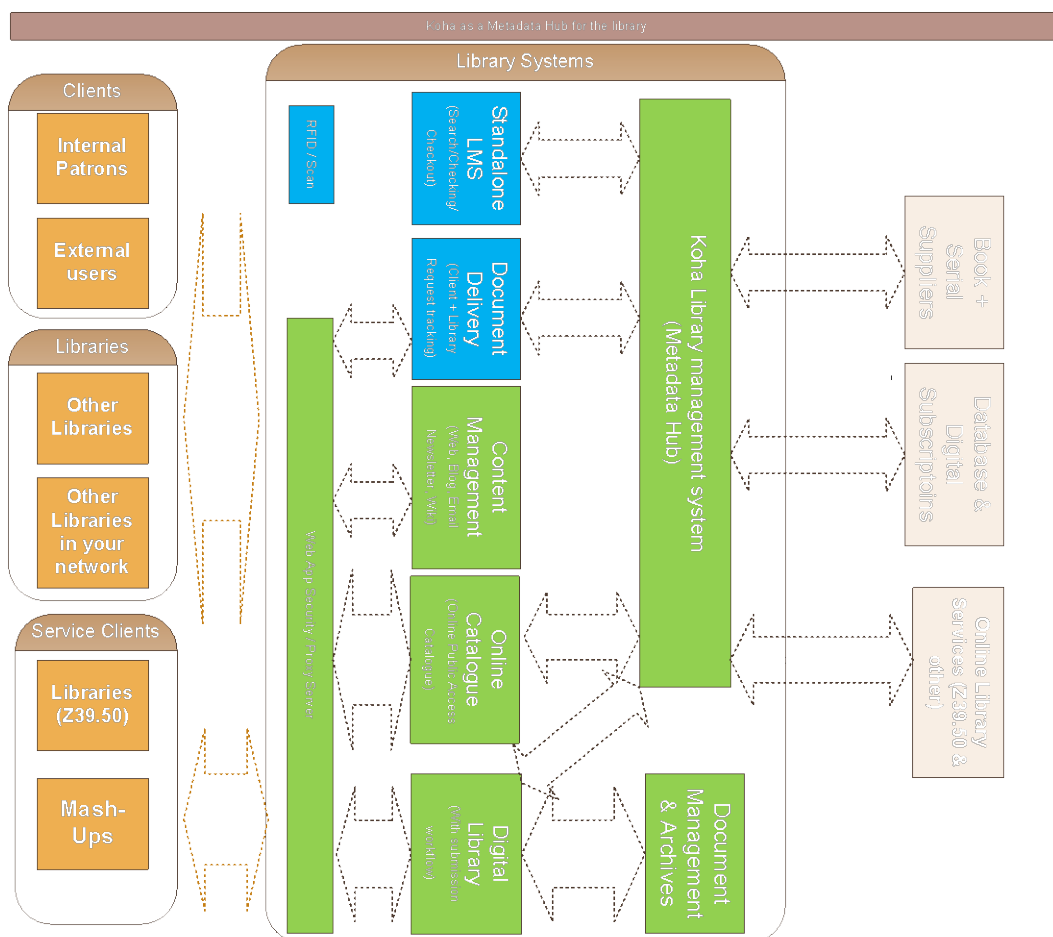


Figure 1: The Inter-Search integrated library management system framework

## Distributed archival models

A final, but more challenging issue for sustainability of open access is the evolution of a robust model for distributed content archives. A responsible model for long term sustainability of open access publication should include robust mechanisms for replication of content between archives. There are models for approach this, either on an escrow basis (Balnaves 2005) or through an open source process of peer-to-peer replication or indeed through periodic crawling of web resources (Balnaves and Chehade 2008) or through a national deposit schemes. The barriers to systematic models for distributed archival management of open access resources centre not so much on the technology but on the copyright issues of the digital objects involved. More work is needed on the management of copyright issues and the copyright “envelope” around each digital object and how copyright metadata can itself be distributed, persisted and honoured between systems.

## CONCLUSION

Prosentient Systems deals every day with Special Libraries, and I am constantly impressed by the extraordinary results they delivery with often meagre budgets. The bulk of the library profession has, of necessity (and perhaps inclination), oriented to setting strong standards and leaving the delivery of information systems in the hands of commercial vendors. The demand

for compliance to strong standards has borne dividends in the facility with which bibliographic data can be exchanged nearly seamlessly. Nevertheless, the nature of information delivery in organisations is gradually but ineluctably changing with the myriad of new ways for the information dissemination and delivery. Information providers, publishers and information agents are offering new services directly to clients where they would once have dealt with such issues through the library - and may now view the library as a dispensable "middle man".

Marginalisation of the library service is a possibility in this context. The degree to which this marginalisation is avoidable depends a lot on the time it takes our profession to reskill in a manner that maintains the relevance of the library and the profession. This re-skilling entails much closer engagement with the technology that library patrons are already immersed in. It also entails projection of the strengths that the profession can historically claim to the coherent organisation of information. Where does open source fit into this? The last decade has opened the possibilities for direct engagement by libraries with the technology that they are working with. A decade ago there were limited opportunities to lift the lid on the technology that commercial library vendors provided their clients, other than demanding compliance with broad library standards. The availability of a wide range of open source tools and technologies offers an opportunity for direct technical engagement in a number of ways: by learning web 2.0 technologies that allow direct re-engagement with the library patron base; by lifting the lid on the technology itself by downloading and installing open source systems - and even changing and developing these systems; by contributing to the community enhancement of the software.

These are real challenges and opportunities - one that Special Libraries are experience right at the coal-face. Special libraries are the canary in the coal mine for the library profession. They can be more easily dispensed of by organisations than public or state/national libraries, if they are seen to be of marginal worth. I have seen the adoption of open source systems happen as a practical response of increasing information capability in a context of a diminishing budget. Open Access is an enabler for better ownership and visibility of organisational resources. There is a very clear role for librarians to ensure that the information management in organisations is coherent, effective and sustainable. There is also a role in engagement across the multiple modes of communication. There is a role also in being the hub of knowledge about the media assets and technologies relevant to the organisation. All of these entail a level of technological engagement that is richest when it is informed by a willingness to experiment with the technology - to try out the tools and the new modes of communication. While it is challenging, it is also pretty interesting. Open source and open access have a pivotal role here, as your institution does not have to wait for a vendor-delivered course in the latest software release. You can get online and try it out, download it, even change it!

## **Biography**

Dr Edmund Balnaves is the Information Officer for the IT Section of IFLA. He is an active proponent of open source solutions for libraries and has established hosting services for special libraries in the Asia-Pacific region to encourage adoption of open source. His doctoral research in the area of systematic content reuse in the world wide web frames his interest in effective, dynamic, information systems for libraries and web-based publishing. He is the editor of the IFLA IT Section Newsletter and has presented on a range of topics in the area of digital archiving, escrow management of subscriptions, evaluation of open source systems and systematic content reuse.

## REFERENCES

- Balnaves, E. (2005). "Systematic Approaches to Long Term Digital Collection Management " Literary and Linguistic Computing **20**(4): 399-413.
- Balnaves, E. and M. Chehade (2008). Digital archiving of e-journals for Special libraries. IFLA World Congress 2008, Quebec city, Quebec.
- Bar-Ilan, J. and B. C. Peritz (2009). "The lifespan of "informetrics" on the Web: An eight year study (1998–2006)." Scientometrics **79**(1): 7–25.
- Corporation for National Research Initiatives. (2010). "Handle System: Unique Persistent Identifiers for Internet Resources." from <http://handle.net>.
- Ekman, R. H. (2000). "Can Libraries of Digital Materials Last Forever?" Change **32**(2): 23-35.
- Lawrence, S., D. M. Pennock, et al. (2001). "Persistence of Web References In Scientific Research." Computer **34**(2): 26-31.
- Longstaff, T. A., C. Chittister, et al. (2001). "Are We Forgetting the Risks of Information Technology." Computer **33**(12): 43-51.
- Phillips, M. E. (1998). "Tomorrow's incunabula: preservation of Internet publications." Lasie **29**: 5-10.
- Rathje, B. D., M. McGrory, et al. (2004). "Designing and Building Integrated Digital Library Systems - Guidelines." Retrieved 10/7/2010, 2010, from <http://archive.ifla.org/VII/s31/pub/Profrep90.pdf>.
- Smith, W. (1998). "Lost in cyberspace: preservation challenges of Australian Internet resources." Lasie **29**: 6-25.