Date submitted: 21/06/2010

Exposing Canada's historical ethnic newspapers through the Multicultural Canada Project

Lynn Copeland
Librarian and Dean of Library Services
Simon Fraser University
Burnaby, BC, Canada

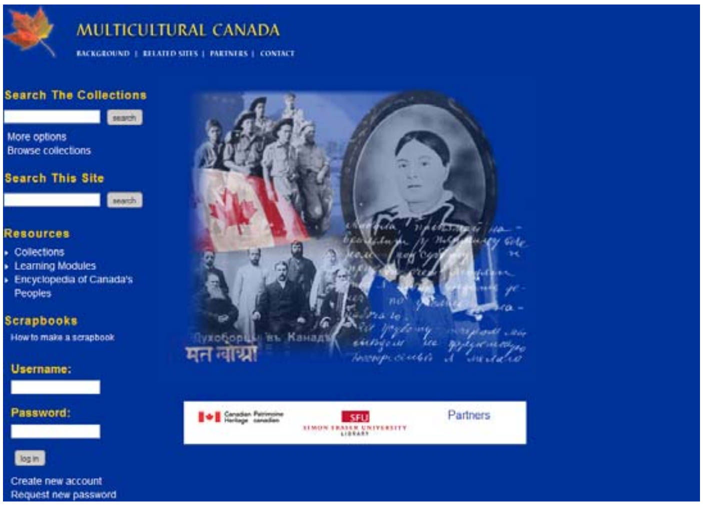
Meeting: **102. Newspapers**

WORLD LIBRARY AND INFORMATION CONGRESS: 76TH IFLA GENERAL CONFERENCE AND ASSEMBLY
10-15 August 2010, Gothenburg, Sweden
http://www.ifla.org/en/ifla76

Abstract:

The purpose of Multicultural Canada (http://multiculturalcanada.ca) is to digitize and provide free access to the heritage of Canada's immigrant peoples. 1.5 million items are now available.

Much of the communication within those new Canadian groups was through the newspapers which Simon Fraser University Library has digitized, some from deteriorating paper copy. A particular challenge was the OCRing and loading of nonroman characters. SFU Library worked closely with OCLC/ContentDM on this. Learning modules and a scrapbooking feature have been developed. Wikipedia links have helped expose the content. SFU Library is using its expertise to host and provide access to other newspaper collections and is contemplating further projects. This paper addresses our achievements and technical challenges.



The purpose of Multicultural Canada (http://multiculturalcanada.ca) is to digitize and provide free online access to the heritage of Canada's immigrant peoples to those communities, scholars, and the world. More than 1.5 million items have been digitized to date.

Significant funding has been provided by the Canadian government through the Canadian Heritage partnerships program. Our partners in creating this wonderful resource have included the Sien Lok Society of Calgary, the Multicultural History Society of Ontario, Jewish Museum and Archives of BC, Universities of British Columbia, Calgary, Toronto, and Victoria, and Vancouver Public Library. It is a commonplace to argue the urgency of capturing and digitizing the stories of our elders because of the aging population, and I very much regret and want to acknowledge the important contributions of Dr. Hari Sharma and Dr. Edgar Wickberg who have passed on since the inception of this project. In 2006 and 2008 I spoke at the IFLA conference/preconference about the progress to date on this project and this talk will not repeat those more general presentations. Instead I will focus on the newspapers which form a significant part of this online collection. This paper was prepared with the extensive help of Melanie Hardbattle, Mark Jordan, Brian Owen, and Ian Song of Simon Fraser University Library.

Nevertheless I would like to say a few words about Canada and about Multicultural Canada to set the context for our newspaper digitization activities. To the present (April 2010), the Canadian Government, though the Canadian Heritage Partnership Programme, and our various partners have invested over $1.2million Cdn in cash and in-kind contributions and produced over 1.4 million items. There are fifty-four collections, including thirty-three newspapers. The newspapers are in fifteen languages, with more than one in each in French (7), English (6), Chinese – traditional or otherwise (5), Urdu (4) and Japanese (2). There are also a number of oral histories, photographs, books and even digital objects and videos in our first partnership with a museum (the SFU Museum of Archaeology and Ethnography). A number of the collections such as the Doukhobor collections of SFU and UT are archival in nature and contain multiple formats. Some sixteen learning modules have been created and the new Multicultural Canada web site includes a scrapbooking feature. This allows individual users to save pages from the website or upload their own images into an electronic 'scrapbook', where they can add comments and create their own arrangement. Scrapbooks can published for all users to view, or kept private for one's own reference. Links from Wikipedia have been particularly successful in exposing these resources. The SFU Library is also home to some hundred digital academic journals, and numerous digital collections including newspapers from the city of Prince George, digitized by libraries in that city.

Most of the Multicultural Canada newspapers had been microfilmed by the MHSO and it was this version which was digitized in most cases.

Much of the communication within those populations took the form of newspapers, for example with the Chinese Times in Vancouver having a life span of seventy eight years. As lead institution on this project, Simon Fraser University Library has undertaken to digitize and present many of these newspapers, some of which were only in deteriorating paper form. A particular challenge has been the OCRing and loading of nonroman characters and SFU Library has worked with OCLC/ContentDM to enable this capability. SFU Library has built on this expertise and is housing other newspaper collections for the Canadian memory community. In this paper, the projects achievements, technical challenges and lessons learned will be addressed.

"If we had only known" might be the motto of the newspaper digitization part of the Multicultural Canada project. Fortunately we didn't or we might never have undertaken the project, because the challenges in this project were indeed serious.

First was the quality of the microfilming. When we received the microfilm, we discovered that several titles were unreadable by any standards and they could not be included at all in the project. In other cases there were missing issues and pages, irregularly arranged sequences and the like. So the filming was a problem. In some instances it appeared the originals were problematic. Nevertheless, where possible, the digitization took place, although in a couple of cases the OCR was too slow to complete in the relatively short timeline of the project (typically CCOP project funding is announced well into the fiscal year, although the projects must be completed within that timeframe). In these and other situations I will mention we will continue the work as time permits.

The Chinese Times in particular presented significant challenges, comprising about 350,000 pages. About thirty years of the paper had never been microfilmed and was stored in non-climate-controlled space. This part of the collection was kindly loaned to SFU by UBC Library (in return for a microfilm copy which was produced as part of the process) and digitized at SFU using our Zeutschel large-format overhead scanner. The microfilm was digitized in part by SFU Library and in part by a local company, Microcom. The Library converted the TIFF master images to JPEG for online display purposes. OCR (Optical character recognition) was undertaken by SFU Library, and with some difficulty we identified three candidates before settling on Abbyy Fine Reader. The other OCR packages were tried on the Chinese and Japanese newspapers and gave poor results, with TH-OCR Pro9 requiring non-UTF8 text be converted to UTF-8, and Readiris which could not handle large batch jobs. This is also a good example of the type of problem that sites need to solve in a short timeframe, and this was where the really interesting activity began! OCLC had agreed that we should be a beta site for their Unicode version of CONTENTdm 5 [1]; 1.5 years later, we can confidently declare that this is an ongoing relationship! Our systems analyst Kurt Bolko and Mark Jordan, Manager of Library Systems, are working closely with Craig Yamashita, Lead OCLC ContentDm developer, and SFU Library continues to have a production environment on the development server supplied by OCLC. To quote Nigel Long, OCLC, "Craig and the team really appreciate the collaborative efforts with SFU in dealing with the linguistic issues developing the software's full UNICODE capabilities and in realizing the kind of data support required for such collections." [2] The Multicultural Canada interface to CONTENTdm was built using Drupal and CONTENTdm's API capabilities. Figure 1 shows the workflow for the project.

The Chinese Times encompasses 2.4GB of storage; far more was needed during the processing stages. More dramatically, the index in ContentDM, which had initially been configured at 2GB is now a massive 8GB, owing to the extremely large number of entries (around 120k). One particularly useful feature for those of us unfamiliar with traditional (or modern for that matter) Chinese is the English word index that has been created for the period to 1950 using the translations which were created by Dr. Edgar Wickberg's (UBC) students during the writing of 'From China to Canada' [1].

Those who have received digitization funding with limited dollars, an extremely tight timeframe, and stringent technological restrictions (to ensure accessibility standards are met) will recognize the challenges faced during this project. This project was significantly more complex because of the wide variety of languages and character sets involved. Our focus during the project was to ensure that the content was created, indexed and loaded, and, in the case of our partner institutions, delivered in a timely manner. There were significant technical challenges but in the end we delivered. We focussed on the long term issues: quality of digitization and planning for preservation of content. Nevertheless we are taking the time 'between projects' to undertake more rigorous testing and to address issues relating to the searching and presentation of the content, improving the quality of the website and providing additional searching help for a fairly complex collection of materials. We also added one important feature from the user perspective, namely the scrapbooking feature.

To give one example of the kind of improvement we are making over the next few months, the default search is for 'all the words' but in Chinese, with pairs or multiples of characters forming a meaningful equivalent to an English word, the 'phrase' search is usually best. We have added text and search features to make this more functional and clear. We have received useful feedback which helps toward the improvement of the site. By the time of the IFLA conference, August 2010, these changes will have been implemented and will be demonstrated during the presentation. As we complete our work as partner to Athabasca University in their 'Connecting Canadians' project and explore two further multicultural digitization projects, one with University of British Columbia, we expect to expand the content of Multicultural Canada and build on the knowledge and expertise we have acquired.

These projects are exciting and important to the understanding of Canada's highly multi-ethnic society and of value from that perspective. They have also expanded our technological capabilities and enabled us to continue to undertake increasingly important and complex projects and we look forward to further challenges. In the end we accomplished more than promised, digitizing almost twice as much content as proposed and overcoming significant technological problems to produce an outstanding digital resource.

References

1. Duke, D. M. (2008, October 2008). CONTENTdm opens global archives with addition of unicode. *NeXTspace*, , 1. Retrieved from http://www.oclc.org/ca/en/nextspace/010/productsandservices.htm

2. Private email.

3. Con, H., & Wickberg, E. (1982). *From China to Canada : A history of the Chinese communities in Canada / Harry Con ... [et al.] ; edited by Edgar Wickberg.* --. Toronto, Ont.: McClelland and Stewart in association with the Multiculturalism Directorate, Dept. of the Secretary of State and the Canadian Govt. Pub. Centre, Supply and Services Canada. Available at http://www.multiculturalcanada.ca/node/4732 .

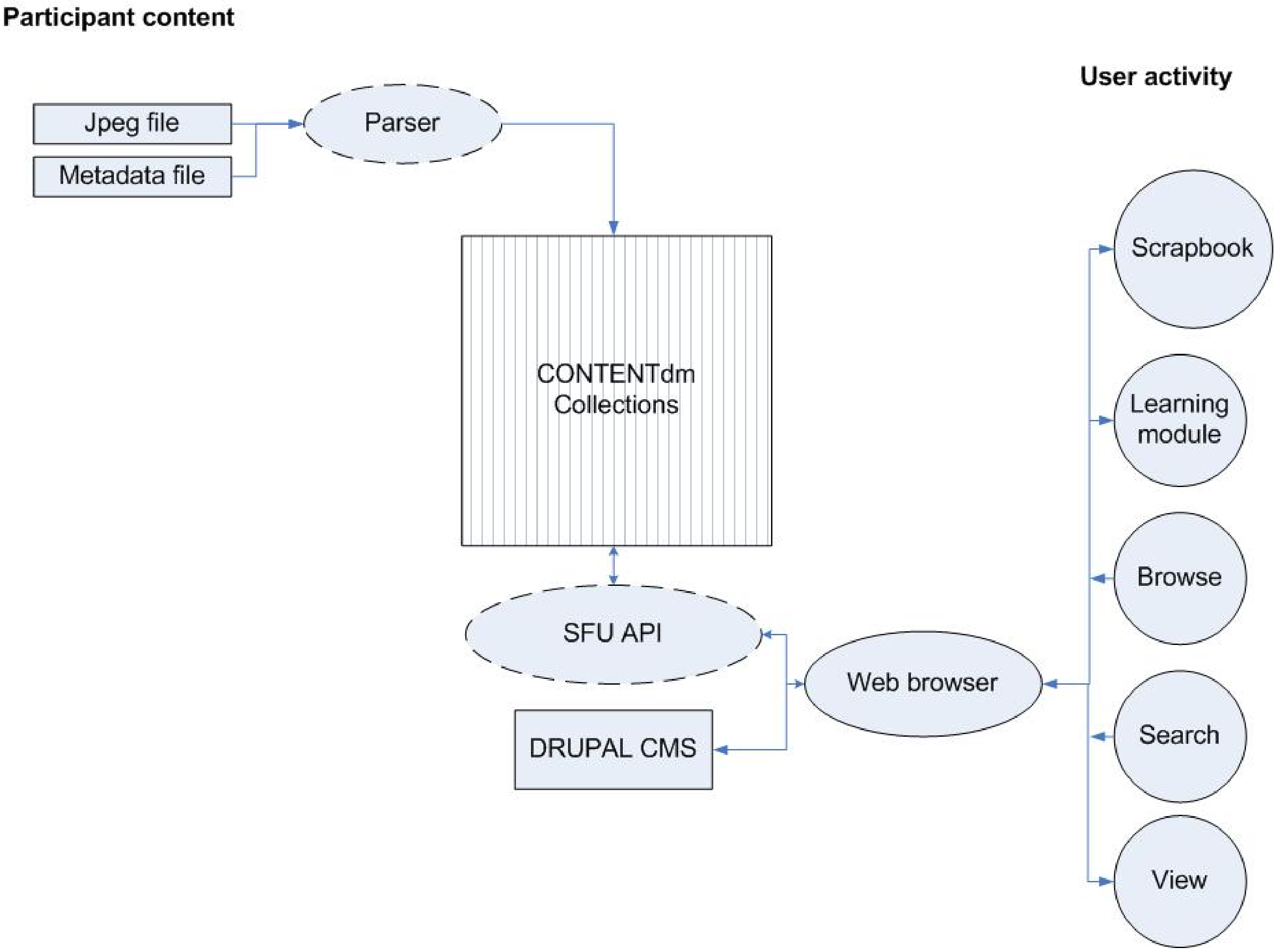


Figure 1. Multicultural Canada Schematic.

***

Short biographical statement :

Lynn Copeland is Dean of Library Services at Simon Fraser University. Lynn has lectured at the UBC School of Library, Archival and Information Studies, where she was a visiting scholar 2006, and engaged in a variety of consultations relating to library technology and collection development. She is the Chair of Canadiana.org, the national organization committed to preserving Canadian heritage in digital form as well as Chair of the Steering Committee of the BC node for the Synergies CFI-funded project and member of the David Lam Centre Steering Committee. She is the lead for the Multicultural Canada digitization project.

Lynn has published a variety of articles and spoken on issues primarily relating to academic libraries, library technology, collection development, multiculturalism and libraries, elearning and libraries, and Interlibrary Loan. She was Chair of the BC Library Foundation and has served on the Boards of the Canadian Association of Research Libraries, BC Library Association, Information Services Vancouver, Vancouver Folk Music Festival, and Chinese Canadian Historical Society of BC.