Digitization for Preservation and Access: a Case Study

Marilyn Berger
Head Librarian (Acting)
Blackader Lauterman Library of Architecture and Art
McGill University
3459 McTavish Street
Montreal, Quebec
H3A 1Y1

Library Hi Tech 1999 17 (2): 146-151

Marilyn Berger is the Head Librarian (Acting) of Blackader Lauterman Library of Architecture and Art, McGill University, Montréal, Québec (E-mail: marilyn.berger@mcgill.ca) She holds a MLS degree from McGill University.

Acknowledgements:

The author wishes to thank Carol Hutchison (MLIS) recent graduate from the McGill University Graduate School of Library and Information Studies who was project manager. Her complete documentation of the project provided much of the background information for this article. In addition, thanks are due to the Blackader Lauterman staff for their energy, creativity and involvement in the project, particulary Kayhan Boncoglu and Sylvain Pimparé for the documentation on Hardware and Software specs.

ABSTRACT

This article describes the process whereby digitization of a journal is used to preserve and provide access to the contents on the World Wide Web. The journal chosen, The *Canadian Architect and Builder*, was the only professional architectural journal published in Canada before World War I. The original printed version is in an extremely fragile state and the digital version will provide accessible, electronic access to the contents. Funded in part by the federal government's Youth Employment Strategy, the Department of Canadian Heritage and the Canadian Library Association, the project objectives and goals are described and the steps taken towards those goals are presented.

Digitization for Preservation and Access: a Case Study

INTRODUCTION

The *Canadian Architect and Builder*, published from 1888 to 1908 is the only professional architectural journal published in Canada before World War I. This journal documents a complete history of the architectural and building crafts in Canada. Profusely illustrated, the journal contains articles and advertisements from the industry of the time, providing an invaluable architectural, social and cultural record of this early period. It is a source of historical background to anyone doing research today in architecture and other disciplines in the humanities and social sciences.

The printed version of the *Canadian Architect and Builder* in its entirety is held by only 3 or 4 institutions across Canada, McGill University being one. The original volumes were in an extremely fragile condition making direct physical access difficult. Most researchers coming to Blackader Lauterman Library to consult the journal were permitted access to the microfilm version only. Research was laborious and time-consuming. Microfilm has been viewed as a medium of storage rather than of access and with increased availability of computer workstations in libraries and connection to the Internet readily accessible, electronic access to research materials is a concept whose time has arrived[i] (Belinger, 1998, p. 178).

The McGill University Digital Collections Program and Blackader Lauterman Library have been involved in several projects using technology for the preservation of library material and then providing access to this material. The aim of the *Canadian Architect and Builder* project was to provide on-line access to the full text of the *Canadian Architect and Builder* through the World Wide Web. The wide community of scholars and researchers in all related disciplines would gain convenient, easily searchable access to the full run of this important journal. Another aim was that if selected, the on-line electronic version of the *Canadian Architect and Builder* would be McGill University's contribution to the **JSTOR** (Journal storage) project which provides desktop delivery of selected journals through the World Wide Web (see *Library Hi Tech*, v. 16, no 1, 1998). There has not been a similar architectural project to date and the *Canadian Architect and Builder* could thus provide a subject specific model in the designated area.

The opportunity to participate in the Young Canada Works in Science and Technology program prompted us to apply for funding from this source. As part of the federal government's Youth Employment Strategy, the Department of Canadian Heritage and the Canadian Library Association would share costs for a portion of the project wages and benefits, transportation and accommodation if applicable. Projects were to focus on providing science and technology graduates with intern experience and opportunity to acquire marketable skills while strengthening Canada's culture and heritage and enhance knowledge about Canada's history, creativity and achievements.

Through the "Young Canada Works in Science and Technology" program and cultural institutions or organizations, technology-based enterprises and scientific research organizations were invited to become partners with the Department of Canadian Heritage and Canadian Library Association and develop internships to help youth acquire the critical experience necessary to enter the labor force.

The *Canadian Architect and Builder* project met all the criteria for the Young Canada Works program and funding, in part, was approved. A recent graduate student from the McGill School of Library and Information Studies was hired to spearhead the project and work began in June 1998 to November 1998.

BACKGROUND INFORMATION

Preliminary research on digitization involved visiting Web sites and reading articles on the subject. A bibliography of Web sites and articles is included for further reading (see Appendix I). A study of various online journal projects, including the **JSTOR** Project[ii] [1] and the European project **Decomate** provided us with the information to compare formats and methods used to mount them on the Web. The **Decomate** project's main goal was to provide end users access – through the library – to copyrighted materials distributed by commercial publishers in electronic form. Attention should be paid to the copyright issue to encourage cooperation between publishers and libraries[iii]. (Dijkstra, 1998, p. 243) The issue of copyright with the *Canadian Architect and Builder* was investigated, but because of the age of the journal no clearance was necessary.

The **JSTOR** Project, funded by the Mellon Foundation is centered at the University of Michigan. With cooperation from publishers, the project grew from an electronic database of the back runs of 10 core journals in the fields of economics and history to 106 participating journals with 64 journals currently available online[iv]. (Shoaf, 1996, p. 232)

What became clear from the literature was that digitization was an exciting preservation option while providing unparalleled access available to all. The technology is advancing rapidly and this raises the question of accepted standards for digital preservation technology. The Blackader Lauterman Library *Canadian Architect and Builder* Project follows the standards for image capture, resolution, data transfer protocols, indexing, access, and file types.

A process was developed for creating an online version of the *Canadian Architect and Builder* (*CAB*). The steps involved included:

1. Image preparation; preparation of the volumes, issues and pages of the journal for scanning
2. Scaning pages of *CAB*
3. Editing images of *CAB*
4. Using Optical Character Recognition Software (OCR) on images to edit text versions.[v] [2]
5. Creating a searchable database of the text
6. Linking text to images
7. Mounting on Web

IMAGE PREPARATION

Due to the historical importance of the *Canadian Architect and Builder*, the first step was to take all measures that were necessary to preserve the hard copy of the volumes. A physical review of the journal revealed much variance in the condition of the pages and that, while some could withstand scanning without any preservation, other pages would not. Consultation took place with the preservation library assistant in the Rare Books and Special Collections Division, Social Sciences and Humanities Library at the University and procedures were set up to follow. The following recommendations were made to preserve the *Canadian Architect and Builder*:

1. Remove issues from their bindings
2. Separate each issue page by page
3. Trim each page so that no rough edges are left
4. Mend rips and tears with acid-free minding tape
5. Put each issue in an acid-free envelope and store each volume in an acid-free storage box

The Work-Study students working in the library during the summer months were trained in preservation techniques to prepare the volumes for scanning and work flow was kept at a steady pace to finish this tedious task. The procedures followed will ensure the preservation of the hard copy for future generation of users and will also make the task of scanning easier.

During the process of removing the volumes from their binding, a note was made of any pages, issues or volumes that were missing, or of any pages that had parts torn or missing that would affect the text or graphics. A list of missing pages, issues and volumes was compiled with the intent of borrowing these items from other libraries that owned them. In particular, the Bibliothèque Nationale du Québec and the University of Toronto Rare Books Department were extremely cooperative and loaned the missing items to McGill for scanning.

Hardware and Software considerations were decided and a list of Hardware and Software programs used is attached. (see Appendix II)

IMPORTANT FACTORS TO CONSIDER FOR FORMAT

At the same time as work was being done to preserve the *Canadian Architect and Builder*, decisions had to be made as to the most appropriate file format for the online version of the journal. The integrity of each page had to be maintained and in order to do this, it was decided to save the scanned pages as images. In consultation with the Digital Collections Librarian, two primary options were identified:

1) Save the image as a PDF (Portable Document Format) file using Adobe Acrobat.

or

2) Save the image as a GIF (Graphical Interchange Format) or JPEG (Joint Photographic Experts Group) file.

A literature search was conducted to obtain information on both the above options in relation to digital collections. Several Web sites were visited, including two that represented the different types of files under investigation. (see Appendix I) **JSTOR** uses GIF files to display the textual images and the *Osler Library Newsletter*[vi] [3] at McGill University stores its information in PDF format. This review led to the conclusion that saving the images as JPEG files would be most appropriate. JPEG will allow the entire project to be uploaded as an HTML file and furthermore, JPEG retains more information as it compresses an image.

- Each page had to be kept intact so that the user could view the page online just as it appeared in the paper copy.
- User-friendliness was important. Therefore, the online version had to be searchable, had to have a minimum of side-to-side scrolling, had to load fairly quickly and the pages had to be clear and easy to read.
- Would Optical Character Recognition (OCR) work on scanned images of the *Canadian Architect and Builder*? This was a consideration as several different fonts were very small, and some were old-fashioned script which cannot always be recognized by OCR programs.
- With standardization the maximum number of users should be able to view it.

SCAN PAGES OF CANADIAN ARCHITECT AND BUILDER

Having determined the type of file to be used for the online version, several preliminary scans were conducted to determine the appropriate image type, brightness, contrast, height and width, as well as the length of time it would take for each page to be scanned. To ensure uniformity in scanning the following parameters were set up:

Height:	12.65 in.
Width:	8.18 in.
Brightness:	125 (approx.)
Contrast:	130 (approx.)
Image Type:	Black and White Photo
File Type:	JPEG
Image Quality:	600 dpi

A file structure was then created to save the images. A directory called "CAB" was set up with sub-directories created for each volume. Sub-directories were then created within each volume for every issue. For example, each page scanned would be named individually (i.e. v11n2p151), each issue would be saved together in a folder (i.e. Issue 2) and all issues would be saved together in a volume folder (Volume 11). This logical progression made it possible to keep the order intact, even when an individual page was missing and had to be scanned at another time. Scanning began with Volume 11 as that volume was previously promised for an exhibition, however, the remaining volumes were scanned in chronological order.

PROBLEMS ENCOUNTERED

Scanning the images at 600 dpi (dots per inch) created huge files, which, in turn, created a storage problem. As well, scanning at 600 dpi was extremely slow and a compromise had to be reached. In the interest of finishing the scanning process, the change was made to 300 dpi and an archival quality copy was not to be had.

Another problem occurred because of scanning the images in grayscale. When users would want a print out or download a page or article the background would be gray and it would take longer to print. The problem was solved by scanning in black and white but the software had to be changed from Adobe Photoshop to Xerox Pagis Pro[vii]. [4] Pagis Pro proved to be an excellent software for text as well as for the drawings and photographs. Previous scanning done on Adobe Photoshop was converted to black and white using Pagis Pro.

Finally, the decision was made no longer to save the images as JPEG's. Although this format is one option for displaying the images on the Web, the original scanned images will be saved as TIFF's[viii] [5] for storage purposes. The images will then be converted to an appropriate format for the Web. Due to the differences in the scanning software, the height, width and other parameters of the scanned images will be determined during the editing phase, rather than during the scanning phase.

Phase II of the project will involve:

1. *Editing of the Images*
   Blemishes must be removed from the pages, the height and width of the images have to be standardized and the images must be blurred or sharpened as necessary.
2. *Optical Character Recognition (OCR) of the Images*.
   In order to make the images searchable, a text version of the *Canadian Architect and Builder* is necessary, and so an OCR versions of the images must be done. Many different fonts and scripts are used in this journal and the OCR software sometimes has a problem recognizing the text. This will have to be solved as part of the editing process.
3. Mount the Images on the Web
4. Link the OCR version to the images
5. Create a searchable database of the journal

CONCLUSIONS

The World Wide Web has enhanced the means to access material previously available only to a select few. Access to the *Canadian Architect and Builder* is a case in point and, whereas in the past, researchers often had to travel many miles to consult the physical volumes, new advances in the technology, namely digitization of images, will allow universal access to the journal.

A project page has been developed to create awareness in the profession that the project is in process. This can viewed on the Web [6]:
<http://blackader.library.mcgill.ca/cab/cabproj.htm>.

Appendix II

HARDWARE AND ACCESSORIES

1. Pentium PC 233 Mhz with MMX support
   - This is a minimum system requirement. MMX is Intel's (maker of the Pentium processors) microprocessor with enhanced multimedia capabilities.

2. 4.3 Gigabyte Hard Disk Drive
   - A big clean hard disk speeds up the entire system. Windows95 uses hard disk space to conduct what is know as *swapping*, which places commonly used tasks on a part of the disk to access easily. This function often extends the limitations of actual RAM memory because eventually, all the memory can be used up if more than one program is working at the same time, which is very often the case. Although the actual images themselves were not saved entirely on our local disk, some of the programs needed for the project are fairly large (i.e. Microsoft Office 97, Pagis Pro 2.0, Adobe Photoshop, etc…)

3. 64 Megabytes EDO RAM
   - RAM (Random Access Memory) Since some images can be fairly large 32 megabytes is an ABSOLUTE minimum. We had 32 and upgraded to 64.

4. ATI 3D Charger 4 Megabytes
   - This is the video adapter card. Nothing less than 4 megabytes of video memory should be used for a scanning project such as CAB. Even 8 megabytes would have been much better in terms of speed and quality of output.

5. Linksys 10/100 (Base T type) Network card
   - A common 10 Base-T style network card was used. This was used to upload our images into a larger server residing on campus with a larger hard disk.

6. 17 inch Relisys Super VGA Color Monitor
   - 17 inch monitors for a project like CAB are not a novelty, but really a necessity. Staring at color images on a smaller screen can be tiring.

7. Microtek ScanMaker III (Flatbed style)
   - 36-Bit Color: 12 bits per RGB color. Approx. 68.7 billion colors
     600 x 1200 dpi. This is perhaps one of the most important considerations when thinking about such a project. It should be high resolution and to be cost effective, FAST. Nothing is more annoying than scanning thousands of pages with a slow scanner.

SOFTWARE

1. Xerox PagisPro 2.0
   (http://www.xerox.com/scansoft/pagis/)
   - The premiere scanning suite full of utilities and programs. This program was used to save the images in .XIF format, a very good quality format which maximizes small file sizes. This is a very important consideration because these scanned images can use a lot of disk space.

2. Adobe Photoshop 4.01 (http://www.adobe.com)
   - This program is used for the actual editing of the images and the conversion into the .GIF format to be used on the Internet. It is a good program for graphics. It has many features and editing tools.

3. Microsoft Word 97
   (http://www.microsoft.com/products/default.htm)
   - Used for any word editing once the OCR (optical character recognition) process is complete. Good spell checker and find functions.

4. Microsoft FrontPage 98
   (http://www.microsoft.com/products/default.htm)
   - This is the HTML (Hyper Text Markup Language) editor to be used to construct the web pages. and in fact manage the entire Web site. A project such as *CAB* will have thousands of web pages with links and files. FrontPage visually shows the links and shows any broken links. Best when used to edit directly over the network (where the files will actually be placed. That way the broken links and the like will be detected and can easily be edited and updated. This program is a full fledged web site management system not to be overlooked and a necessity. Another similar program is NetObjects Fusion.

5. Norton Anti-Virus 5.0
   (http://www.symantec.com/nav/index.html)
   - Used to scan all files for viruses. Since network connections are unavoidable, a virus scanning program is necessary.

ADDITIONAL CONSIDERATIONS

1. A Network drive
   - Depending on the size of the local hard disk and the magnitude of the project, a network drive may be needed. If the local hard disk is 12 gigabytes there is little need for a network drive. However, a network drive is needed when local hard disk space runs out and/or to make backups of scanned images.

2. Backup system
   - It is recommended to backup onto tape which can automatically be programmed to do so every night.

BIBLIOGRAPHY

"Building Large-scale digital Libraries." *Computer* 29:5 (May 1996). Available online: <http://computer.org/computer/dli/index.html>.

"The JSTOR Production Process." [Online], 1998. Available: http://www.jstor.org/about/production.html [1998, May 14].

Canadian Initiative on Digital Libraries. Available: http://www.nlc-bnc.ca/cidl/cidle.htm [1998, May 11].

Chesnutt, David R. "The Model Editions Partnership: "Smart Text" and Beyond". *D-Lib Magazine* [Online], July/August 1997, 5 pages. Available: http://www.dlib.org/dlib/july97/07chesnutt.html [1998, May 19].

The Getty Information Institute: the Imaging Initiative. Available: http://www.gii.getty.edu/index/imaginit.html [1999, January 25]

Guthrie, Kevin. "JSTOR." February 18, 1998. Available: http://www.mellon.org/jsback.html [1998, May 14]

JSTOR Available: <http://www.jstor.org/>.

Kirchhoff, Amy and Mark Ratliff. "JSTOR Expanding: New Title and Tools." [Online], September 2, 1997, 4 paragraphs. Available: http://www.princeton.edu/cit/campus_com/jstorsep.html [1999, January 25].

Kuny, Terry. "An Introduction to Digitization Technologies and Issues." *Network Notes*[Online], no.14, October 1, 1995, 9 pages. Available: http://www.nlc-bnc.ca/pubs/netnotes/notes14.htm [1998, May 11].

Osler Library Newsletter Available: <http://imago.library.mcgill.ca/osler/>.

Ratliff, Mark and Amy Kirchhoff. "JSTOR Celebrates its First Birthday." *CIT Info* [Online], February '98, 10 paragraphs. Available: http://www.princeton.edu/cit/campus_com/jstor2.html [1998, May 19].

RLG preserv project website. Available:
www.rlg.org/preserv/ [May 13, 1998].

Steinberger, Mark. “Making Optimal Use of the Electronic Environment.” *Proceedings of the Conference on Electronic Communication in Mathematics* [Online], May 29-June1, 1997. Available:
http://www.geom.umn.edu/docs/cecm/steinberger/talk.html
[1998, May 14]

Xerox Pagis Pro 2.0 Available:
<http://www.xerox.com/scansoft/pagis/>.

Bellinger, Meg. “The Transformation from Microfilm to Digital Storage and Access.” *Journal of Library Administration* 25:4 (1998): 178

Dale, Robin. “Selection for Preservation and Access: Notes from the PARS/CMDS Program.” *ALCTS Newsletter* 6:6 (1995): 178.

Dijkstra, Joost. “Journals in Transition: From Paper to Electronic Access: the Decomate Project.” *Serials Librarian* 33:3/4 (1998): 243-270

Shoaf, Eric. “Preservation and Digitization: Trends and Implications.” *Advances in Librarianship* 20 (1996): 232

NOTES

[i] Meg Bellinger. “The Transformation from Microfilm to Digital Storage and Access.” *Journal of Library Administration* 25:4 (1998): 178.
[ii] JSTOR can be accessed at<http://www.jstor.org>.
[iii] Joost Dijkstra. “Journals in Transition: From Paper to Electronic Access: the Decomate Project.” *Serials Librarian* 33:3/4 (1998): 243.
[iv] Eric Shoaf. “Preservation and Digitization: Trends and Implications.” *Advances in Librarianship* 20 (1996): 232.
[v] OCR is the process whereby a computer program “reads” the text from an image of a document and converts it into ASCII text.
[vi] Osler Library Newsletter can be accessed at<http://imago.library.mcgill.ca/osler>.
[vii] Xerox Pagis Pro 2.0 can be accessed at<http://www.xerox.com/scansoft/pagis>.
[viii] TIFF – Tagged Image File Format.