Mass Digitization

Sian Meikle
University of Toronto Libraries
Outline

- Context
- History
- Products and process
- Observations about use, objectives arising
- What’s next?
University of Toronto Libraries

- 18 million volumes
- 45 libraries, 3 campuses
- 78,000 FTE users (staff, students, faculty)
Partnering with the Internet Archive

The University of Toronto is one of the five largest academic libraries in North America.

The Internet Archive is a non-profit organization, based in San Francisco, that was founded in 1996 to build an ‘Internet library,’ with the purpose of offering permanent access for researchers, historians, scholars and the general public to historical collections that exist in digital format.

www.archive.org
Internet Archive
Book Scanning experience

- 1996 registered as a non-profit
- 2003 (India) Million books project
- 2004 Sloan grant, equipment evaluation, trial scanning
- 2006 Production scanning, 3 sites
- 2007 8 sites; 5 million pages or 12-15,000 books each month
- 2008 18 sites; 10 million pages or 25,000 books each month
Open Content Alliance: Preservation & Access

2.5 petabytes and growing

How much is that?
One mp3 =~ 3-4 megabytes

2 petabytes = 2,684,354,560 Mbytes

1.5 million downloads a day
(one of top 350 global sites)

3 storage facilities: San Francisco, Amsterdam, and Alexandria, Egypt

Experience with multiple formats
Phase one (Pilot): Autumn 2004-Autumn 2005


Phase three (OCA+?): May 23, 2008-
Phase One, Sept 2004 – Sept 2005
University of Toronto Collections

- Evaluate technology & workflow
- Scan selections from:
  - Centre for Renaissance and Reformation Studies
  - Centre for 19th century French Studies
  - Pontifical Institute of Mediaeval Studies
  - Circulating collection
  - Records of Early English Drama
    
(© University of Toronto Press)
Phase Two U of T Collections

- Most ranges of LC
- Focus on religion, history, Canadiana (when possible), (some) literature, science
- Mostly English language
- Mostly pre-1923
- Multiple libraries
- Some special collections
- Circulating pre-1923 materials
Phase Three U of T Collections

- Most ranges of LC and other schema
- Most subject areas
- Focus on other languages
- Mostly pre-1923
- Multiple libraries including many external partners
- Some special collections
- Circulating pre-1923 materials
Toronto Post MSN Partners

- Memorial University: Newfoundland Quarterly, Newfoundland materials
- McMaster University: Selections from First World War Collection
- Ryerson University: Selections, *Yellow Book: an illustrated quarterly*
- University of Ottawa: 18th & 19th century faculty selections: history, French, music, history of medicine, jurisprudence, nursing
- Ontario College of Art and Design
- Ontario Council of University Libraries
- Library and Archives of Canada: 0.5M pages, Canadian government publications
- Legislative Assembly of Ontario Library
- Toronto Public Library: local history and genealogy
- University of Alberta: Canadiana
- Tufts University, Boston, USA (Mellon and other grant funds)
- Other: Havergal College, U of T Faculty, Federally funded publisher, test scans for other OCA partners, individual researchers, National Institute of Newman Studies
Scanning Centre Capacity

“Scribe” scanning station capacity

- 500 pages per hour
- 14 hours per day, 5 days per week
- 7,000 pages each day per scribe

“Scribe” Centre capacity

- 161,000 pages per day (23 scribes)
- 805,000 pages per week (23 scribes)
- 2,683 books per week (23 scribes)

If an average book is 300 pages

- 100,000+ books per year
What goes in?

- **Books:**
  - not too big and not too small: 3”x3” to 14.5”x9.5”
  - not too old and not too new
    6% get rejected for hard living; 1922 cut-off

- **MARC metadata**
  - z39.50 is used to fetch MARC data, and so…

- **An identifier to tie book to its metadata**
Some books aren’t mass-digitized

Books scanned: 143,380

Books rejected: 12,424

Rejection rate: 9%
Mass digitization: some Q&A

Duplication

Q: How do we guard against duplication?
A: It might be cheaper just to scan duplicates.

Omissions

Q: What about fold outs, uncut pages, tightly bound books, print running into margins…
A: Mass digitization works *because* it is efficient. A parallel process should handle exception cases.
What comes out?

- **JPEG 2000s:**
  - Raw (~900KB)
  - Crop, skew, light-compensated (~800KB)
- **PDF**
  - Page images with embedded OCR
    - Colour (~100 KB)
    - Black and white (~60KB)
- **Metadata (xml)**
  - Descriptive (catalogue record)
  - Operational (scanning information)
  - Structural (# pages, covers, title page, etc.)
- **OCR (UTF-8)**
  - ABBYY, DjVu
- **Flip book (~35KB)**
  - [Constructive Anatomy](#)
What do we get? For each book scanned, we receive

- Acrobat file of page images… and page text (OCR)

THE MEN
WHAT a reign! Was history ever better dressed?
1 never waver between the cardboard figures of the great Elizabethan time and this reign as a monument to lavish display, but if any time should beat this for quaintness, colour, and variety, it is the time of Henry VIII.

Look at the scenes and characters to be dressed: John, Duke of Bedford, the Protector, Joan of Arc, Jack Cade, a hundred other people; Crevant, Verneuil, Orleans, London 176
Comparison: scanned & born-digital

<table>
<thead>
<tr>
<th>Scanned from print</th>
<th>Born-digital</th>
</tr>
</thead>
<tbody>
<tr>
<td>Page images</td>
<td>E-text</td>
</tr>
<tr>
<td>Search uncorrected OCR</td>
<td>Search text</td>
</tr>
<tr>
<td>TOC, title page, index are marked</td>
<td>Can be highly segmented, linked</td>
</tr>
<tr>
<td>Literature, history, …</td>
<td>STM, social sciences, reference, …</td>
</tr>
</tbody>
</table>

---

**DOCUMENTARY HISTORY.**

dence during the summer, and learning that it was probable that they would *winter* in our neighbourhood, they arrived here, without having lost the hope of seeing those for whom they had been particularly sent. Their hopes have not been frustrated, for these savages shortly afterwards arrived here, to the number of about two hundred and fifty souls.”—Page 89.
Open Content for UTL

Internet Archive also provides XML metadata & images

<title>The drama of the forests: romance and adventure</title>
<description>Northwest, Canadian—description and travel</description>
<language>English</language>
<page_height>9.5</page_height>
<page_width>5</page_width>
Constructing the online book

Scan Center

Assign unique id

Get metadata via z39.50

Scan book

Perform QA

Upload scans

UTL script:
- barcode in
- identifier out
- tracks scan decision

Internet Archive

Make book available online

Approve book

Create derivatives
IA scanning for other institutions

- Ship books
  - Send marc record file with books
  - Request marc records from another source
    - Library and Archives Canada, Library of Congress, other library catalogues...
  - Arrange z39.50 access for IA
    - OCAD

- Sponsor books
  - Select area of interest for scanning
  - Sponsor scanning
    - Tufts Perseus collection, Library and Archives Canada
# How is it used? The current top 10 list

<table>
<thead>
<tr>
<th>Uses</th>
<th>Year</th>
<th>Work</th>
</tr>
</thead>
<tbody>
<tr>
<td>116,574</td>
<td>1475</td>
<td>St. Augustine. <em>De civitate Dei</em></td>
</tr>
<tr>
<td>22,911</td>
<td>1920</td>
<td>Bridgman, GB. <em>Constructive anatomy</em></td>
</tr>
<tr>
<td>13,120</td>
<td>1592</td>
<td>Colonna, Francesco, d. 1527. <em>Hypnerotomachia</em></td>
</tr>
<tr>
<td>10,066</td>
<td>1925</td>
<td>Powell, John Benjamin. <em>Who's who in China; containing the pictures and biographies of China's best known.</em></td>
</tr>
<tr>
<td>9,236</td>
<td>1904</td>
<td>Gallonio, Antonio, d. 1605. <em>Traité des instruments de martyre et des divers modes de supplice employés par les païens.</em></td>
</tr>
<tr>
<td>9,203</td>
<td>1910</td>
<td>Schopenhauer, Arthur. <em>The world as will and idea</em></td>
</tr>
<tr>
<td>6,676</td>
<td>1831</td>
<td>Shelley, Mary Wollstonecraft. <em>Frankenstein, or, The modern Prometheus</em></td>
</tr>
<tr>
<td>6,584</td>
<td>1884</td>
<td>Abbott, Edwin Abbott. <em>Flatland: a romance of many dimensions</em></td>
</tr>
</tbody>
</table>
### How is it used? General statistics

<table>
<thead>
<tr>
<th></th>
<th>Scanned books</th>
<th>Print books</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Avg use</td>
<td>Min use</td>
</tr>
<tr>
<td>Top 100</td>
<td>5590</td>
<td>2821</td>
</tr>
<tr>
<td>Top 1,000</td>
<td>1943</td>
<td>1096</td>
</tr>
<tr>
<td>Top 10,000</td>
<td>699</td>
<td>371</td>
</tr>
<tr>
<td>Top 50,000</td>
<td>307</td>
<td>136</td>
</tr>
</tbody>
</table>
How is it used?

General print vs online comparison

![Chart showing comparison between scanned books and print books across different top categories.](chart.png)
Which classes are used more?
Impact of schema
Comparison by class
Online or print?

- Readers prefer print:
  - LC classes B,C,D,E,F (religion, philosophy, history)

- Readers prefer online:
  - LC classes H,J,Q,T,Z (social sciences, science, technology, library science & bibliography)

  Christianson, M. and Aucoin, M., 2005

- “Even though the use of electronic sources and online reading habits vary by discipline, the frequency of printing out electronic documents is surprisingly similar across all disciplines.”

  Liu, Z. 2006
Comparison

B class

More popular

- Judaism
- Buddhism
- Bible
- Doctrinal & practical theology
- Christian denominations
Comparison
P class

Dictionaries
- 0.6% total scans
- 1.5% top titles
- 3.1% downloads
Popular languages

- English  -11%
- French  35%
- Latin   49%
- Italian 56%
- German  81%

[Bar chart showing popularity of languages]
How do people read?

Intentional reading

- Attentive, sustained, linear reading of text
- Heavily influenced by printed-book culture
- Dominant in classical and scholarly literature

Functional reading

- Manipulating different content types
- Web browsing, text database searching
- Most screen reading is functional

# Reading online

<table>
<thead>
<tr>
<th>Percentage of time spent on</th>
<th>Increasing</th>
<th>Decreasing</th>
<th>No change</th>
<th>Don’t know</th>
</tr>
</thead>
<tbody>
<tr>
<td>Browsing and scanning</td>
<td>80.5</td>
<td>11.5</td>
<td>8.0</td>
<td>0</td>
</tr>
<tr>
<td>Keyword spotting</td>
<td>72.6</td>
<td>2.7</td>
<td>16.0</td>
<td>8.8</td>
</tr>
<tr>
<td>One-time reading</td>
<td>56.6</td>
<td>8.0</td>
<td>29.2</td>
<td>6.2</td>
</tr>
<tr>
<td>Reading selectively</td>
<td>77.9</td>
<td>2.7</td>
<td>16.8</td>
<td>2.7</td>
</tr>
<tr>
<td>Non-linear reading</td>
<td>82.3</td>
<td>0</td>
<td>15.9</td>
<td>1.8</td>
</tr>
<tr>
<td>Sustained attention</td>
<td>15.9</td>
<td>49.6</td>
<td>29.2</td>
<td>5.3</td>
</tr>
<tr>
<td>In-depth reading</td>
<td>26.6</td>
<td>45.1</td>
<td>23.0</td>
<td>5.3</td>
</tr>
<tr>
<td>Concentrated reading</td>
<td>21.2</td>
<td>44.2</td>
<td>26.5</td>
<td>8.0</td>
</tr>
</tbody>
</table>

*Note: Figures given are percentages; figures may not add up to 100 percent because of rounding*
How do people know what they’ve read?

[A] strong relationship…exists between the sensory motor representation of the user and his/her treatment of the information content of the paper book or e-book…

Because an electronic book is functionally closer to a computer than a traditional book […] it does not provide the external indicators to memory that the classical book does…

Morneau et al, 2005
Delivering the book to the user

<table>
<thead>
<tr>
<th>User tasks</th>
<th>Printed books</th>
<th>Online books</th>
</tr>
</thead>
<tbody>
<tr>
<td>Make use copy</td>
<td>Make discovery surrogate</td>
<td>Make discovery surrogate</td>
</tr>
<tr>
<td>Make discovery surrogate</td>
<td>Search surrogates, choose candidates</td>
<td>Search surrogates, choose candidates</td>
</tr>
<tr>
<td>Search surrogates, choose candidates</td>
<td>Examine candidates</td>
<td>Examine candidates</td>
</tr>
<tr>
<td>Examine candidates</td>
<td>Browse more candidates</td>
<td>???</td>
</tr>
<tr>
<td>Browse more candidates</td>
<td>Choose material</td>
<td>Choose material</td>
</tr>
<tr>
<td>Choose material</td>
<td>Make use copy</td>
<td>Make use copy</td>
</tr>
</tbody>
</table>
Delivering the book to the user

- **Discovery:**
  - target TOC and index for indexing & correction

- **Use:**
  - support production of good print copies for use
  - granular linking
  - browse functions
  - Highlighting, bookmarking…
Odd ideas

- Evidence Based Librarianship
- One size doesn’t fit all (reproduction, artifact)
- Library processes focused on containers
- Libraries have cultural heritage role; we must avoid content being locked in proprietary formats that are managed by businesses
- (ebook reader formats – epub, kindle, google, ???)
Odd ideas cont

- Donald Waters, Managing Digital Assets in Higher Education: An Overview of Strategic Issues. ARL 244 Feb 2006:
  
  - p. 1 The touchstone question [...] must be: How well does this resource, or that system or feature, advance scholarship?
  
  - p. 5: The central issue is whether scholars can advance knowledge in ways that were not previously possible.
How are we using it?

- Scholar's Portal E-book platform
  - integrates licensed and free content
  - pdf-like reader
- open access to IA content
Scholars’ Portal

- 21 Ontario universities, 392,000 FTEs
- Locally loaded content:
  - Now:
    - 14 million journal articles in 8000 journals
    - 200 million citations in 110 abstracts & indexes
    - ODESI (data sets, GIS data)
  - Next:
    - 300,000 e-books
- Locally operated services:
  - ILL (VDX), ERM (SFX/Verde),
    Bibliographic Management Software (RefWorks)
Why load it locally?

- **Safekeeping**
  - Lots of copies keep stuff safe

- **Discovery**
  - Integration with licensed books
  - Integration with non-book content
  - Local subject specialization

- **Services**
  - LMS, bibliographic management, …
Next steps

- Print on demand
- Scan on demand
- Enriched structural metadata to improve discovery
  - Current structural metadata: pagination, covers, title page, copyright page
  - Desired structural metadata: Table of contents, index, images, maps
Discovering the special collections

Local: discovery layer
National: canadiana.org

Doctor James Bertram Collip, Edmonton, Alberta, 1892-?
 WORKES WITH DOCTOR BEST AND DOCTOR BANTING ON DISCOVERY ON INSULIN.
Glenbow Archives

Chart of insulin assays [1923]
UNIVERSITY OF TORONTO LIBRARIES

Insulin, diabetes, and rewards for discoveries
NATURE

University of Toronto Libraries

Locations

Burlington (1)
Canada - Alberta (1)
London (1)

Item Types

Archival finding aid (2) document (768)
Discovery layer

- Faceted search using Endeca
- stretch “catalogue” to include:
  - metadata for all books, not just our books
  - web site
  - A&Is
  - Full text journals
  - Full text books
How can libraries use it?

- link to Internet Archive
  - repository of 1 million online books
- add marc records to catalogue
  - metadata integrated with local collection
- add full text books to collection
  - full text search
References


References


