Date : 28/05/2008

Digital archiving of e-journals for Special libraries

Dr Edmund Balnaves
Prosentient Systems
Sydney
Australia 2007
ejb@prosentient.com.au

Mark Chehade
Prosentient Systems
Sydney
Australia 2007
chehade@prosentient.com.au

Meeting: **159. Information Technology**
Simultaneous Interpretation: Not available

WORLD LIBRARY AND INFORMATION CONGRESS: 74TH IFLA GENERAL CONFERENCE AND COUNCIL
10-14 August 2008, Québec, Canada
http://www.ifla.org/iv/ifla74/index.htm

Abstract

Special libraries are not well resourced to undertake their own e-journal archiving initiatives, and are consequently vulnerable to supplier changes in e-journal supply. Partially funded by the ALIA 2007 study grant, this paper reports on a proof-of-concept research into the design of a "Smart Client" application (easily deployed, supports both online and offline operation and uses a blend of local and web-based resources) for digital archiving of e-journal subscriptions held by special libraries. This research comprises an initial survey of archiving experience in Australian special libraries, and architectural design and proof-of-concept implementation of an archiving application. Responses from 164 Australian special libraries indicate that many libraries already encounter loss of subscription coverage from a range of causes, and that very few have an e-journal archiving strategy. Our objective is to implement an archiving application that is suitable for installation in a special library context where information technology support may be minimal. The paper presents results of the archiving survey, the architectural design of the smart client application, and outcomes of proof-of-concept trials. The research outcomes indicate that a simple-to-use cross-platform "Smart Client" approach to e-journal archiving is viable and can be distributed in an open source framework.

Introduction

Australian special libraries are facing the challenge of managing the transition to e-journal collection building. With this transition comes the issue of archival management of journal holdings, traditionally a role of the library rather than publisher. Larger national and university institutions are undertaking their own research projects and co-operating in building archival services. However special libraries rarely have resources to undertake e-journal archiving projects. In 2007 Dr Balnaves proposed a research project to explore e-journal archiving for special libraries. In this paper we report on the results of a survey into current practice in Australian special libraries and also report on a of a proof of concept implementation built around open source tools to provide a Smart Client application for digital archiving of e-journal subscriptions for special libraries.

The transition to e-journal collections

Researchers consistently show a preference for electronic resources (Brown 2007), and this is driving a progressive transition to electronic delivery of research materials. In 2007 Prosentient Systems undertook a survey of the 263 member libraries of the GratisNet Inter-library loan network on their current practice in electronic subscriptions and e-journal archiving. 164 responses were received. Only 7 of these libraries reported any current work in e-journal archiving. Unlike academic institutions, special libraries still retain a large percentage of print subscriptions, with only 28 respondents (17%) having more than half of their collection in electronic-only form. However 58% of libraries had at least 30% of their collection in both electronic and print form, and 38% of libraries have at least 10% of their subscription in electronic form only.

Unlike traditional print subscriptions, access to E-Journals is not guaranteed and in the case of extended service outages, libraries who have cancelled their print subscriptions for the alternative of an electronic subscription face the dilemma of losing access to the historical and current collection. This risk is more than theoretical: 57% of libraries reported a loss of access to an e-journal subscription (e.g. through supplier changes). The 14 libraries that indicated that e-journal archiving was a priority were also those that have substantial e-journal-only collections, indicating that these libraries are conscious of the risks associated with this subscription method.

Neither the publisher nor the subscription agent necessarily takes responsibility for indefinite continuity of access for electronic subscriptions, although some now make deposit arrangements with organisations such as OCLC (Online Computer Library Center). This presents a significant dilemma for subscribing institutions:

> "The archiving of digital materials is generally not provided for in license agreements. Publishers generally draft these license agreements and frequently do not discuss this issue. The monetary value of the license agreement is in its ability to generate recurring annual fees. Archiving of materials previously licensed tends to decrease the amount of fees available under a license agreement. Librarians generally will have to ask for the right to archive digital materials." (Alford 2002)

The short-term risks of moving to e-journal subscription include the application stability and change management of the providers service, network access and operational continuity of the client and internet technology infrastructure. Long-term risks are systemic to information technology, not the least of which is long-term certainty in resource delivery (Balnaves 2005). The record of long-term persistence of Web-based resources is generally poor even over durations as short as five years (Lawrence, Pennock et al. 2001).

Previous work

There are many approaches to managing the archiving challenge. A centralised approach is most common and assumes reliance on a central (generally internet-accessible) archive available to all participating libraries. This approach has the advantage of cost-efficiency and centralises the resolution of IP issues and management of the technical an systems issues An independent approach aims to deliver capability to the library itself to manage and control a local archive. Such an approach places a greater burden on the individual library but has the advantage of freedom from wider connectivity and service risks.

OCLC has had a central role in providing centralised e-journal access and archiving. They currently provide access to almost 6,000 journals from 70 publishers (Machovec 2006). OCLC has an online journal archive system which allows libraries to sustain persistent access to their electronic subscriptions. OCLC acts as a third party archiving agent by establishing an access account on behalf of the subscribing library which allows users to view full text of articles from their subscribed journals. The OCLC Electronic Collections Online program takes the approach of creating a central journal archive, and has gone to some lengths to ensure that licence agreements have been made with the various publishers.

Another online archiving initiative is PORTICO, which stems from JSTOR, also an electronic journal archive. "The mission of Portico is to preserve scholarly literature published in electronic form and to ensure that these materials remain accessible to future scholars, researchers, and students"(PORTICO 2007). As with the OCLC Electronic Collections Online program, PORTICO is a central online scholarly journal archive and electronic document resource.

The reliance on a central, possibly trans-national, e-journal archive may not be satisfactory to some institutions, for reasons of cost or reliability of online access. A risk analysis incorporating issues of business continuity and long term certainty of collection preservation may also mandate local archiving for some institutions (Balnaves 2005).

There are many other institutional archiving projects associated with Digital Library initiatives, notable among them being the Yale Electronic Archive (YEA) is a joint project with the publisher Elsevier, directed to establishing a substantial archive for escrow and archival management of their digital journal collection. The YEA project demonstrates that collaboration between publishers and subscribers for the establishment of a digital archive is economically feasible (Yale University Library 2002) and can enhance the reputation of the supplier by providing continuity of service assurance to the client. There are also many Digital library initiatives in academic and national institutions directed to the preservation of core national collections (Oltmans and van Wijngaarden 2004).

There is one significant local archiving initiative. The LOCKSS (Lots Of Copies Keeps Stuff Safe) project is a free peer-to-peer solution to digital preservation and access. LOCKSS was developed at Stanford University and is an open source application that allows libraries to build their own digital collections and provides librarians with a way to collect, store, preserve, and provide access to their own, local copy of authorized content they purchase. The LOCKSS system converts a computer into a digital preservation "appliance" in the library that, with a publisher's permission, non-invasively collects specific content to which the library has access. If content is not available to the user from the publisher's site, it can be delivered transparently from the stored content, with no need for intervention by publisher or librarian (Kubilius and Walton 2005).

The LOCKSS program is run from a self-booting disk containing a minimal version of the Linux operating system, without installing the operating system on the hard drive. It assures archival continuity both through local archiving and peer-to-peer distributed archival management (Maniatis, Roussopoulos et al. 2005). By using a peer to peer auditing protocol, LCAP (Library Cache Auditing Protocol), LOCKSS is able to run checks and comparisons between caches to repair any damaged data that may exist in one of the participating web caches.

The LOCKSS system requires a level of technical skill that may limit its adoption by smaller libraries that lack information technology skills. The peer-to-peer approach also has several risks, including cache corruption, or deliberate injection of incorrect data that spreads through the peer-to-peer architecture. Viral and injection attacks are common in the Internet and cannot be excluded as a risk to this architecture. Peer-to-peer architectures may not be well regarded by publishers, who face the risk of distribution of journal articles beyond the licensing terms or boundaries that a library may properly hold. There is room for further experimentation with local archiving approaches, and alternative approaches can help to add to the open source resource base in this area.

Design principles for the Inter-Store Smart Client

In 2006 the Australian Library and Information Association (ALIA) awarded the ALIA Research Study award to Edmund Balnaves for proposed research into proof-of-concept for an e-journal archiving system suitable for Special libraries. The subsequent research was jointly funded through this study award and Prosentient Systems. The target open source application is intended to be deployed in libraries as a lightweight, “Smart Client” that can be installed on most desktops. A “Smart Client” is an application that optimises resources locally and through web services connectivity, and can operate in both an offline and online manner. Development frameworks and techniques supporting Smart Clients have matured considerably since the first concepts of Smart Client emerged in 1997 (Yoshikawa, Chun et al. 1997).

The project, with a working title of "Inter-Store", is designed to capitalise on Smart Client design approaches to achieve a blend of local and networked operation in environments where technical skills are not necessarily common. There is clear applicability of the design concepts in this proof-of-concept model to the deployment of library services applications in developing countries, where similar resource and connectivity constraints apply. The following considerations framed the application architecture:

- To provide a single install process, and involve minimal or no configuration.
- The application must install and deploy all components within its own framework (that is, it must not rely on other external components such as Apache).
- The user interface must be unambiguous: with the least amount of pages possible and with a common structure across all pages.
- The system must be able to run on an average specification PC.
- The user interface must display on multiple monitor sizes

Key functional elements of the local element of the application are:

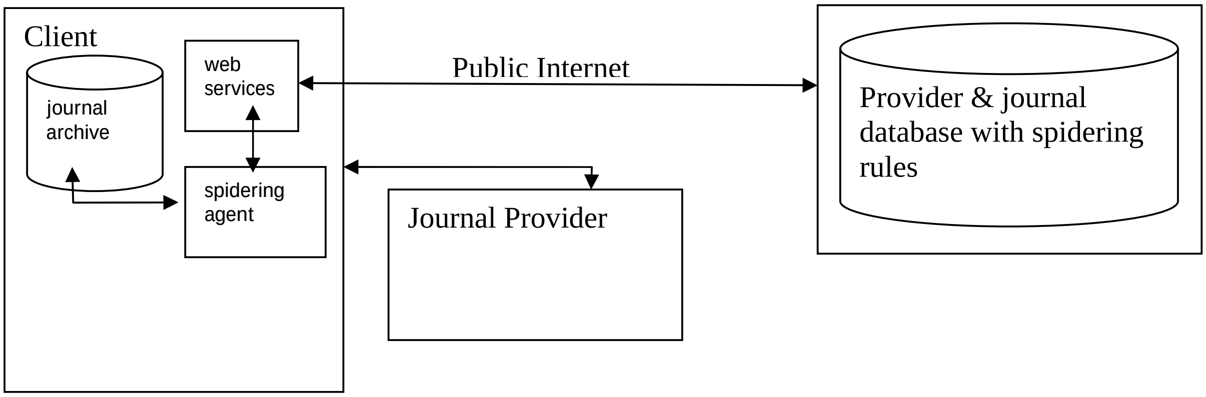
- The ability draw application and metadata updates from web services provider
- The system must be a configurable archive that builds a store of the journals that exist in the current electronic subscriptions. The system must also provide a means for viewing the archived content.

The Smart Client application has been developed in Java to allow portability across different platforms. It makes use of two crucial open source tools:

- The Jetty embedded web server component that allows the Smart Client to provide a browser based interface for the client application
- The Derby embedded database developed by the Apache foundation.

Designed properly, much of the operation an archiving engine can be autonomous, requiring only periodic user intervention. Access to an archive can also be designed to require minimal technical end-user skills, through search engine-style interfaces. However, the metadata information relating to journal collections, suppliers, crawling rules and locations, and also those relating to journal licensing and IP restrictions, is extensive and complex. Each journal provider typically has different methods of presenting their journal holdings, different methods for user authentication and a range of different ways for the presenting their downloads. It would be very difficult for a single special library to collect and design or specify this wide range of download requirements. In developing Smart Client application we have therefore had two goals: firstly, making use of existing open source tools for building the client application, and secondly to ensure that information about journal providers is collected through general purpose metadata that can be used by all of the clients.

Diagram 1: InterStore Client & Metadata Server



The client maintains an independent archive, stored locally on the computer that it is currently running on. The client acts as an autonomous agent, crawling the journal sites based on the user configured subscriptions, gathering the full text journal for local storage. The client implements a generic spidering pattern that has been developed to support multiple journal provider website structures through the use of provider specific spidering rules, which are not built into the client, but rather accessed from the metadata server through the use of web services. Not only does the metadata server provide the spidering rules, it also provides an index of the journals offered by each supported provider.

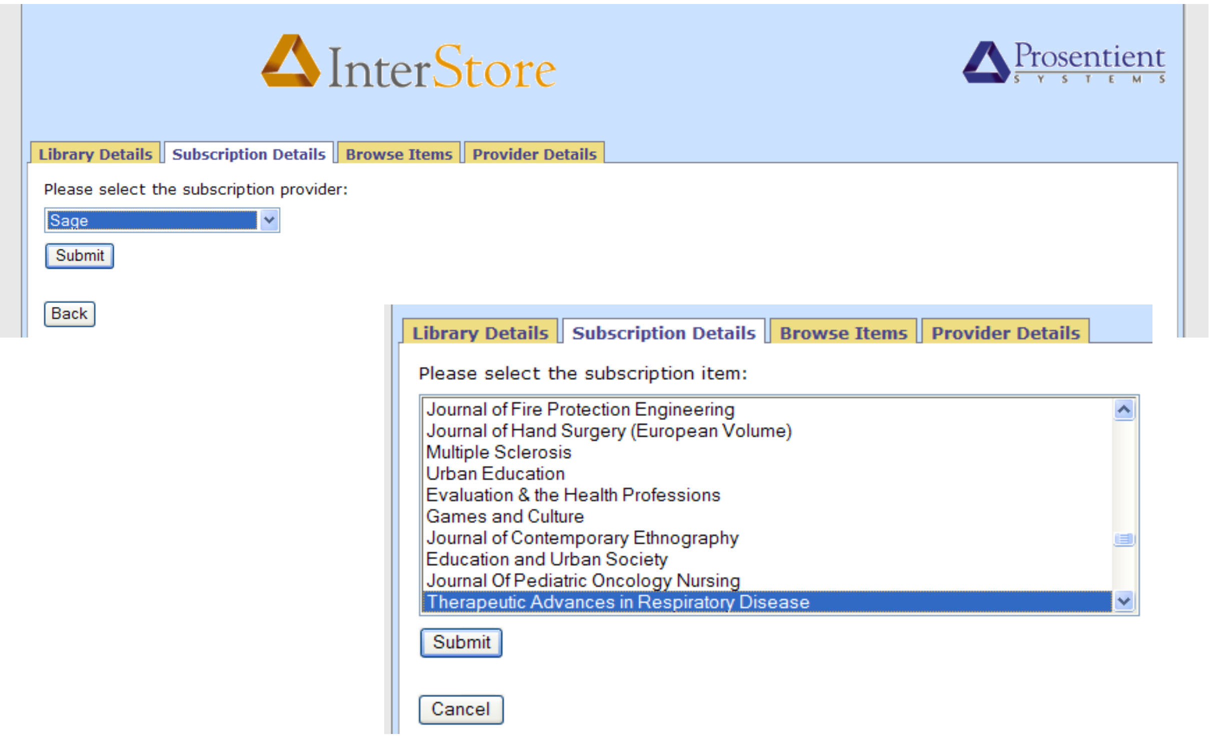
The novelty of the metadata server lies in its ability to separate the task of archiving from journal indexing. This system does not aim to build a local index of all the journals that are offered by a provider, but rather to build an accessible local archive. The extensive metadata associated with journals, journal providers and web crawling rules would be expensive, laborious and difficult to repeat in for a special library. The metadata server is a simple, reproducible web services provider that encapsulates these rule sets and associated journal and provider metadata. Such a metadata server could be hosted in many locations nationally.

An obvious strength of this design is the self-maintaining properties of both the client and the metadata server. Changes made to spidering rules can be propagated through the web service metadata server. The local client can constrain its crawling activity to the minimum number of pages required to keep its local archive current within the bounds of the particular institution's subscription.

Web Crawling

To create a local archive of documents, the InterStore system uses website crawling techniques based on crawling rules supplied from a metadata server. The spidering agent of InterStore is designed to crawl journal provider sites as infrequently as possible to reduce the load placed on both the client and the host. Use of a journal metadata repository considerably reduces the need to search & determine website structures to be crawled. The system makes no use of threading, to reduce the impact of the crawling process, and will look only for those issues that are within the defined library subscription period. Web crawling is performed in two phases: initial subscription, and periodic updates.

The InterStore client transmits a request to the metadata server for the list of available journals for the selected provider. The metadata server responds by returning a list of journals and their corresponding id and crawling/discovery rules. The journal id is used to build a URL string corresponding to that journal.


InterStore
Prosentient SYSTEMS
Library Details
Subscription Details
Browse Items
Provider Details
Submit
Back
Library Details
Subscription Details
Browse Items
Provider Details
Journal of Fire Protection Engineering
Journal of Hand Surgery (European Volume)
Multiple Sclerosis
Urban Education
Evaluation & the Health Professions
Games and Culture
Journal of Contemporary Ethnography
Education and Urban Society
Journal Of Pediatric Oncology Nursing
Therapeutic Advances in Respiratory Disease
Submit
Cancel

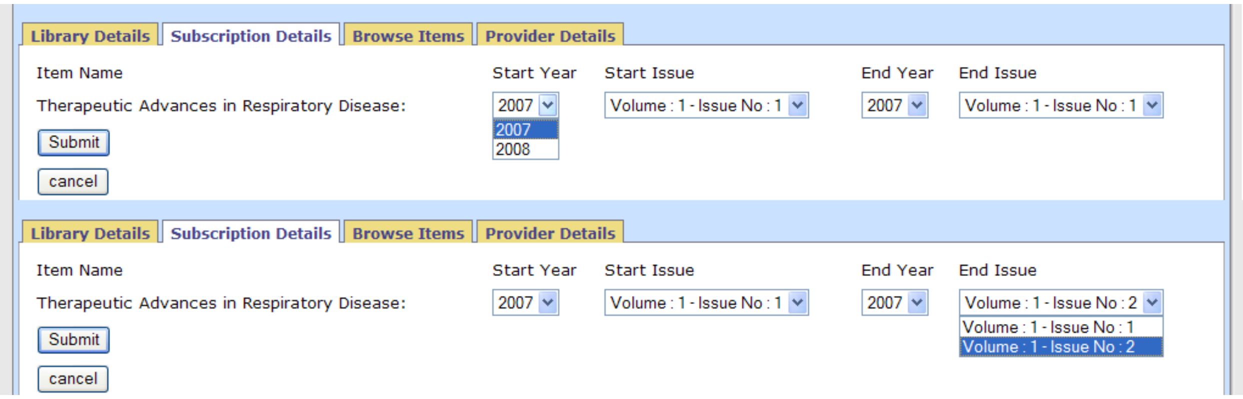
Please select the subscription provider: Sage

Please select the subscription item: Therapeutic Advances in Respiratory Disease

The client derives a Table of Contents using the browsing metadata – for instance a “Year” table of contents page contains a link to the issues table of contents for each year that the journal was published. The client can spider this page to dynamically build form select menu’s containing the available years by extracting the relevant links. The same process is followed to create the form select menu for the available issues

Next a structured list of issues is presented, again derived dynamically from the website based on crawling metadata:

The same process is followed for setting the end subscription year and issue, except the web crawler dynamically validates the subscription. To ensure that a valid subscription is entered, the Client intelligently constructs the form select menus to only allow the entry of valid values. For instance, it does not allow the subscription end year to be before the start year. The images below illustrate the dynamic subscription validation, which only allows the selection of years and issues that will not compromise the validity of the subscription format.



Item Name	Start Year	Start Issue	End Year	End Issue
Therapeutic Advances in Respiratory Disease:	2007	Volume : 1 - Issue No : 1	2007	Volume : 1 - Issue No : 1

Item Name	Start Year	Start Issue	End Year	End Issue
Therapeutic Advances in Respiratory Disease:	2007	Volume : 1 - Issue No : 1	2007	Volume : 1 - Issue No : 2

Intelligent spidering of the table of contents pages has allowed the Client to dynamically build all the forms used during the subscription configuration process. All values were retrieved at the time of configuration, hence ensuring that all journal information is current while abolishing the need to constantly update an index of provider-specific journal holdings.

The web crawler can implement the same spidering pattern for multiple journal providers simply by inserting the provider specific spidering rules, which are dynamically retrieved from the metadata server. This level of abstraction allows for the re-use of code, as well as easing the task of updating and developing additional spidering modules. The electronic document is stored in the nominated directory in the local file system. It is also stored in WARC format for archival management and optional interfacing using other Digital Archive management tools

Currently, updates to the archive are user initiated. Scheduling would be possible through relevant operating system scheduling features (for example cron in Linux and the windows scheduler).

Web Services

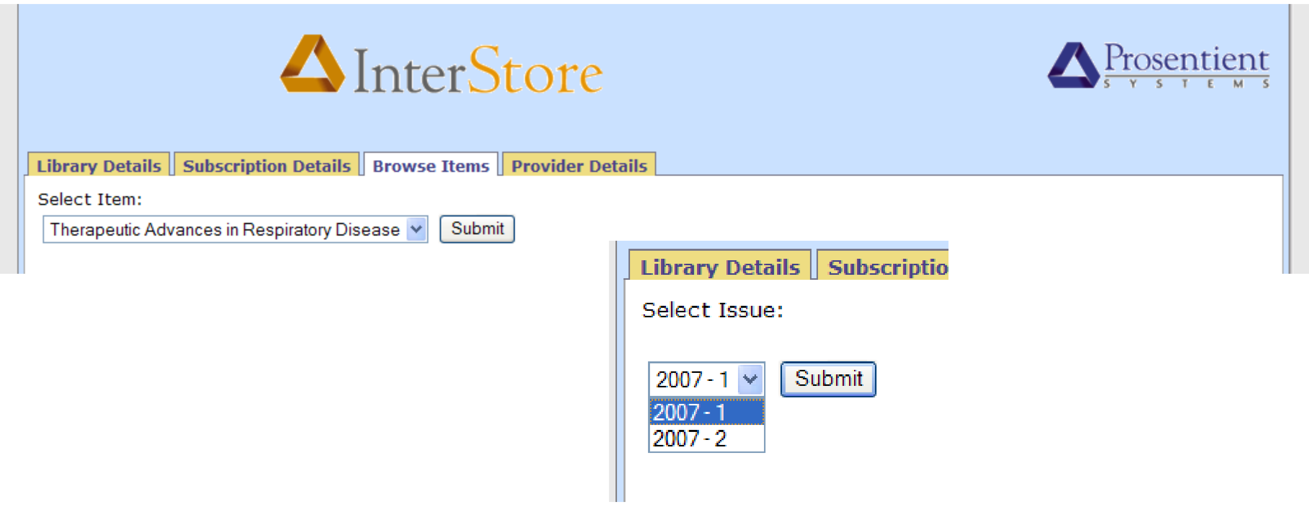
The InterStore Client uses web services to communicate with the metadata server. A web service is a software system that allows the communication between servers and clients through the use of XML files and the SOAP standard. The metadata server acts as the web services server and provides data to the InterStore Client which acts as the web services client. Web services utilize XML, as XML currently provides the most universal means for modelling and exchanging data. Web services are an excellent tool to accommodate the ever changing, flexible IT infrastructure and hence, the use of web services provides a means of easily creating a distributed application.

In terms of the InterStore system architecture, web services allow for the creation of the distributed network of InterStore Clients and metadata servers, with web services providing client to metadata server communication.

Browsing the Archive

The InterStore client also acts as a portal to the archived journals. The client allows local users to browse their archive and view the desired article locally, even if access to the internet is unavailable. While the client may rely on the metadata server for spidering rules, it does not require the metadata server to browse and view the local archive. The embedded web server provides the browsing method for access to the local archive, independent of the metadata server.

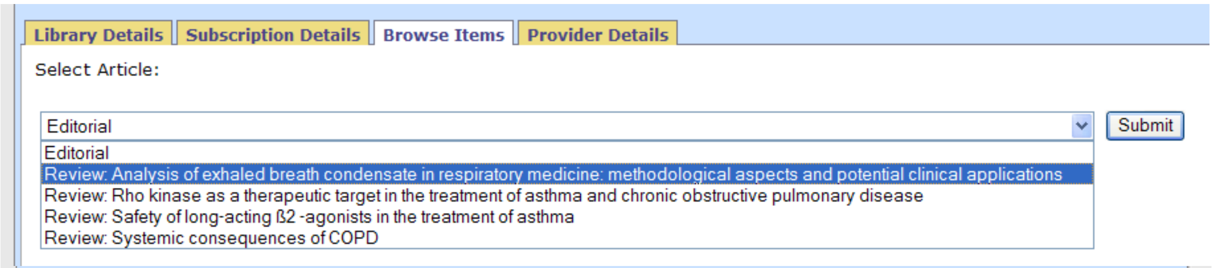
This proof-of-concept implementation provides a very simple browsing interface, and access is limited to the local machine as a means of limiting the copyright implications. The embedded web server could very easily be extended to provide public browsing access to the archive, or proxy based access. Other archive searching agents could also be deployed.



Select Item: Therapeutic Advances in Respiratory Disease

Select Issue: 2007 - 1

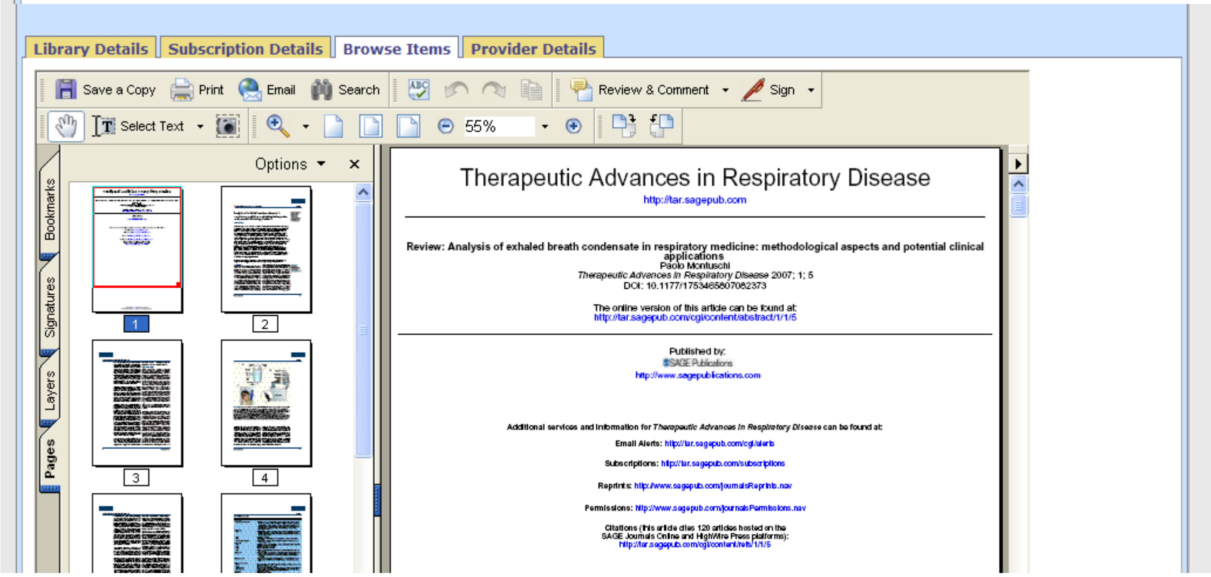
2007 - 1
2007 - 2



Select Article:

Editorial

Editorial
Review: Analysis of exhaled breath condensate in respiratory medicine: methodological aspects and potential clinical applications
Review: Rho kinase as a therapeutic target in the treatment of asthma and chronic obstructive pulmonary disease
Review: Safety of long-acting ß2 -agonists in the treatment of asthma
Review: Systemic consequences of COPD



Therapeutic Advances in Respiratory Disease
http://tar.sagepub.com

Review: Analysis of exhaled breath condensate in respiratory medicine: methodological aspects and potential clinical applications
Paolo Montuschi
Therapeutic Advances in Respiratory Disease 2007; 1; 5
DOI: 10.1177/1753465807082373

The online version of this article can be found at:
http://tar.sagepub.com/cgi/content/abstract/1/1/5

Published by:
SAGE Publications
http://www.sagepublications.com

Additional services and information for *Therapeutic Advances in Respiratory Disease* can be found at:

Email Alerts: http://tar.sagepub.com/cgi/alerts

Subscriptions: http://tar.sagepub.com/subscriptions

Reprints: http://www.sagepub.com/journalsReprints.nav

Permissions: http://www.sagepub.com/journalsPermissions.nav

Citations (this article cites 120 articles hosted on the
SAGE Journals Online and HighWire Press platforms):
http://tar.sagepub.com/cgi/content/refs/1/1/5

Archival formats

The focus of this project was the archiving of PDF e-journal content. E-journal content is almost universally in one of three formats:
* text only
* Portable Data format image
* HTML full text

WARC (Web ARChive file format) provides a means of consolidating digital objects and metadata in a single compressed file format (Library of Congress 2009). There is a Java-based open souce toolkit for this format developed by the *The Laboratory for Web Algorithmics* (LAW)

E-journal metadata is stored in the embedded database and is attached to each archived document using WARC archiving. The digital documents were stored "as is" in the archive with no further translation.

The design of the Inter-Store application attempts to ameliorate issues of obsolescence in two ways: storage of the archival documents in the file system rather than database; and use of a portable java application design. By storing the digital objects "natively" in PDF format, and also archived in WARC format, the project hopes to optimise the future accessibility of the digital archive objects to other applications as required.

Analysis

While it is early days in the life cycle of this new application, results to date have been positive. Installation is very easy, with the only pre-requisite being the Java runtime engine. Two small libraries form the core of our test community. Profiles for fifteen large & small e-journal providers have so far been defined and tested in the metadata repository. These libraries have fewer than ten e-journal subscriptions - but even this small number is spread across five providers.

The web service has successfully operated through proxy firewalls and the quite low bandwidth settings. Local archiving is functioning successfully. To emphasise the archival nature we have not integrated a search interface, only a local Table of Contents-based view of archival material.

Since the solution is developed as a web application, it can be deployed on any computer with a web browser but does not require permanent web connectivity. Since the web server and database are both embedded, they do not require any user intervention or configuration.

To treat continuity of internet access as a certainty belies risks to continuity of access over the long term. The Distribution Denial of Service (DDOS) attack which was launched against the Estonian internet infrastructure illustrates the ways in which internet infrastructure can be impeded at the national level (Halpin 2007). Small libraries and those libraries in developing countries may face other institutional and cost impediments to continuity of internet access or access to trans-national archival resources.

There are also limitations to the scope of local archiving. In some cases subscriptions are being replaced with license-based models for access, where a wide collection of journals is accessed on a pay-per-view or annual fee model, but with no associated subscription rights. Such a model does not lend itself to any form of local archiving.

One of the challenges to digital archiving is the pace of change in intellectual property and digital rights management. Some publisher licensing models will simply prevent local archiving of digital content. Digital rights management systems have the potential to block archiving of content or to limit its scope. Publishers will need ongoing encouragement to engage positively in approaches to archiving in the e-journal era. Web-2.0 based web services standards for e-journal discovery & download would make this task easier.

This research offers the example of a smart client approach to local archiving. Innovation in archiving methods is essential in order to work with an increasingly electronic research environment. The pace of transition to e-journal collection building is progressing in advance of the mechanics for e-journal archiving. Open source approaches encourage such innovation by providing a variety of models to systems issues that encourage incremental enhancement.

We anticipate that our approach to journal archiving could be applicable for both special libraries and libraries in developing countries that have the resources and poor Internet connectivity. At this point we are focusing on further trials in Australian special libraries while we build additional provider profiles for the metadata server. We would be delighted to partner with other institutions in developing this approach to Journal archiving.

References

Alford, D. (2002). "Negotiating and Analyzing Electronic License Agreements." Law Library Journal **94**(4): 621-644.

Balnaves, E. (2005). "Systematic Approaches to Long Term Digital Collection Management " Literary and Linguistic Computing **20**(4): 399-413.

Brown, J. (2007). "Researchers and librarians – Worlds apart? ." Chartered Institute of Library and Information Professionals Update, July/August 2007 Retrieved 2008-07-01, from http://www.cilip.org.uk/publications/updatemagazine/archive/archive2007/july/brown.htm.

Halpin, T. (2007). "Estonia accuses Russia of ' waging cyber war'." TimesOnline **May 17, 2007**(http://www.timesonline.co.uk/tol/news/world/europe/article1802959.ece last accessed 2007-06-01).

Kubilius, R., K. and L. J. Walton (2005). "Seize the E-Journal: Models for Archiving symposium: report." Journal of the Medical Libraries Association **93**(1): 126–129.

Lawrence, S., D. M. Pennock, et al. (2001). "Persistence of Web References In Scientific Research." Computer **34**(2): 26-31.

Library of Congress. (2009). "Sustainability of Digital Formats Planning for Library of Congress Collections: WARC, Web ARChive file format." Retrieved 1-05-2008, 2008, from http://www.digitalpreservation.gov/formats/fdd/fdd000236.shtml.

Machovec, G. S. (2006). "E-journal Archives and Preservation An Executive Overview." The Charleston ADVISOR **7**(4): 51-51.

Maniatis, P., M. e. Roussopoulos, et al. (2005). "The LOCKSS peer-to-peer digital preservation system, ." ACM Transactions on Computer Systems **23**(1): 2-50.

Oltmans, E. and H. van Wijngaarden (2004). "Digital preservation in practice: the -Depot at the Koninklijke Bibliotheek." VINE **34**(1): 21 - 26.

PORTICO. (2007). "Portico’s Archival Approach." Retrieved 2007-06-01, from http://www.portico.org/about/approach.html.

Yale University Library (2002). YEA: The Yale Electronic Archive - One Year of Progress: Report on the Digital Preservation Planning Project. New Haven, CT, Yale.

Yoshikawa, C., B. Chun, et al. (1997). Using smart clients to build scalable services. Proceedings of the USENIX 1997 Technical Conference, Anaheim, California