CANADIAN NEWSPAPERS ONLINE: A NATIONAL CONSULTATION

Toward a national strategy to strengthen access to Canadian newspapers online

Library and Archives of Canada, Ottawa
October 7-8, 2002

REPORT

The National Library of Canada, in association with the Canadian Initiative on Digital Libraries (CIDL), the Canadian Newspaper Association (CNA), and the Association for Canadian Studies (ACS) hosted a national consultation on Canadian online newspapers in October 2002. Only days before the consultation, on October 2, the Minister of Canadian Heritage had announced that the National Library and National Archives of Canada were joining together to create the Library and Archives of Canada.

The Library and Archives of Canada's own collections of newspapers in print, microform and on CD-ROM are among the largest of their kind in the country, but its ability to build and sustain a central national collection of online newspapers is less clear. The consultation aimed to explore cooperative strategies to strengthen, on a national basis, online access to contemporary and historical newspaper content for Canadians. The ultimate goal was to begin to articulate a national strategy to strengthen access and preservation of Canadian online newspapers.

Approximately 80 participants came from all sectors, including newspaper publishers, commercial information providers, researcher institutions, and libraries.

Summary and Outcomes

Summary

Newspaper users, including an academic researcher, a genealogist, and a librarian, outlined the wide range of casual and formal research uses that are made of online newspapers. They expressed a desire to see the widest possible range of newspaper content made readily available on a minimum number of separate sites. The pared-down online editions of most current major dailies were seen as insufficient for most research purposes. For retrospective digitization projects, users voiced a need for complete newspaper issues and comprehensive back-runs. They require sophisticated search capabilities, including full-text searching, some structured search options, and the ability to view the original newspaper page image, where articles can be seen in their original context. They also emphasized that online newspapers are a new medium worthy of preservation, as is the Web in general.

The newspaper industry outlined its challenge: to find a business model for online newspapers that will either be profitable or offset production costs, and will not jeopardize print subscription revenues. Newspaper publishers are not committed to assuring long-term access and preservation of their online versions, but neither are they currently prepared to hand off responsibility to a third party such as the Library and Archives of Canada.

Libraries situated themselves as access providers, supporting research uses of newspapers. When demand merits, they are willing to pay for core newspaper sources on behalf of their user constituencies, but likely at the sacrifice of print subscriptions. They do not see themselves being able to take on responsibilities for long-term preservation.

National libraries (not only the National Library of Canada, but also participants from the Library of Congress and the British Library) expressed their concern that long-term access and preservation of digital newspapers cannot be assured without industry cooperation and a collaborative national effort.

Presentations from aggregators and vendors illustrated the procedures, issues and problems involved in digitization. Public sector selection criteria and objectives differed from those of the private sector: the former emphasized completeness and totality while the latter focused on papers with broad national or international readership, or those having an existing microfilm client base. Technical challenges and the high cost of these projects were common concerns. Particular emphasis was placed on the need to plan and assess requirements and costs and to develop standards before undertaking a newspaper digitization project. Different information retrieval models were presented. Commercial products of high-profile daily newspapers such as the *Times* (London) and *New York Times* feature both full-text keyword and article-level indexing. Some large non-profit projects used software with automatic zoning of pages for more precise keyword retrieval. Others currently offer page-level keyword access, but are looking at developing an automatic zoning capability. One project provides page images only, with page-level access by newspaper title/place and date but not by keyword.

Outcomes

Participants identified four areas of particular concern: the selection of content for long-term access, the need for leadership, the need to define business models and an approach to funding, and technical issues. Extensive discussion coalesced around the following recommendations:

<u>The type of leadership that is needed, and from whom:</u>

- In the short term, the Library and Archives of Canada should initiate a working group of representative stakeholders to continue the dialogue begun with the consultation. This working group would flesh out the issues, define required actions, and develop a national digital newspaper preservation and access strategy.
- Consider establishing a Canadian Newspaper Centre. With the new Library and Archives of Canada in a leading role, such a Centre could address on a more permanent and high-profile basis the strategies developed by the working group, including legislation, standards, and preservation. It would lobby for funds, be a research centre of excellence, and develop digitization toolkits and other resources.
- Promote the value of digital newspapers as a source of information for Canadian education, research and innovation at all levels through advocacy by a coalition of leaders.

- Monitor newspaper access and preservation initiatives in other countries, including the goals and activities of the American working group on newspapers formed under the OCLC Digital and Preservation Co-op.

To ensure more Canadian digital newspaper content is produced and accessible over the long term:

- Comprehensive historical coverage should be the overarching goal, both for the preservation of currently produced online newspaper content and for retrospective digitization of historical Canadian newspapers.
- Respect the current complementary approach, whereby the private sector digitizes those Canadian dailies for which they see a market, and public-sector-based projects digitize newspapers of more local or limited reach.
- Strengthen pan-Canadian coverage by developing a strong provincial approach, building upon and extending the scope of the original Decentralized Plan for Canadian Newspapers.
- The Library and Archives of Canada or Canadian Initiative on Digital Libraries (CIDL) should inventory all online newspapers and retrospective digitization projects, including those of libraries, historical societies, genealogical and other associations so that the full extent of gaps in coverage can be better known.
- Prepare an integrated, pan-Canadian strategy to address retrospective newspaper digitization. The focus should be on needs assessment, funding priorities, a business model, coordinated collaborative applications for funding, and a plan for funding the digitization of newspapers of limited or niche-market interest. Such a document would assist in the development of private/public sector partnerships.

Funding a national strategy:

- Compulsory legal deposit of electronic newspapers is recommended.
- A business case for the digitization of Canadian newspapers should be prepared, and special government funding sought.

Toward national technical standards:

- Inventory standards employed in current projects.
- Develop a set of recommended practices. For example, digitization projects in Canada should support, as a minimum, full-text keyword searching and the ability to view the original page image.
- The Library and Archives of Canada should continue to provide leadership in dialogue/discussions on technical standards, which should be vendor neutral to allow maximum interoperability.

Stakeholders agreed that unless new relationships are established, collections are at risk and a body of valuable Canadian digital content will be lost. The importance of partnerships and of complementary activity between private and public sectors was underlined, and meeting participants affirmed that sustained leadership from the Library and Archives of Canada was required.

Detailed Proceedings

Day 1 "Born Digital" - Identification and discussion of issues related to newspapers that are created in digital format

Welcome and opening remarks by Dr. Roch Carrier, National Librarian, were followed by a panel discussion moderated by Dr. Ian Wilson, National Archivist. Speakers had been asked to consider models for online newspapers and their stability; whether more online newspaper content should be archived and made accessible in perpetuity, and whose responsibility this should be; and whether intellectual property rights and access to archival newspaper content were reconcilable.

There were common threads throughout the presentations. High priority was given to the need for leadership and coordination of roles and responsibilities, the need to re-visit newspaper preservation as a whole, and the need to establish legal deposit for electronic newspapers. Concerns were expressed regarding costs and funding models: owners require a return on their investment, while access providers, whose funds are limited, seek the lowest cost model, access to complete newspapers and comprehensive archives. The academic community's desire for downloading and emailing options elicited some discussion, since this would result in users being able to infringe on the proprietor's copyright. Academic, research and genealogical users stressed the importance of capturing full-page images in addition to article text. Most presentations acknowledged the technological challenges involved.

- Owner's Perspective - Andrew Martin, President of Infomart Ltd. and Financial Post DataGroup, and Director of Licensing for Southam Publications

  Models are still evolving. Generally, online newspapers are produced because publishers are trying to build financially viable business models around the technology, not in response to demand. Since revenues do not offset the costs of mounting an online presence, the prevailing approach is to offer decreasing portions of newspapers. The needs of advertisers and users continue to be researched, with consideration being given to challenges posed by demographics, education, technology, and user training and knowledge.

  The industry is not yet ready to take a position on increased archiving in perpetuity. Financial considerations, legal deposit obligations to the National Library, as well as the longevity, use and support of this new medium are among the problems that must be resolved. In addition, many publishers do not consider indefinite retention of their website material to be their responsibility.

  It may be necessary to assess the commercial value of various methods of dealing with web pages and their content before determining if intellectual property rights and access to archival newspaper content are reconcilable.

- Access Provider's Perspective - Nancy Peden, Carleton University Library

  Academic researchers have greater need for news databases than for the daily "freebies" on newspaper websites, which usually represent only a percentage of the print paper content. They require completeness, with information on national and international issues, content analysis over extended periods, and Canadian perspectives on foreign events and issues. A "smorgasbord" approach to selection, rather than fixed packages is preferred; interest is in "newspapers of record" vs. local dailies of small and medium towns and cities. Research would be facilitated by advanced searching capabilities, downloading/emailing options and timely -- preferably same-day -- access. Vendor stability and long-term access is a concern.

Individual universities with limited budgets do not have the resources to assume responsibility for preservation but can pay maintenance fees to ensure product availability and standards.

- User's perspective - John Reid, British Isles Family History Society of Greater Ottawa

  Genealogists need online newspaper archives with accurate search capabilities, all occurrences captured, and the ability to view images of the original articles. They require speedy delivery, complete coverage of all sections, access to international as well as national newspapers with a comprehensive transaction point, and assured long-term access. It was further noted that public awareness of the availability and potential of online newspapers should be promoted through increased marketing.

- User's perspective - Mike Gasher, Dept. of Communications, Concordia University, representing the Association of Canadian Studies

  Users have many questions about the future models for online newspapers. Will they be free, subscription based, specialized or general? Will online news sites demarcate their geography differently from their hardcopy editions? To researchers, online editions are not second-rate to their hardcopy counterparts; they consider online news sites to be distinct media of significant value. Dr. Gasher suggested that the National Library should collect online newspapers comprehensively. If this is not possible, a selection strategy should be developed for newspaper retention, and for websites and their internal links. Canadian archives and libraries must work together to develop a strategy for the preservation -- which may be decentralized -- of online material. The newspaper industry also has responsibilities in this area. Furthermore, Dr. Gasher suggested that multi-organizational licenses could be developed to facilitate national and international collaboration between researchers.

During the question period following the presentations, a number of further observations were made:

- The industry is still considering ways of allowing simultaneous access to a product by multiple users; possible options are price increases to the licensee or the use of technology to control access.
- Decentralized archiving could involve provincial and local libraries as part of a national system.
- Partnerships between the private sector and users may be needed to make models work. Various models may be established for specific groups of users and constituents.
- The reliability of non-traditional news items on the fringe of the commercial models may be determined with the passage of time; those that are effective and seem likely to survive should be selected for preservation.
- Paper may be bypassed when committing items to a more long-lasting medium, e.g. direct conversion from PDF to microfilm.
- Digital management should permit the preservation of various editions of the same paper.
- Committing websites to print will not preserve the ability to navigate through internal and external links.
- Decisions will need to be made regarding the frequency of capture, as news sites are updated continuously.
- Subscription online newspapers have not been studied as to their financial viability, size of readership, etc.

In the afternoon of Day 1, participants in breakout groups were asked to identify the key issues related to accessing Canadian newspapers online. Suggested considerations were roles and responsibilities, increasing the amount of content archived, funding, coordination, and aggregation. A representative of each group reported during a plenary session. The consensus appeared to be that content, leadership and business models/funding were paramount, closely followed by technical issues. The following summarizes discussions:

- Roles and responsibilities: A need for leadership and coordination of roles to ensure long-term preservation was identified. It was generally felt that the National Library should undertake this role. A clear delineation of roles is required, as well as guidelines and standards to assure inclusiveness, to foster cooperation, and to increase awareness of the value of digitization. Libraries can play a cooperative role at the local level, as can volunteers. As a major publisher in Canada, the Canadian government should also be an important player. The National Library, CIDL and publishers should work together to lobby for funds.

- Business models/funding: Public and private partnerships and approaches should be developed; hybrid models would accommodate the private and public sectors, libraries and publishers. The development of funding incentives (e.g. tax breaks) would encourage businesses to participate, since the relatively small market size of Canada is a significant challenge. Other challenges are the need to balance the commercial value of information against its value over time as a public good, and the lack of resources dedicated to the preservation and maintenance of digital content.

- Selection and aggregation: An inventory of newspapers and existing digitization projects is necessary to eliminate duplication. Comprehensive geographic coverage and inclusiveness in selection was recommended; the thematic approach should be avoided. However, practical compromises may be necessary (e.g. possible focus on the ethnic press, and/or smaller players). There is a need to define "digital newspaper" and the extent to which it reflects the "real" newspaper; the technology and the ease with which providers can put news on websites has blurred the boundaries. Online material, as opposed to that which is digital, could be considered a subset of content which exists elsewhere.

- Copyright: The Library and Archives of Canada should champion a change in the current state of copyright legislation in Canada. Due to the lack of legislation covering copyright in the digital environment, it is unclear whether archival copies may be made and what type of access to them is permitted.

- Standards: Standards and best practices should be established for the creation, preservation and re-purposing of content. There is some distrust of metadata standards; government-developed standards are not always fully practicable. There is evidence that institutions and companies are willing to cooperate and share practices already developed. Standards would also make feasible the use of centralized tools for data entry. Since content quality is variable, standards for the preservation of newspaper material must be flexible.

Day 2 "Re-born Digital" - Digitization of historical Canadian newspapers

Claude Bonnelly, Université Laval, moderated the morning presentations on non-Canadian and Canadian digitization and access projects. The focus was on models and lessons learned. Summaries of presentations are as follows:

- British Library Newspaper Project - Ed King, British Library

The British Library is committed to keeping all UK newspapers in their original format, but recognizes the need to facilitate access to them. The objectives of this pilot project were to test scanning from negative microfilm, supplying images/text in full via Web browsers, and the accessibility of the resulting product. Preservation and copyright issues were not part of this pilot. The cross section of titles chosen ensured a variety of fonts and subject searching. Microfilm reels were sent to OCLC for processing and Olive Software was used for automatic de-segmentation of pages, where zoning and data capture, as well as indexing of every word were instituted. Next steps will be to evaluate the selection process, examine costs, consider issues related to preservation/metadata, assess the ability of the software to process a large volume of material, tackle copyright and plan subsequent projects.

That partnerships between the private and public sectors are essential was evidenced by the involvement of OCLC, Olive Software and King's College, London, who hosted the data. Other conclusions and lessons learned are that open communication facilitates rapid progress, that limited funds make careful selection a necessity, and that copyright, standards and system interoperability must be taken into consideration.

- Times Digital Archive - Salvy Trojman, Gale Group

  Gale's strategy for digitization is to seek newspapers with broad national or international appeal, those with an existing client base in microfilm, or with potential for library sales or national licensing. In the case of the *Times* of London project, the objective was to make the historical issues attractive and timely and to produce an easy-to-use that would return high-quality search results. Issues were complete and presented in a searchable facsimile format. Procedures included acquiring rights from publishers, consultation with customers, scanning from analogue to digital, and image post-processing such as cleaning, zoning, article clipping, the addition of category type and metadata.

- Proquest Historical Newspapers - Stephen Abram, Micromedia Proquest

  When planning a digitization project, primary considerations should be an assessment of the source material (format, condition, rights, etc.), of potential uses, as well as of the users and their experience. In addition, technical standards should be defined, the method of accessing the material should be determined and cost estimates doubled. It is also important to determine infrastructure needs, select manufacturing options and monitor quality control.

  Ten major newspapers will probably be selected for conversion during this project, including regional, ethnic, and local newspaper components. Cost considerations made stringent selection criteria a necessity. Experience shows that, while article zoning produces faster image download, better OCR rates, article level blocking and searching, the associated costs may make it impracticable except for major papers.

- OCLC Digitization and Preservation - Robert Harriman, OCLC

  This presentation focused on the activities of the Historic Newspaper Archive, a working group within OCLC's Digital and Preservation Co-op. The group created a vision, or list of desirable elements for digitized newspapers. Their goals and objectives include considering the best development model to serve the nation's research needs, providing a framework for negotiating rights, and identifying/developing metadata standards. High priority is also given to exploring interoperability issues, developing criteria for funding, and promoting advocacy at local, state and national levels. A Technical Advisory group has been formed to move the vision forward by identifying issues/problems, developing pilot projects, preparing papers and reports, cost models, etc.

- Alberta Heritage Digitization Project - Tim Au Yeung, University of Alberta Library

  Digitization of provincial historical materials, with particular emphasis on perceived gaps in accessible materials, is the project objective. Totality and completeness, meaning facsimile presentation and the inclusion of every newspaper regardless of size, are the governing principles. The collection spans daily and weekly newspapers from major urban centres and smaller towns. Procedures include the use of user help to assist in quality control and for input on interfaces. Conclusions and lessons learned include the need for a substantial infrastructure to support this type of project and that manual intervention is expensive. Prior to any future format translation, it will be necessary to balance the risk of file migration against re-digitization. Incorporation of locally created indexes into search mechanisms, and a platform allowing full-text searching with more sophisticated interfaces are planned.

- Paper of Record - Bob Huggins, Cold North Wind

  The Cold North Wind recognizes the importance of newspapers as a record of continuous daily life during the past 500 years. Its vision is to create a database of knowledge that allows individuals to look at historical events day by day from different perspectives. Accordingly, the company does not concentrate on just the top ten or twenty newspapers, but takes a broader approach. It is building a substantial body of North American newspapers, with a number of major goals already achieved. Paper of Record is an historical archive of full-page newspaper images, in their original format, dating from the 1700's. The project is built on partnerships with organizations that own valuable collections of historical newspapers on microfilm. Its interface and search capabilities are similar to those used in projects described previously. The company's philosophy is to use a more automated process; it believes in volume, and in speed to market and to mass. The underlying technology is Zylab, from the Netherlands. Processing involves scanning the original records from microfilm; the image collection is searchable through optical character recognition (OCR) software. Because of this project's magnitude, assistance was requested -- particularly in establishing standards -- in recognition of the importance of metadata and to promote data sharing.

- Pages of the Past (*Toronto Star* and *Globe and Mail*) - Stephen Abram, Micromedia ProQuest

  In addition to outlining the digitization of the *Toronto Star* and *Globe and Mail*, the major challenges of the new medium were emphasized. Librarians must master new search and conversion paradigms, develop image manipulation skills and skills for searching huge databases. Micromedia ProQuest is focusing on the market and sales aspects (e.g. providing what the library market needs and managing the adoption of the products into the libraries and institutions of subscribers). Entertainment, sociological research, advertising, and a view of contemporary life in another era are the primary uses of digitized products. Future projects should take into consideration unintended consequences (i.e. easy access to this information could have an impact on theses and dissertations, school textbooks, Canadian perspectives on world issues, as well as public opinion).

- HALINET Halton Newspaper Index - Brian Bell, Oakville Public Library

  The project's objective is to create an index of local community newspapers and local community content, giving clients one place to look for information. Goals are the integration of public library catalogues with commercial databases, with aggregator sources, with community information databases, and with photograph databases, etc. An open-source consortial, cooperative approach has been taken in developing the application. Progress to date has been

reached using local budgets and much volunteer help. HALINET is committed to complete a similar module which libraries may freely use to index images.

The possibility of distributing this product as a CIDL benefit was mentioned. Decisions would be required as to storage of the centralized index module and database. Responsibility for the archival unit, the use of originals vs. microfilm as the image sources, and the format of digital originals (which increasingly are in PDF at the community newspaper level) are other concerns. Better local scanning management, FTP and secure archiving of originals, shared thesaurus maintenance, and name control are needed.

A national vision of the Library and Archives of Canada managing a Dublin-Core-based, harvested central database pointing to local or regional databases and national data providers was proposed.

Breakout groups in the afternoon session were asked to recommend strategies for future action. Their discussions centered on content selection, leadership and coordination, technical standards, and funding. Group representatives reported their discussions to a plenary session. The following is a detailed list of recommended actions:

Content (2 groups)

- Inventory all online newspapers and current digitization projects, including those of libraries, historical societies, genealogical and other associations.
- Aim for comprehensive geographic coverage, both for the retention of currently published online newspaper content, and for retrospective digitization projects.
- Support full-text keyword searching and the ability to view the original page image for digitization projects.
- Prepare a pan-Canadian, comprehensive scoping document to assess all aspects of digitization, including funding, and assist in the development of private/public sector partnerships.
- Explore the adaptability of the CIHM model.
- Establish a base model for generalized descriptors to enable users to quickly identify relevant content.

Leadership (2 groups)

- Establish a Canadian Newspaper Centre, representing sectors as well as users, to provide a physical presence and virtual focal point for cohesive strategy. With the Library and Archives of Canada in a leading role, the Centre would address issues such as legislation, standards, and preservation. It would lobby for funds, promote theme building or centres of excellence, and develop toolkits and other resources.
- Prepare a business case for the digitization of Canadian newspapers.
- Promote the value of digitized newspapers as a source of information for Canadian research at all levels. Advocacy by a coalition of leaders is recommended.

Technical Standards

- Inventory standards of current projects as well as standard practices.
- Develop an inventory of recommended practices.

- Recognize that long-term preservation could require different media for different purposes. The image preserved may be separate from the image seen by the end-user. Both microfilm and digitization may be needed.
- The Library and Archives of Canada should continue to provide leadership in dialogue/discussions on technical standards, which should be vendor neutral to allow maximum interoperability.
- Establish a system of decentralized repositories with description standards, etc. Standard metadata may avoid unnecessary duplication of files.

Funding

- The Library and Archives of Canada should form a working group to address the question of newspaper digitization using an integrated approach. The focus should be on needs assessment, funding priorities, a business model and coordinated collaborative applications for funding, and funding for the digitization of newspapers with limited or local markets.
- Recommend compulsory legal deposit of electronic newspapers.

Questions and discussion following the plenary reports:

- Funding can be addressed at three different levels; there are options for individual projects.
- Decentralized deposit of newspapers might be a solution.
- It is assumed that users will access digitized products through institutions, and that there will not be individual user fees.
- Consortial buying is already in place in certain areas.
- There may be artificial distinctions between a consumer and a citizen.
- The private sector has led digitization to date, but leadership by government is recommended.
- The focus should be on promoting digitized products rather than gathering and vaulting the information.
- Coalition of leadership can lead to a single voice using the marketing and communication skills of its various components.

Closing

Susan Haigh, Library and Archives of Canada, provided a brief summary of the two days and thanked the organizers, speakers and participants who had made the event so stimulating and successful.

Prepared by Janet Martin and Susan Haigh, November, 2002.