NEWSPAPER DIGITAL LIBRARIES NEWS-CLIPPING SERVICES AND LONG TERM ARCHIVING USING GREENSTONE

M.G. Sreekumar

R. Biju

T. Sunitha

P. Sreejaya

Joshy Kuriakose

and

K.M. Sudheesh Kumar

ABSTRACT

The importance of newspapers for people of all walks of life, institutions, government and countries in general, need not be overemphasized. It is truly a reflection and narrative of a society, status and its culture. Researchers too turn to newspapers for casual as well as quality research. It is interesting to note the continued relevance of the newspapers in the traditional paper format even when the publishing and transmission of news as well as newspapers across the world have migrated to newer paradigms and platforms such as the Web, Blogs, PDAs, Mobiles, eReaders etc. At the same time, newspaper management in libraries pose plethora of problems to librarians. Requirement of vast space, large size racks, extraordinary weight of bound volumes coupled with the damage that the paper undergoes over time are just a few to mention them. Recent developments in technology, especially the digital technologies, have a lot to offer in efficiently combating the above problems. The current information environment prompts libraries to leverage on the latest digital technologies towards building newspaper digital libraries and in setting up dynamic newspaper access and retrieval systems. These newer breeds of information services offer lots of power and visibility to newspapers, especially the old collections. This paper highlights the power of Greenstone software in developing newspaper digital libraries. It also reports the success stories of two large scale newspaper digital libraries - the Papers Past newspaper digital library, a collection of 19th and 20th century newspapers from the National Library of New Zealand and the Singapore National Library. The experience of the Center for Development of Digital Libraries (CDDL) of IIM Kozhikode in setting up a digital library of newsclippings, the digital equivalent of the traditional press clipping service, using Greenstone software is also shared.

Key Words (newspapers, newspaper digital libraries, newsclipping service, digital library software, Greenstone)

NEWSPAPERS

Newspapers are considered to be the first draft of history, while at the same time, they are part of a country's cultural heritage [Gatos]. Whether it is casual reading, random search or serious research, one of the best places to look for quality information is newspapers. They are rich with local news, long hours of journalism and well researched analysis. For more than 400 years, they have carefully and faithfully chronicled every sphere of life – be it society, politics, education, business or sports. Its value cannot be judged by its enterprise nor its importance derived by its age. Older the newspaper, higher is its research value propositions. Even before computers were born, librarians have developed excellent tools and technologies for storing and retrieving information from the traditional print based collections. The comprehensive indexing and search mechanisms deployed by librarians to retrieve information have been quite accurate and quick. Newspapers are part of the prominent collections of most libraries and the newspaper section is probably the users' first port of land in a library.

Changing Service Models

Even though the manual efforts are commendable, it is simply not enough in an age which is monopolized by search engines and intelligent information systems. Over the years, users also have become very keen in getting instant access to information and they look for micro-second search results. Inefficiencies of manual access, risks of natural hazards and physical storage costs still remain a challenge for the traditional systems to deal with. As technology marches past at tremendous pace, the traditional meaning and definitions of a library's collection range also undergo a great deal of change. In the current information environment libraries need to leverage on the latest digital technologies towards building newspaper digital libraries. By converting newspaper archives to digital collections the dual goal of digital libraries in terms of preventing paper deterioration as well as providing online full-text access of the archives by all interested parties is achieved.

Digital Libraries Carved out of Open Source Software

Digital Libraries (DL) are now emerging as a crucial component of global information infrastructure, adopting the latest information and communication technology. Digital Libraries are networked collections of digital texts, documents, images, sounds, data, software, and many more that are the core of today's Internet and tomorrow's universally accessible digital repositories of all human knowledge. According to the Digital Library Federation (DLF, USA - http://www.dlf.org), "Digital libraries are organizations that provide the resources, including the specialized staff, to select, structure, offer intellectual access to, interpret, distribute, preserve the integrity of, and ensure the persistence over time of collections of digital works so that they are readily and economically available for use by a defined community or set of communities".

Digital libraries offer new levels of access to broader audiences of users and new opportunities for library and information science field to advance both

theory and practice. They contain information collections predominantly in digital or electronic form. Electronic publications have some special management requirements as compared to printed document. They include infrastructure, acceptability, access restrictions, readability, standardization, authentication, preservation, copyright, user interface etc.

Digital libraries do enable the seamless integration of the scholarly electronic information, help in creating and maintaining local digital content, and strengthen the mechanisms and capacity of the library's information systems and services. They increase the portability, efficiency of access, flexibility, availability and preservation of digital objects. Once the information is made digital, it could be stored, retrieved, shared, copied and transmitted across distances without having to invest any additional expenditure. Digital Libraries can help move the nation towards realizing the enormously powerful vision of 'anytime, anywhere' access to the best and the latest of human thought and culture, so that no classroom, individual or a society is isolated from knowledge resources. Digital library brings the library to the user, overcoming all geographical barriers.

World over there is increasing appreciation of the Open Access movement and the Open Source Software philosophies and for may a libraries it is a chosen decision, be it technical or financial reasons, not to go for a proprietary digital library software. One needs to evaluate some of the popular Open Source Software for digital libraries, which are in use internationally. 'Dienst', 'Eprints', 'Fedora', 'Greenstone' etc. are among the candidates for the preferred software. Obviously Greenstone outscores the group as a general purpose digital library software from the point of view of a multi-publication type, multi-format, multi-media and a multi-lingual practical digital library [Greenstone]. And once finalized, it could be formally adopted as the software for creating the digital library.

Greenstone Software Features

Greenstone is a suite of software for building and distributing digital library collections. It is not a digital library but a tool for building digital libraries. It provides a new way of organizing information and publishing it on the Internet in the form of a fully-searchable, metadata-driven digital library. It has been developed and distributed in cooperation with UNESCO and the Human Info NGO in Belgium. It is open-source, multilingual software, issued under the terms of the GNU General Public License. Its developers received the

2004 IFIP Namur award for "contributions to the awareness of social implications of information technology, and the need for an holistic approach in the use of information technology that takes account of social implications". Again in 2008, Greenstone was bestowed the prestigious Mellon award, which it richly deserved.

Greenstone software along with Java Run Time Environment (JRE), Ghostscript and ImageMagick are deployed for building digital libraries. The software suite is available at the open source directory 'Sourceforge' [Sourceforge].

Figure 1: IIMK Digital Library

The salient features of Greenstone are basically taken from two of the official publications of the software development team appeared in D-Lib Magazine during the year 2001 [Witten, 2001] and 2003 [Witten, 2003]. Greenstone builds collections using almost popular and standard digital formats such as HTML, XML, Word, Post Script, PDF, RTF, JPG, GIF, JPEG, MPEG etc. and many other formats which include audio as well as video. It is provided with effective full-text searching and metadata-based browsing facilities that are attractive and easy to use. Moreover, they are easily maintained and can be augmented and rebuilt entirely automatically. The system is extensible: software "plug-ins" accommodate different document and metadata types. Greenstone incorporates an interface that makes it easy for people to create their own library collections. Collections may be built and served locally from the user's own web server, or (given appropriate permissions) remotely on a shared digital library host.

End users can easily build new collections styled after existing ones from material on the Web or from their local files (or both), and collections can be updated and new ones brought on-line at any time. The Greenstone Librarian Interface (GLI) is a Java based GUI interface for easy collection building. Greenstone software runs on a wide variety of platforms such as Windows, Unix / Linux, Apple Mac etc. and provides full-text mirroring, indexing, searching, browsing and metadata extraction. It incorporates an interface that makes it easy for institutions to create their own library collections. Collections could be built and served locally from the user's own web server, or (given appropriate permissions) remotely on a shared digital library host. The other set of features include OAI plug-in (introduced since the 2.40 version) and DCMI compliance, UNICODE based multi-lingual capabilities and a user-friendly multimedia interfacing

[Unicode]. Further more, it is powered by robust indexing systems such as 'Managing Gigabyte' Plus-Plus ('MG' PP) and Lucene. A very interesting feature of Greenstone is its exhaustive set of well documented and articulated manuals (http://www.greenstone.org/ cgi-bin/library?e=p-en-docs-utfZz-8&a=p&p=docs) such as

'Installer's Guide', 'User's Guide', 'Developer's Guide', and 'From Paper to Collection' a document describing the entire process of creating a digital library collection from paper documents. This includes the scanning and OCR process and the use of the "Organizer". There is one more interesting documentation 'Inside Greenstone Collections' which clarifies most of the trickier parts of using Greenstone, especially dealing with configuration file for the collection in question. The recent introduction of the 'realistic book' feature of Greenstone is very exciting and shall enthrall the users. It provides collection builders the facility to serve the eBooks in the real-life print counterpart format (http://nzdl.org/Books). It is truly exciting with its quick, easy-to-use, and responsive page-turning mechanism, and combines the ability to include hyperlinks and animated media.

There are presently two versions of Greenstone going around, Version 2 and 3, and they are generally referred to as Greenstone2 and Greenstone3. The latest in Greenstone2, as on February 2010, is V.2.83 and that of Greenstone3 is V.03. Greenstone2 will be there for some more time, but ultimately Waikato/Greenstone see that Geenstone3 will replace it. IIMK is presently the UNESCO host for the Greenstone Support for South Asia. The web site http://greenstonesupport.iimk.ac.in provides a compendium of information on the software. The eList greenstonesupport@iimk.ac.in offers regular online support to around 200 professionals from the South Asia region.

There is a belief that Greenstone software is ideal only for small to medium size digital libraries. Stefan Boddie et. al., colleagues from DL Consulting and the University of Waikato, the founding members of the Greenstone project, have since clarified that Greenstone is now being used to produce large newspaper collections for the National Libraries of New Zealand and Singapore [Stefan].

'Papers Past' Newspaper Digital Library of New Zealand

Papers Past is a Greenstone-based newspaper digital library of the National Library of New Zealand. Papers Past contains more than one million pages of digitised New Zealand newspapers and periodicals. The collection covers the years 1839 to 1932 and includes 52 publications from all regions of New Zealand. When fully built, the scanned and OCRed collection would have around 20 GB of raw text, 2 billion words, with 60 million full-text searchable unique terms. All images will be fully-searchable. The collection would consist of over 6.5 million newspaper articles, each with its own metadata (much of it automatically generated); and the total volume of metadata is 50 GB. Before being built into a digital library collection the metadata is stored in XML format, which occupies around 600 GB, slightly less than 1 MB per newspaper page

[Stefan].

Figure 2: Papers Past Homepage

Papers Past involves a massive amount of metadata. Its specification demands that newspaper articles be viewable individually as well as in their original context on the page. *Papers Past* is presently available at http://paperspast.natlib.govt.nz. There are two main ways to find information in Papers Past. The searching mode let users enter a query term and retrieves articles that contain that term. The browsing feature let users look at all the newspapers, starting with a year, a region, or a newspaper title. All the newspaper titles on the site can be searched and browsed.

Newspaper Digital Libraries of Singapore National Library

The newspapers digital library at the National Library Singapore is again powered by a Greenstone-based commercial product named Veridian. The front-end user interface as well as the retrieval engine are developed by the National Library Board. The collection is a vast one, comprising around 200 Singapore and Malaya titles published since 1806. It is rich in content and is quite important one, as it reflects the political and social life of Singapore since the country's founding in 1819. English language titles, representing the official language of Singapore, as well as other prominent languages of the ethnic groups in the country such as Malay, Mandarin and Tamil are also there in the collection. The collection is hosted at the Lee Kong Chian Reference Library in the National Library Building. The project is a joint initiative of the National Library Board (NLB)'s National Library and the island-wide network of Public Libraries.

The service, named NewspaperSG, was released on the Web in March 2009. Both onsite and offsite users can search over 550,000 pages of the digitised Straits Times (1845 –

1989) and the microfilm holdings of the newspapers from a single website at http://newspapers.nl.sg. Users will also be able to obtain some brief information such as the article title, date of publication, and a 50-words extract from the search results. Full article views are, however, only available through the library's multimedia stations (PC stations). To aid the discovery

of archival news content from the Web, a "Table of Contents" is also generated for each newspaper issue and then submitted to major search engines (Google, Yahoo, MSN) for indexing.

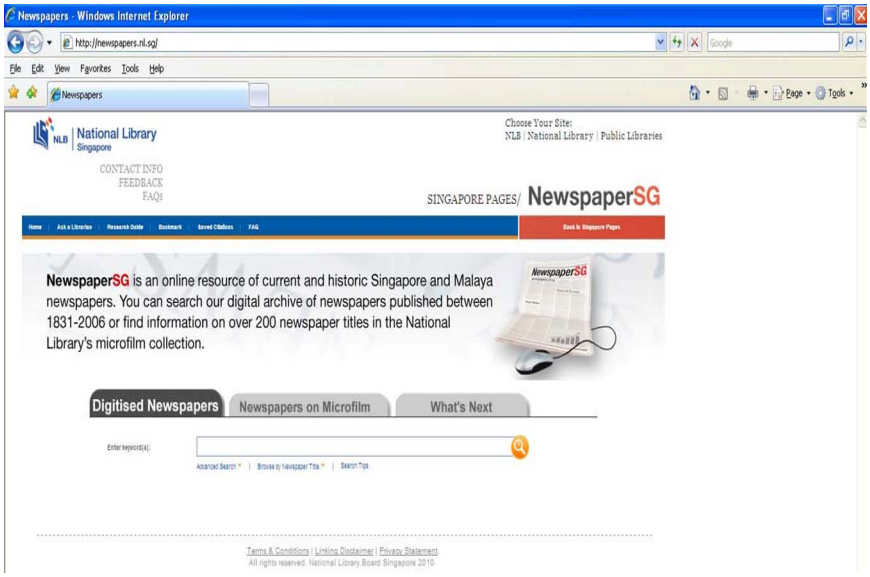

Figure 3: NewspaperSG Homepage

Technical Considerations

Newspapers are pretty complex objects to deal with and they indeed pose a host of challenges to the digital library management and the collection development team. Firstly, in most cases, they manifest in A2 size and that itself is a point of debate as to what should be the mode of presentation in terms of size, in the digital interface. There are multiple service models found in usage by different online newspaper libraries such as ePaper format bundled with proprietary eReaders, PDF/JPEG/TIFF files of different sizes and resolutions or in any other suitable image format. While dealing with quite old and backdated newspapers, the low quality of originals and the old font styles could give substantial labour. OCRing these documents could be even more a mammoth task.

Search tools should be carefully designed, keeping in mind the different category of users. Provision should be there for advanced searching using intelligent algorithms, exact word searching and Boolean queries, while the basic free text searches on simple as well as compound terms are well addressed. Vector space approach, extended Boolean models and weighting schemes are other possibilities for increased search and retrieval efficiency. For non-English and vernacular collections, NLP and its advanced models could be explored and implemented.

Metadata which are searchable by the available indexing systems for each and every article, forming part of the newspaper items in the collection, is the next important challenge. Systems like Greenstone has mechanisms to

generate automatically extracted metadata from most of the text oriented file formats. Bi-tonal digital images in TIFF format are the one recommended for scanned versions and then they are later OCRed for indexing purpose. While Dublin Core, ONIX, METS and ALTO (XML Schema for technical metadata used with OCR scanning output) are the recommended metadata standards, XML is the recommended encoding or representation markup format.

Newsclipping Service of IIM Kozhikode

Inspired by the seamless features of Greenstone software in terms of interface presentation, search and retrieval, customization and configurations, we embarked on creating a digital library of newsclippings. Traditionally, pressclipping services have long been well known artifacts offered by the library fraternity. Named as 'IIMK in Media', the collection aims at containing all news items on IIMK that have appeared in newspapers across the world. A reverse chronology collection development approach has been adopted presently, while parallel efforts will be tried for capturing the news archives available in print format.

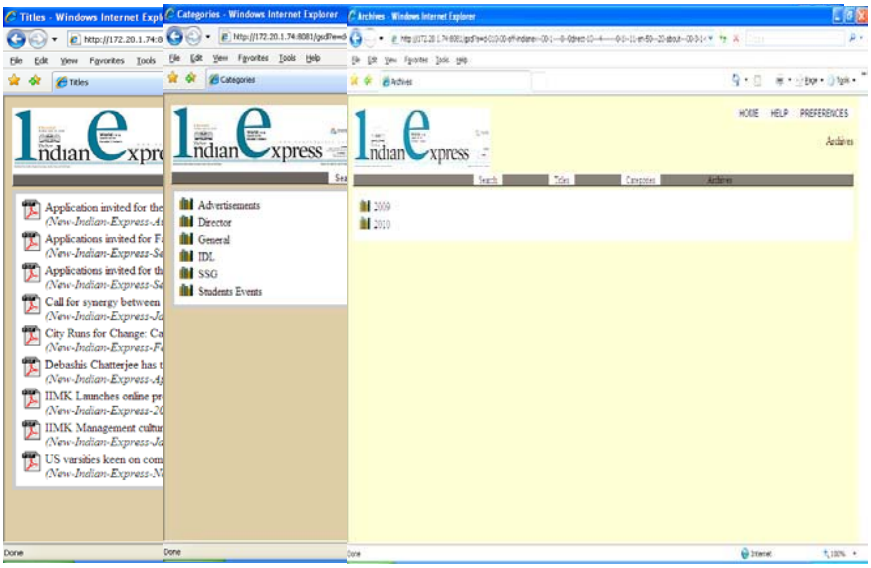

Figure 4: News-clipping Service Homepage

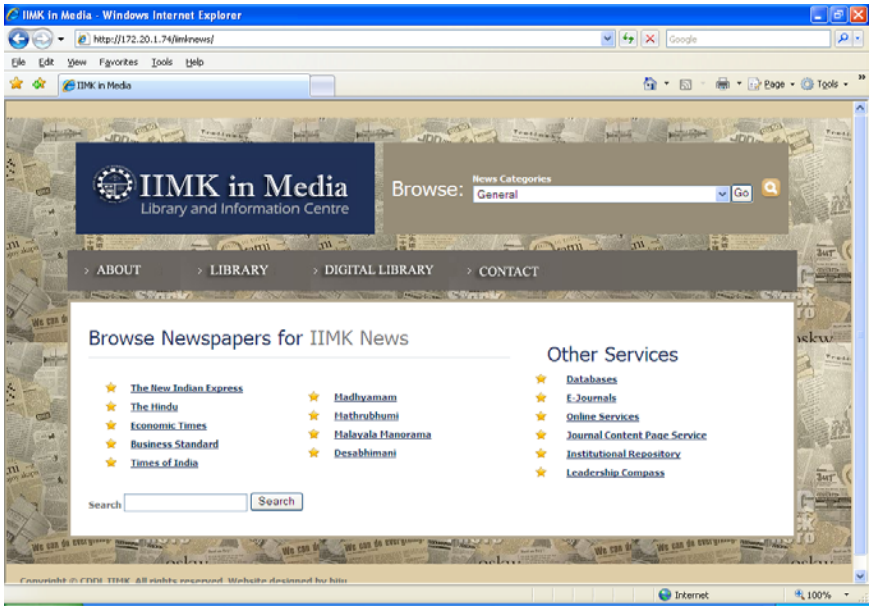

Figure 5: Browsing Classifier Screens : Title, Categories and Archives

The information model chosen for this service at the front-end interface include, in addition to the generic Browse and Search, a pull-down menu of News Categories, and the Newspaper Title-based collections. The collection level (within each newspaper title) browsing classifiers include Titles (the news headings), Categories (Advertisements, General, Programmes, Students, Faculty etc.) and Archives (year-wise). Hierarchy metadata descriptions, using Dublin Core, facilitate browsing on multi-layer subjects, categories as well as years and dates. The search interface could be configured for single collection, multiple collections or searches at the global level.

CONCLUSION

Newspapers play commendable roles in the socio-cultural-political fabric of a society, and preserving them as part of our heritage is of paramount importance. Maintaining them in the traditional format and methods are not viable any more. Digital libraries offer enormous opportunities, options, convenience, flexibility and improved efficiency to the users. Availability of robust software in the open source domain provides an unprecedented opportunity to libraries across the world, especially in catching up with the latest trends and technology. Greenstone provides a vast amount of freedom to the users to leverage on, and being a componentization model, it has the advantage of experimenting with the emerging and cutting edge applications and getting them incorporated into it. The model newspaper digital libraries showcased in this paper are just two examples and one can see more such examples in the Greenstone example collections as well as elsewhere. Also, the newsclipping service mentioned is simply one among such many applications one could think of configuring using Greenstone. The message

that we wish to underscore in this paper is that there are a whole lot of powerful open source software solutions that are built on latest software architectural standards and technologies, and the information science fraternity may find, evaluate, choose and use them as appropriate.

ACKNOWLEDGEMENT

The authors would like to acknowledge the unstinted support extended by Stefan Boddie of DL Consulting, New Zealand, and the development team of the Papers Past the NLB collections.

REFERENCES

1. Gatos, B. et. al.
2. An integrated system for creating a Digital Library from Newspaper Archives http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.61.282&rep=rep1&type=pdf
3. 2. http://www.dlf.org
4. IIMK Digital Library < http://iimk.ac.in/gsdl/cgi-bin/library >
5. 4. htttp://www.greenstone.org
6. SourceForge.net (world's largest Open Source software development website)
7. <http://www.sourceforge.net/>
8. Witten, Ian H. et al. 2001
9. Greenstone : Open-Source Digital Library Software
10. *D-Lib Magazine*, 7 (10): 1-16.
11. 7. Witten, Ian H. 2003
12. Examples of Practical Digital Libraries : Collections Built Internationally Using
13. Greenstone
14. *D-Lib Magazine* 9 (3): 1-15.
15. 8. Stefan, B et. al.
16. Coping with very large digital collections using Greenstone http://www.google.co.in/url?q=http://www.dlconsulting.com/marketing/vldl_boddie_et_al_final.pdf
17. 9. http://paperspast.natlib.govt.nz
18. 10. http://newspapers.nl.sg