Date : 16/07/2008

Canadian Mass Digitization: the University of Toronto Libraries partnership with the Internet Archive – historical overview, recent issues, and future implications

Jonathan Bengtson,
Associate Librarian for Scholarly Resources,
University of Toronto Libraries, Canada

&

Robert Miller,
Director of Books,
Internet Archive (IA),
San Francisco, California, United States

Meeting: **139. Section on Acquisition and Collection Development: "In and Out (of Copyright): Contrasting Perspectives on Digitization of Library Collections"**

Simultaneous Interpretation: English, Arabic, Chinese, French, German, Russian and Spanish

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

The University of Toronto Libraries (UTL) and the Internet Archive (IA) have been collaborating on mass digitization projects since late 2004. Currently the Toronto scanning centre maintains 23 IA "Scribe" machines, capable of digitizing between 40-50 million pages of text per annum.

The projects undertaken by UTL over the past four years have fallen under two general categories: 1) pre-1923 materials digitized as part of the Microsoft libraries partnership, and 2) materials digitized by UTL and other Canadian institutions using using the UTL/IA scanning centre under the principles of the Open Content Alliance (OCA). In the latter category the majority of the work has been on pre-1923 materials but there are also numerous examples of more recent materials as well. Pioneering efforts are also under way for scan-on-demand; an interactive service feature to broaden the reach of digitized content to remote users.

This paper will review the history of UTL's involvement with the IA in mass digitization within the wider context of how the various projects (including the MSN partnership) relate to the Google books initiative and the ultimate potential that mass digitization presents to scholarship. An overview of recent and ongoing discussions at the IA, Open Content Alliance (OCA) and MSN of how to move into scanning more recent materials (especially the 1923-1963 period) will be covered. An overview of issues related to technical aspects of the projects will also be included, with an assessment of areas of future opportunity in the scanning, preservation and presentation of digitized materials.

The bulk of the paper will discuss what the potential implications are for scholarship now that libraries and their partners are building massive collections of digitized materials. Topics covered will include: what tools need to be designed to make the most use of materials, what technological barriers remain to be overcome, what partnerships (public and private) will help university libraries exploit digital texts, how public domain (i.e. out of copyright) digital texts relate to licensed (i.e. in-copyright) e-books and other electronic resources, and what are the implications to scholarship for the changing dynamic between electronic information and print-only information

Paper Details and Presentation format- A face-paced Q&A will be conducted between Jonathan and Robert; both experts in their respective areas; both committed to building an open on-line digital library.

1.Why do we need a digital library and who should be its audience?
Answer- Discussion on real and anticipated needs, review of audience/users.

a. k-12,
b. university and researchers
c. general audience; namely those that never will walk into a library and the general public – including the international perspective.

2. What should a digital library be comprised of?
Answer-

a. from the user perspective, it should be a collection reflecting all materials that may be accessed by digital means, it should allow for the linkage to various formats, it should be searchable (included federated search), it should allow for user/professional updates. It should be flexible and dynamic; allowing for contribution and updates. It should be open.

b. from the preservationist perspective it should utilize digital formats that have perceived longevity; allow for forward migration of data, reflect resolution levels that fit modern preservation guidelines, offer dynamic meta-tagging, allow for citation linking, broad indexing and post capture updating. i.e. jpg 2000s for text and still images, mp3/4 for video, audio files, etc. It should be open and have redundancy of storage.

c. in short, a digital library should be a community; a balance between structure and vibrant change.

3. What should the philosophical underpinnings of a digital library be?
Answer-. Discuss various models-

a. Google

b. MSN

c. Wikipedia

d. OCA

4. What are the structural issues that should be considered?

a. storage

b. selection

c. access

d. standards

e. future potential for service layers

f. need for collaboration and multiple aggregators or contributors of material

g. in short, most of these are in place already. These should not be a limiter.

4. Who should be the stake holders in this effort?
Answer-

a. libraries themselves; part of preservation/collections budget
b. special collections holders; archival collections; i.e. LBI institute
c. public, university, state libraries, national libraries; support their user community
d. funders- Sloans, Mellons, Moores; goal congruent with their mission
e. communities of users- examples- i.e. genealogists, scientists

5. Why have existing initiatives fallen short or not met the library needs?
Answer-

a. Google/MSN- limits of what they are doing and why their vision might not match the libraries vision.

b. Individual library efforts- i.e. NYPL photo project.

c. OCA- a good start in terms of principles, but it is not a funded model.
d. In sum, individual efforts are extremely resource greedy; back end storage, process to capture images, content selection; not very efficient!

6. Are there any examples of projects that begin to come close to doing this successfully-
Answer-

a. BLC- self funded- 17 libraries, pluses and minuses, uses OCA principles
b. BHL- foundation funded- 10 libraries, portal design, uses a service layer approach for access, IA does back end, uses OCA principles.

c. Show example of digital scanning center- U of T.
d. Show evidence of progress- reference download stats. i.e. Adams collection

7. What prevents efforts like these from being collaborative?
Answer-

a. libraries aren't organized to typically share well- goals are aimed internally
b. funding doesn't reward sharing
c. preservation and collection budgets aren't aligned necessarily with new opportunities; e.g. digital library
d. new paradigm should be reviewed to make this happen
e. it might take an outside group, project or vision to tie them together-
f. example of efforts that could/should be collaborative- OCA efforts, Million books project, Google project

8. So going forward, what is needed?

Answer- We need an idea that is big enough to capture the imagination, yet simple enough to fund to fund/support
a. Project concept- immigration; include gov docs, oral histories, maps, music, texts.
b. Project concept-what is the inalienable right (or expectation) that every student from kindergarten through graduation should have in their digital backpack).
c. Project visionary- who will be the modern day Andrew Carnegie; funding the creation of the 21st century digital library? What a legacy!
d. Project concept- Hidden Collections

e. We must dream big and dream past the boundaries of specific institution

9. How could this be funded?
Answer-

a. Gov't- pros/cons, time line
b. Private- pros/cons, time line

c. Sugar daddy/moma
d. Peanut and butter approach
e. Other.

Bios:

Jonathan Bengtson is the Associate University Librarian for Scholarly Resources at the University of Toronto. He oversees the collection development, technical services, ordering, and serials departments, as well as provides leadership in scholarly communications, preservation and digital initiatives. Jonathan has held various senior positions in academic, research, and nonprofit libraries in Canada, the United States, and the United Kingdom—including Executive Director of the Providence Athenaeum (founded in 1753) in Providence, Rhode Island; Head Librarian of the Queen's College, Oxford (founded in 1341); and, Chief Librarian of the University of St. Michael's College, University of Toronto. Jonathan has been a member of the United Kingdom Library Association's National Council and served as Vice President/President-Elect of the Consortium of Rhode Island Academic and Research Libraries. He is currently on the Board of Directors of the Society for the History of Authorship, Reading, and Publishing and is the coordinator for the University of Toronto's partnership with the Open Content Alliance and the former Microsoft Live Books mass digitization projects. He holds a summa cum laude BA in history from the University of California and post-graduate degrees from Oxford University (medieval history) and University College London (library studies). Jonathan is active in publishing and teaching on topics in the history of the book, the history of libraries, modern library management, and medieval history.

Robert Miller- Presently heads the Global Digitization program with the Internet Archive. He built the operation that presently runs in five countries and in 16 locations. Prior to joining the Internet Archive, Robert founded or co-founded five start up companies; including CEO of the Internet Search company, Focus Engine. Prior to that, Robert held senior management positions with Mattel Toys and Tyco/AMP Electronics, both industry leaders. Robert holds a BSIE from Lehigh University and set up sales, marketing and operations globally with long term assignments in Asia and Europe. Robert has lived in Germany and Afghanistan. It is with great pride, that Robert notes his first volunteer job was in a public library when he was 10 years old! It's nice to be back in the library world again!