

---

# Mass Digitization

---

Sian Meikle

University of Toronto Libraries

---

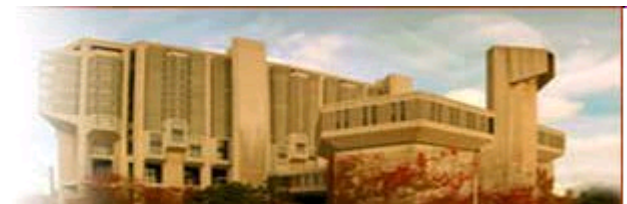
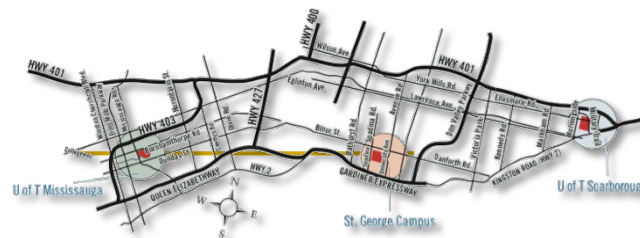
# Outline

- 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](http://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  $\approx$  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  
Francis

c

o, Amsterdam, and Alexandria, Egypt



Experience with multiple formats

---

# University of Toronto Mass Digitization

- Phase one (Pilot):  
Autumn 2004-Autumn 2005
  - Phase two (MSN/OCA):  
Autumn 2005-May 22, 2008
  - 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

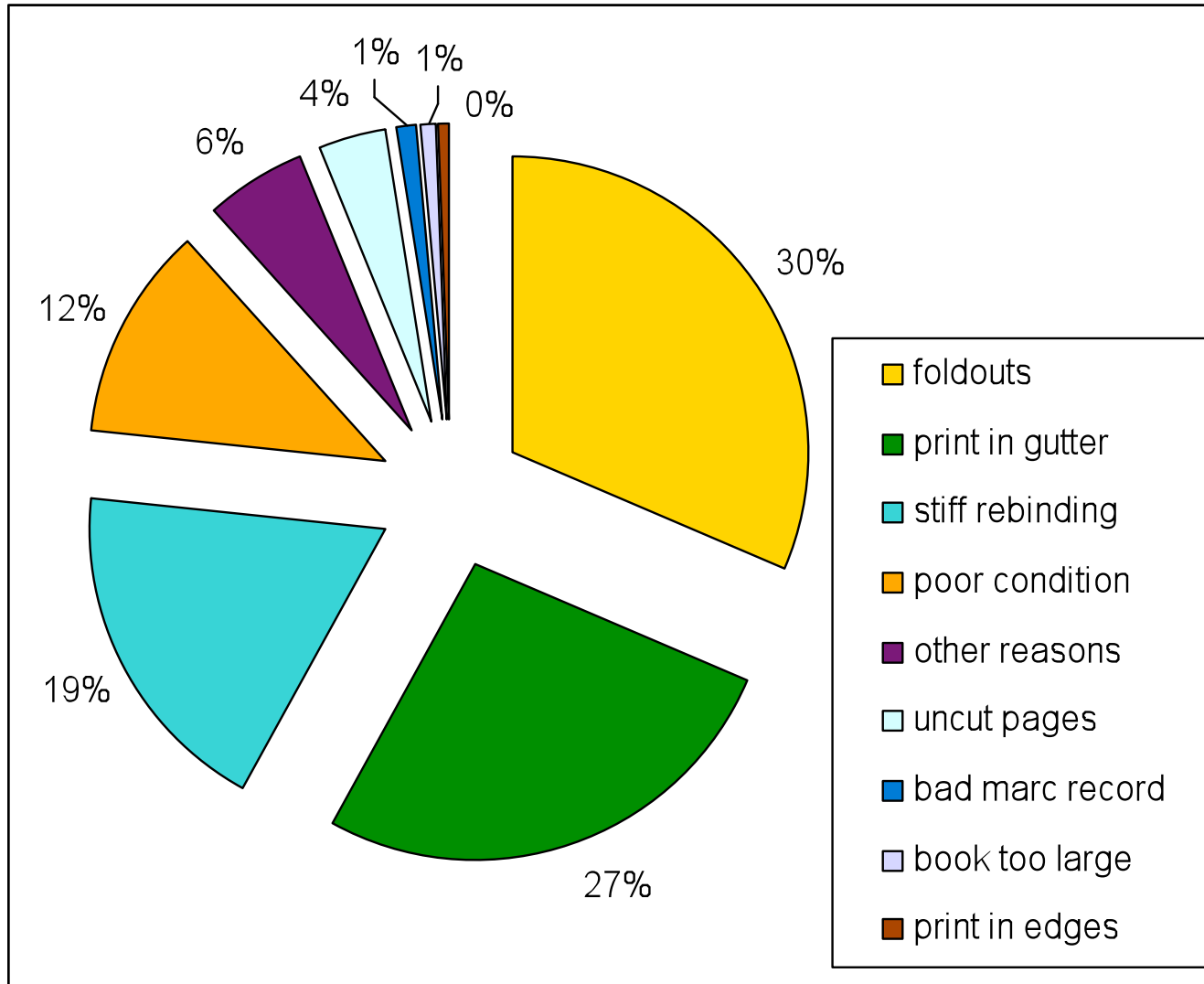
- 100,000+ books per year
-

# 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...

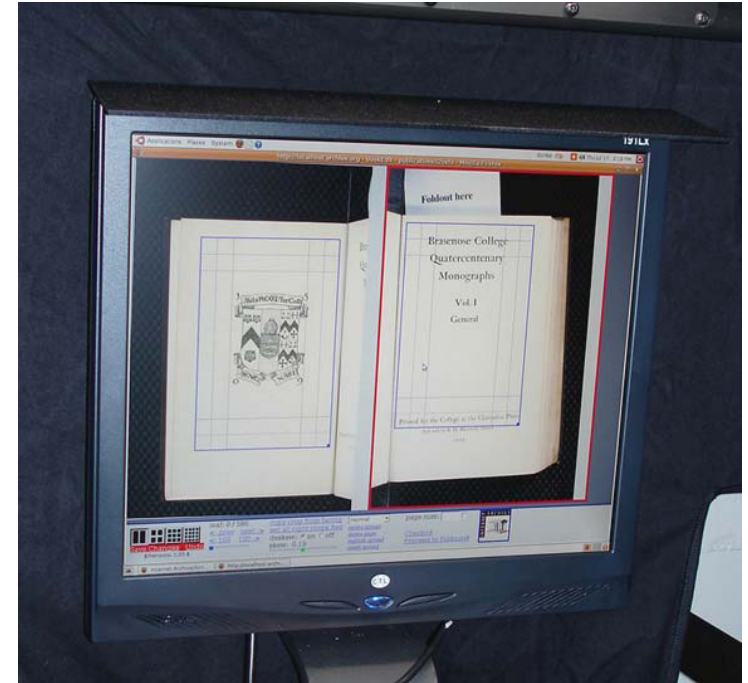
A: Mass digitization works *because* it is efficient.  
A parallel process should handle exception cases.

---



# What comes out?

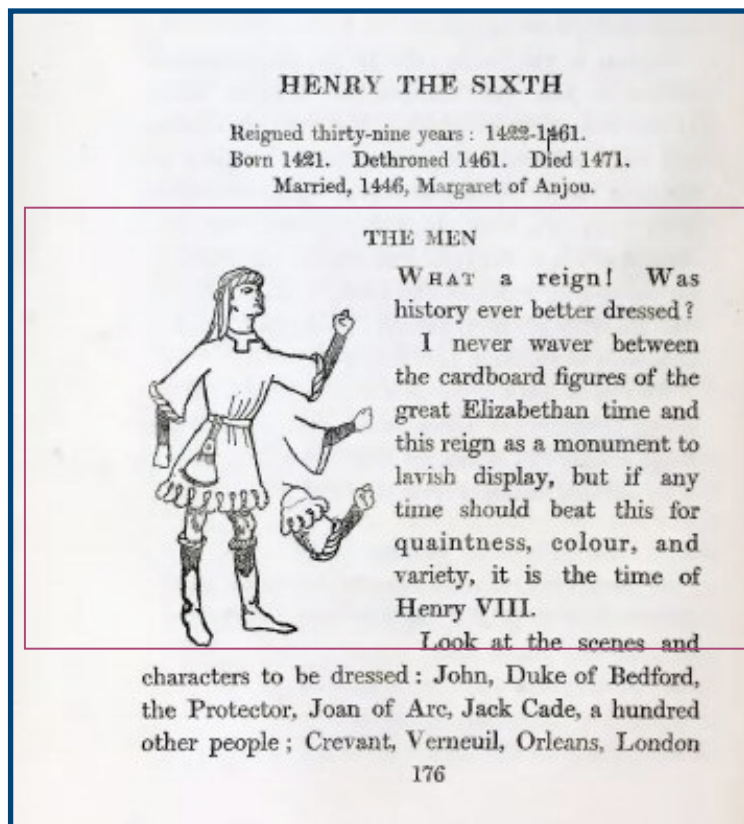
- JPEG 2000s:
  - Raw (~900KB)
  - crop, skew, light-compensated (~800KB)
- PDF
  - page images with embedded OCR
    - colour (~100 KB)
    - black and white (~60KB)
- Metadata (xml)
  - Descriptive (catalogue record)
  - Operational (scanning information)
  - Structural (# pages, covers, title page, etc.)
- OCR (UTF-8)
  - ABBYY, DjVu
- Flip book (~35KB)
  - Constructive Anatomy



# Open Content for UTL

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 ?  
I 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.

# Comparison: scanned & born-digital

Scanned from print	Born-digital
Page images	E-text
Search uncorrected OCR	Search text
TOC, title page, index are marked	Can be highly segmented, linked
Literature, history, ...	STM, social sciences, reference, ...

14

## 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.

**Samter's Immunological Diseases**  
> Table of Contents > SECTION II - EFFECTOR MECHANISMS: INNATE AND ADAPTIVE > 18 - T-LYMPHOCYTE EFFECTOR ACTIVITY  
T-LYMPHOCYTE EFFECTOR ACTIVITY

Search:    ☒ Check Spelling

[Back](#) [Save](#) | [Print Preview](#) | [Email](#) | [Email Jumpstart](#)  
[What's New in Books](#)

**18**  
**T-LYMPHOCYTE EFFECTOR ACTIVITY**

Thomas J. Braciale M.D., Ph.D.  
Steven M. Varga Ph.D.

The vertebrate immune system has evolved to counter infestation and invasion by foreign prokaryotic and eukaryotic microorganisms. The immune system is composed of two distinct but related components, the innate and the adaptive immune systems. The *innate immune system*

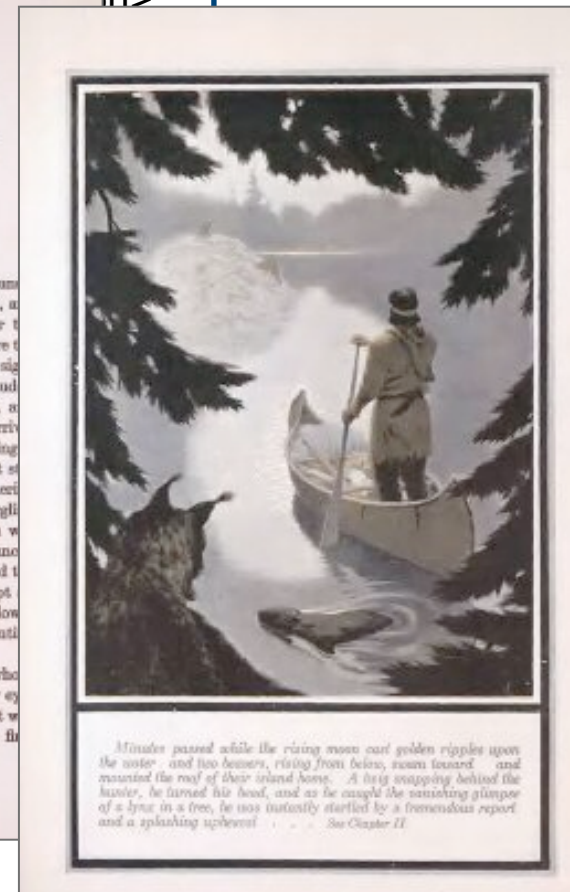
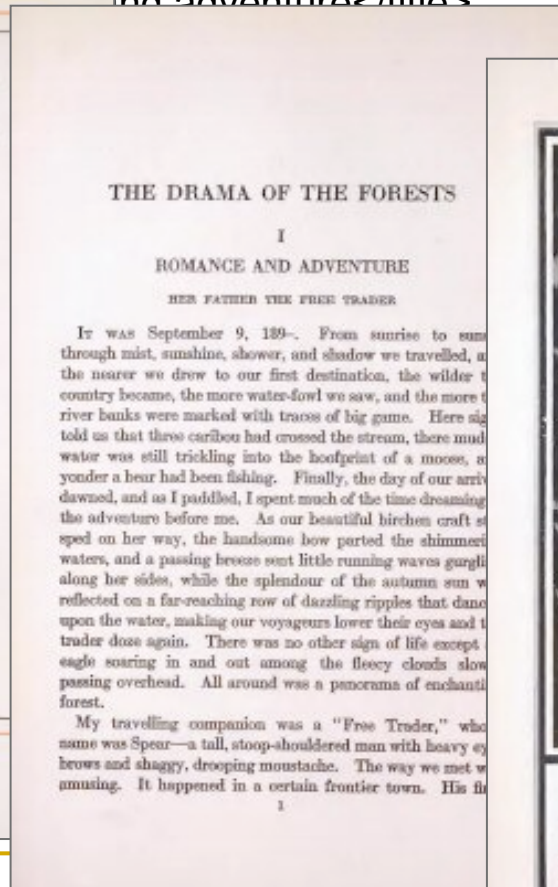
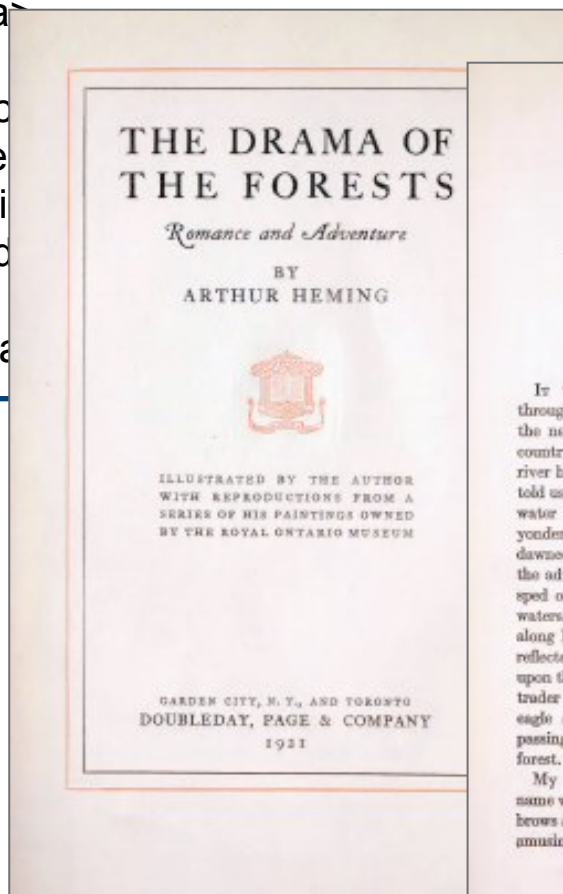
TABLE OF CONTENTS

- [+] SECTION I - RECOGNITION: INNATE AND ADAPTIVE IMMUNITY
- [+] SECTION II - EFFECTOR MECHANISMS: INNATE AND ADAPTIVE
  - [+] 18 - T-LYMPHOCYTE EFFECTOR ACTIVITY
    - [+] 18.1 - EVENTS IN THE INDUCTION AND EXPRESSION OF T-LYMPHOCYTE EFFECTOR ACTIVITY

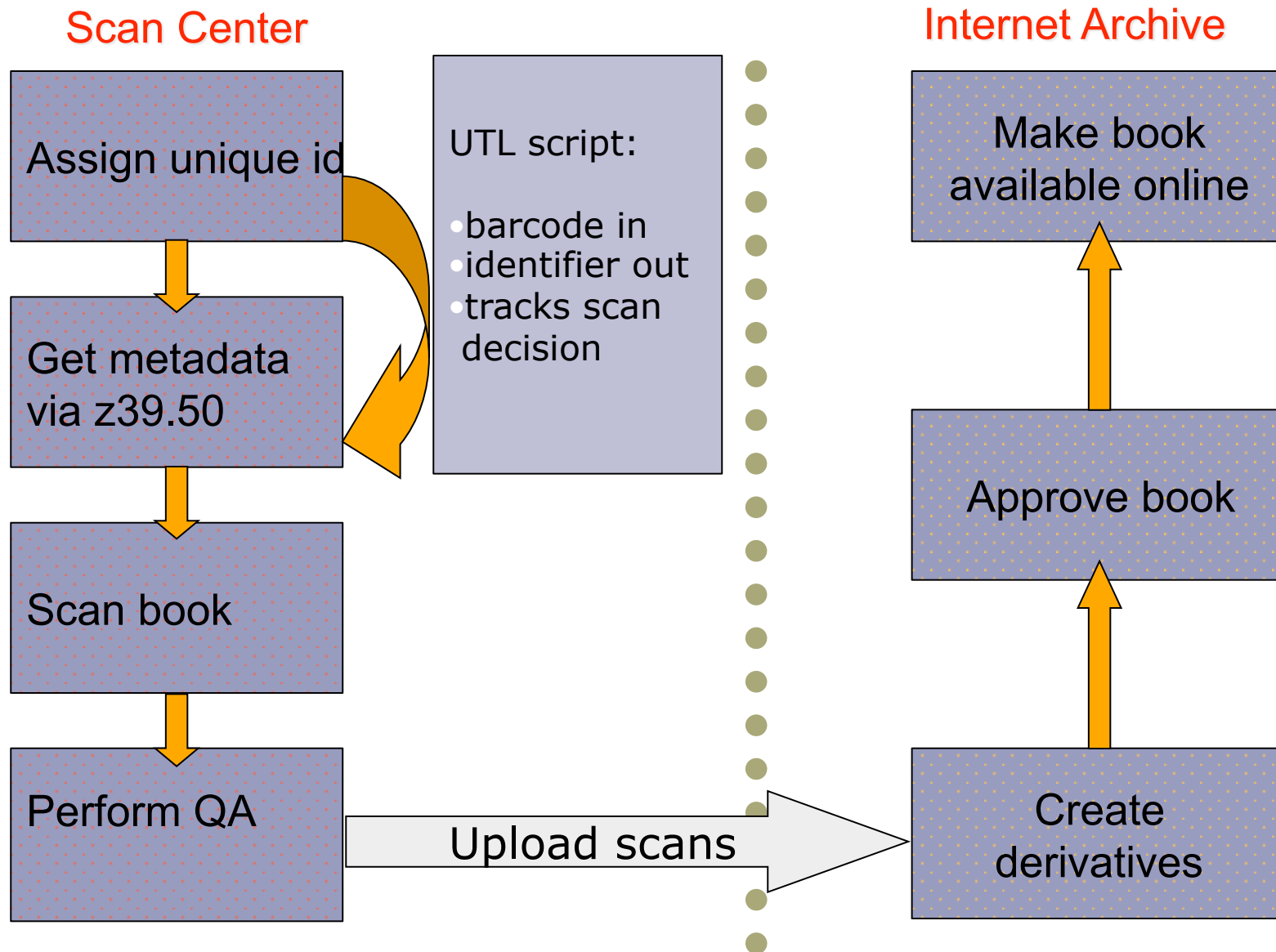
# Open Content for UTL

Internet Archive also provides XML metadata & images

```
<metadata>  
<title>The Romance and adventures</title>  
<description>  
<language>  
<page_height>  
<page_width>  
[...]  
</metadata>
```



# Constructing the online book



# 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

Uses	Year	Work
116,574	1475	St. Augustine. <i>De civitate Dei</i>
22,911	1920	Bridgman, GB. <i>Constructive anatomy</i>
13,120	1592	Colonna, Francesco, d. 1527. <i>Hypnerotomachia</i>
10,066	1925	Powell, John Benjamin. <i>Who's who in China; containing the pictures and biographies of China's best known..</i>
9,236	1904	Gallonio, Antonio, d. 1605. <i>Traité des instruments de martyre et des divers modes de supplice employés par les païens ...</i>
9,203	1910	Schopenhauer, Arthur. <i>The world as will and idea</i>
8,065	1894	Mosselman, Gustave. <i>Manual of veterinary microbiology.</i>
7,367	1910	Descartes, René, et al. <i>French and English philosophers: Descartes, Rousseau, Voltaire, Hobbes.</i>
6,676	1831	Shelley, Mary Wollstonecraft. <i>Frankenstein, or, The modern Prometheus</i>
6,584	1884	Abbott, Edwin Abbott. <i>Flatland : a romance of many dimensions</i>

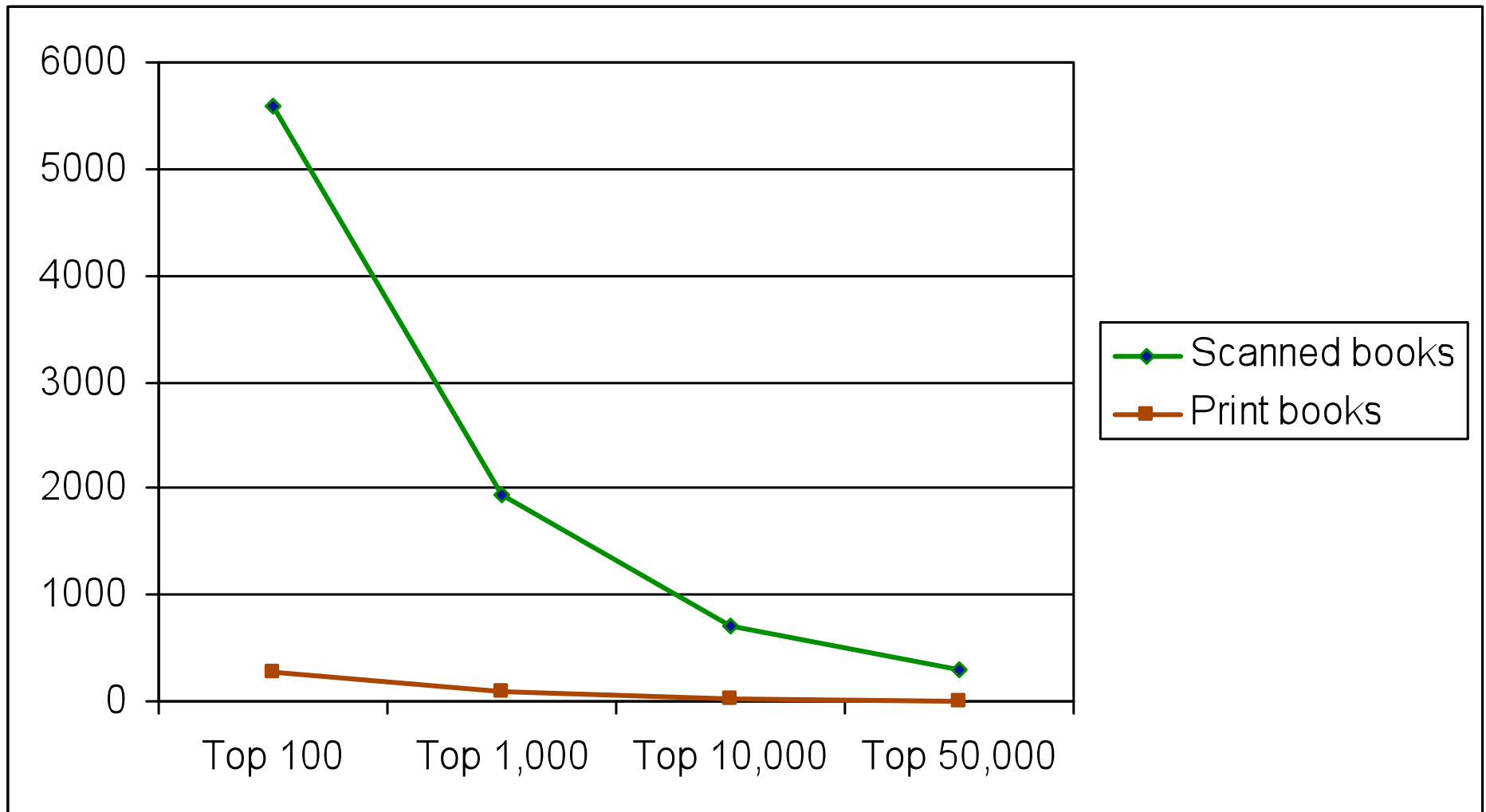


# How is it used? General statistics

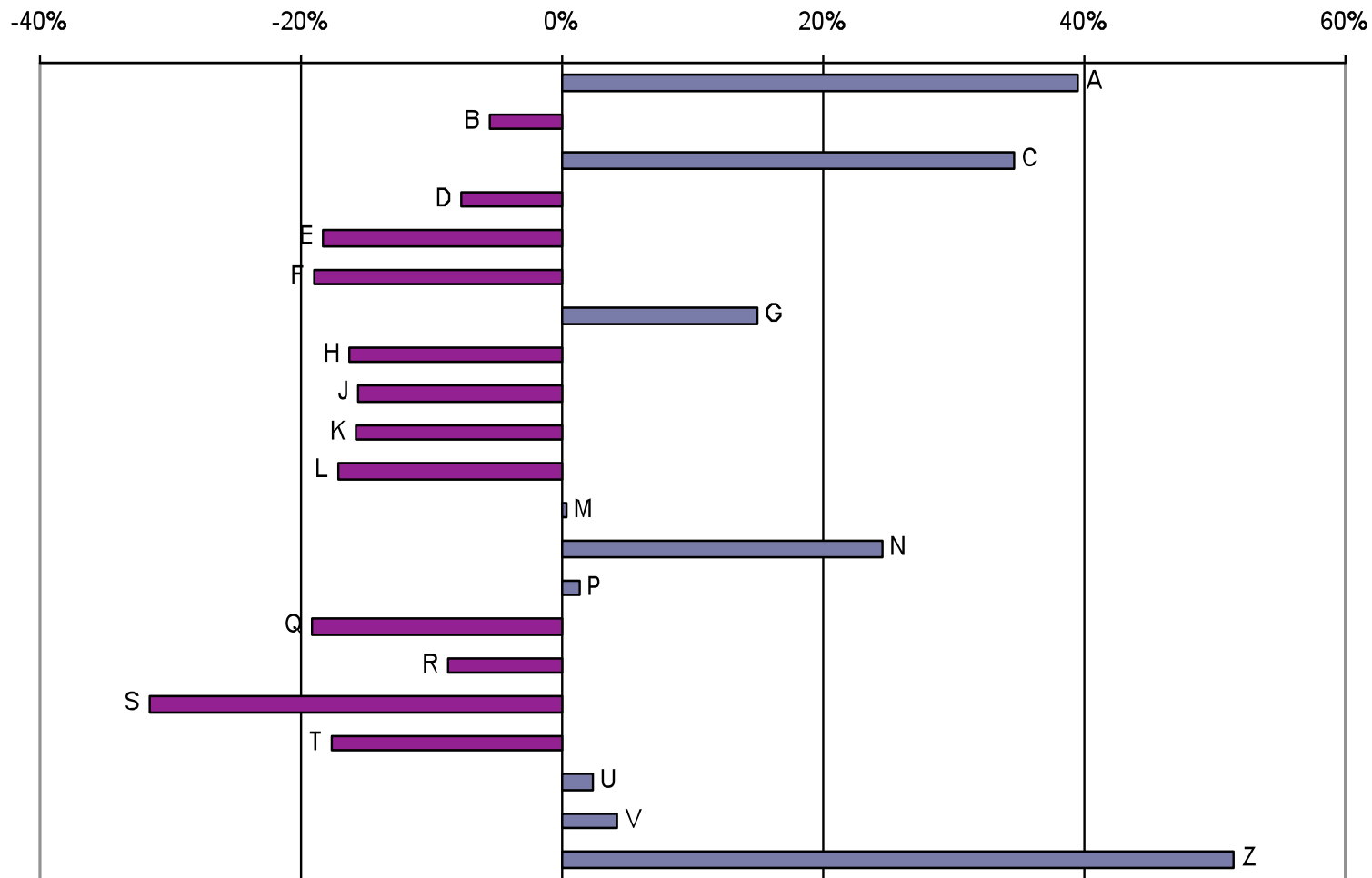
	Scanned books		Print books	
	Avg use	Min use	Avg use	Min use
Top 100	5590	2821	273	163
Top 1,000	1943	1096	98	52
Top 10,000	699	371	30	12
Top 50,000	307	136	10	2

# How is it used?

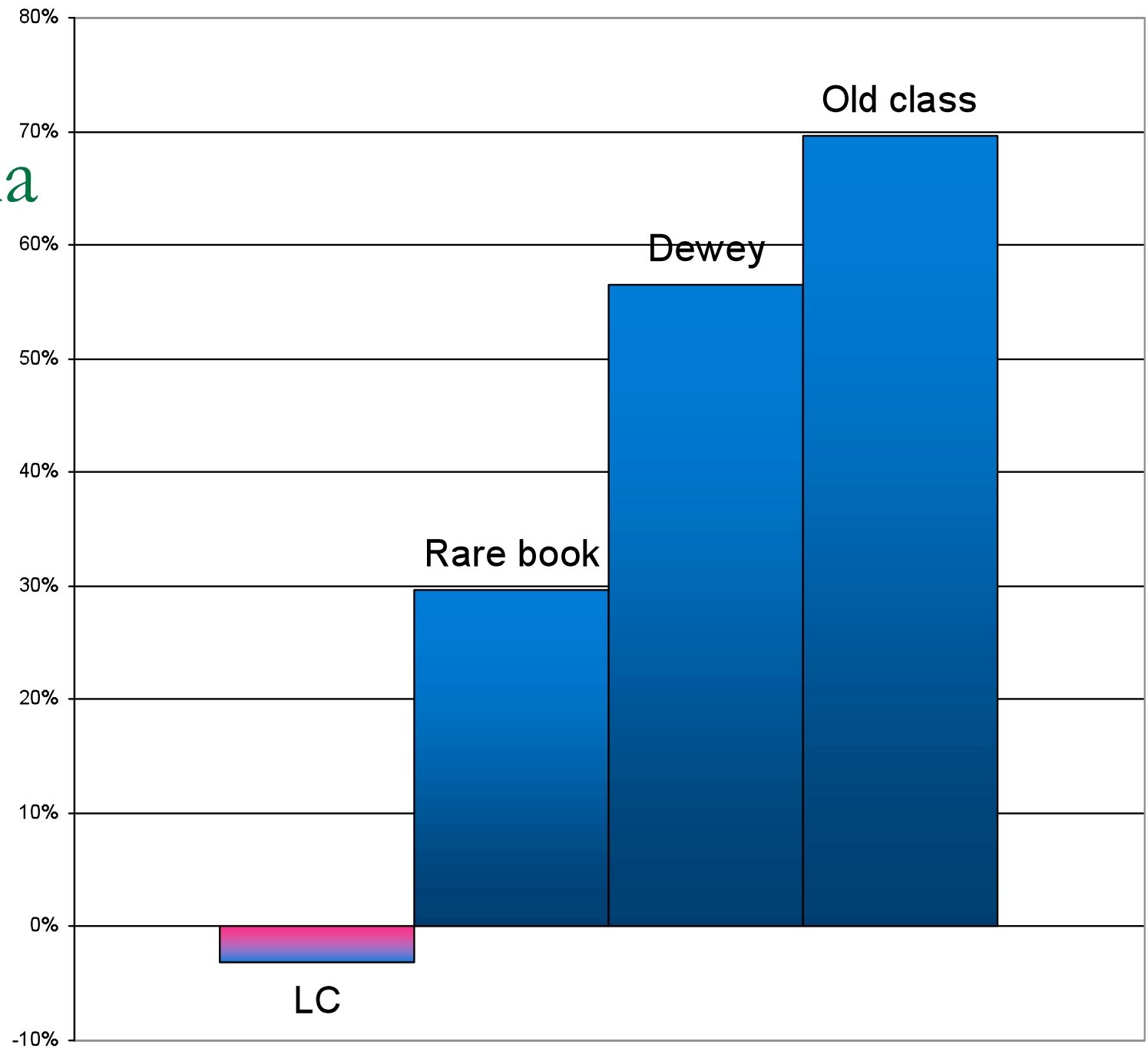
## General print vs online comparison



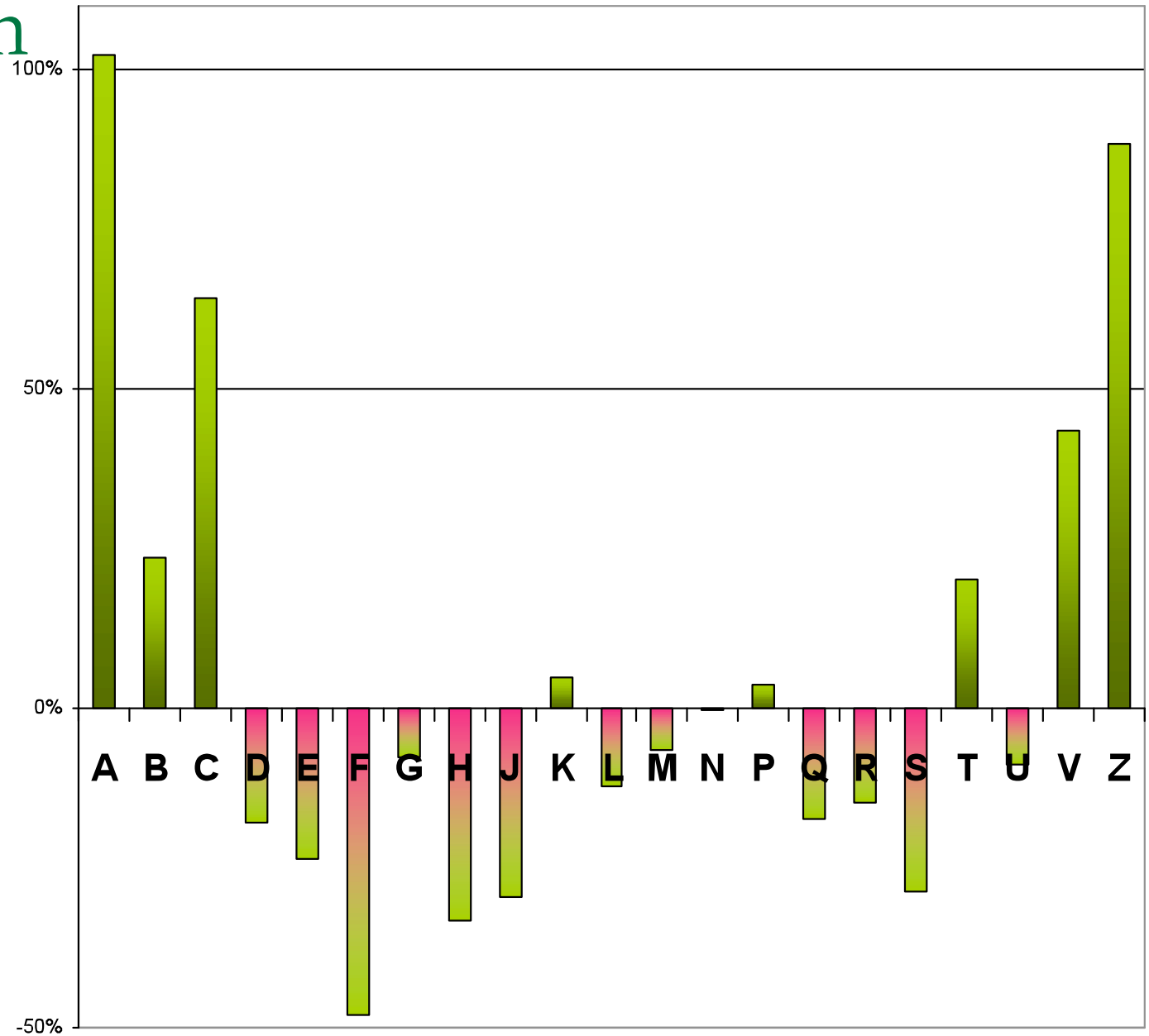
# Which classes are used more?



# 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.”

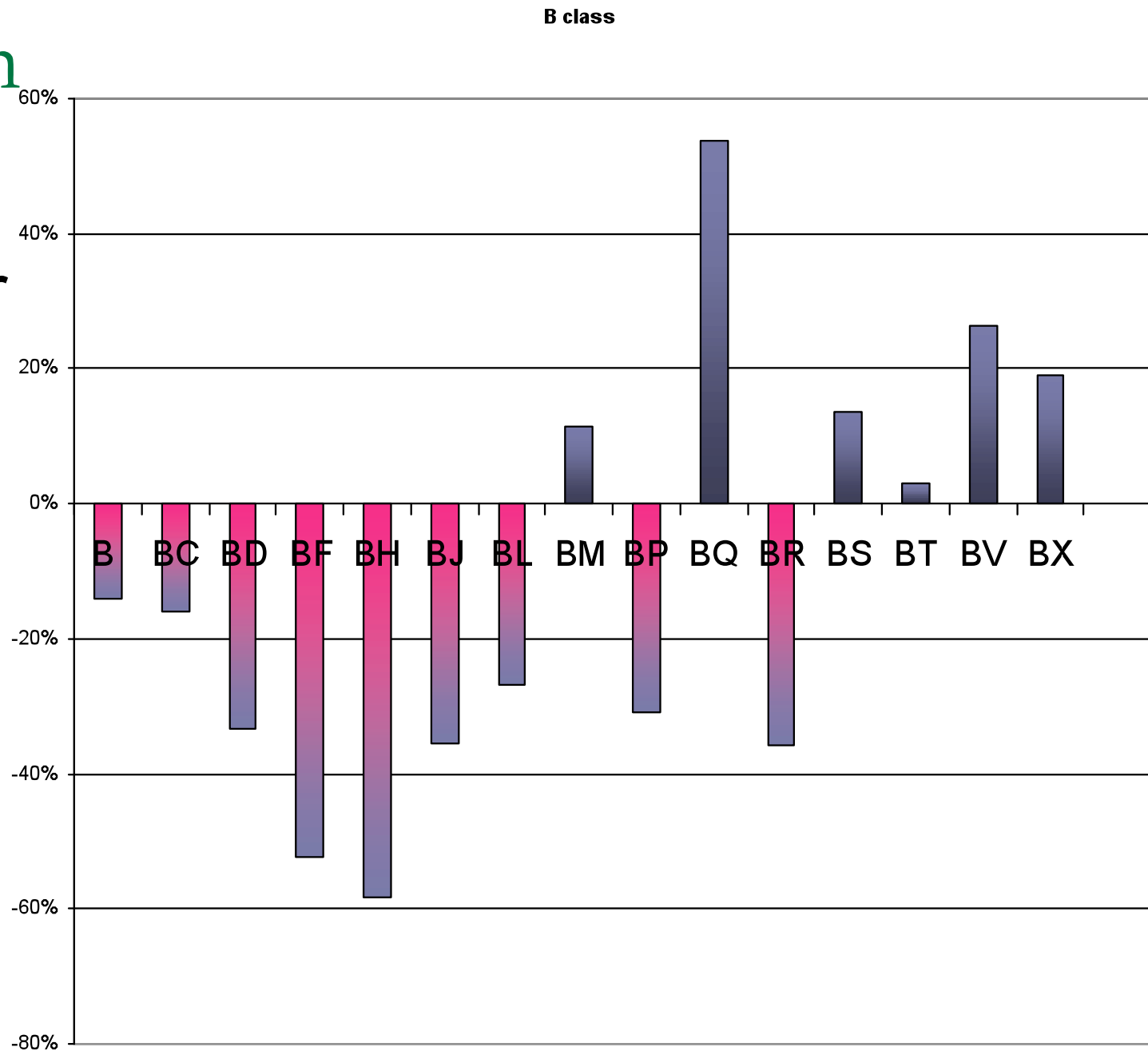
Liu, Z. 2006

# Comparison

## B class

More popular

- Judaism
- Buddhism
- Bible
- Doctrinal & practical theology
- Christian denominations



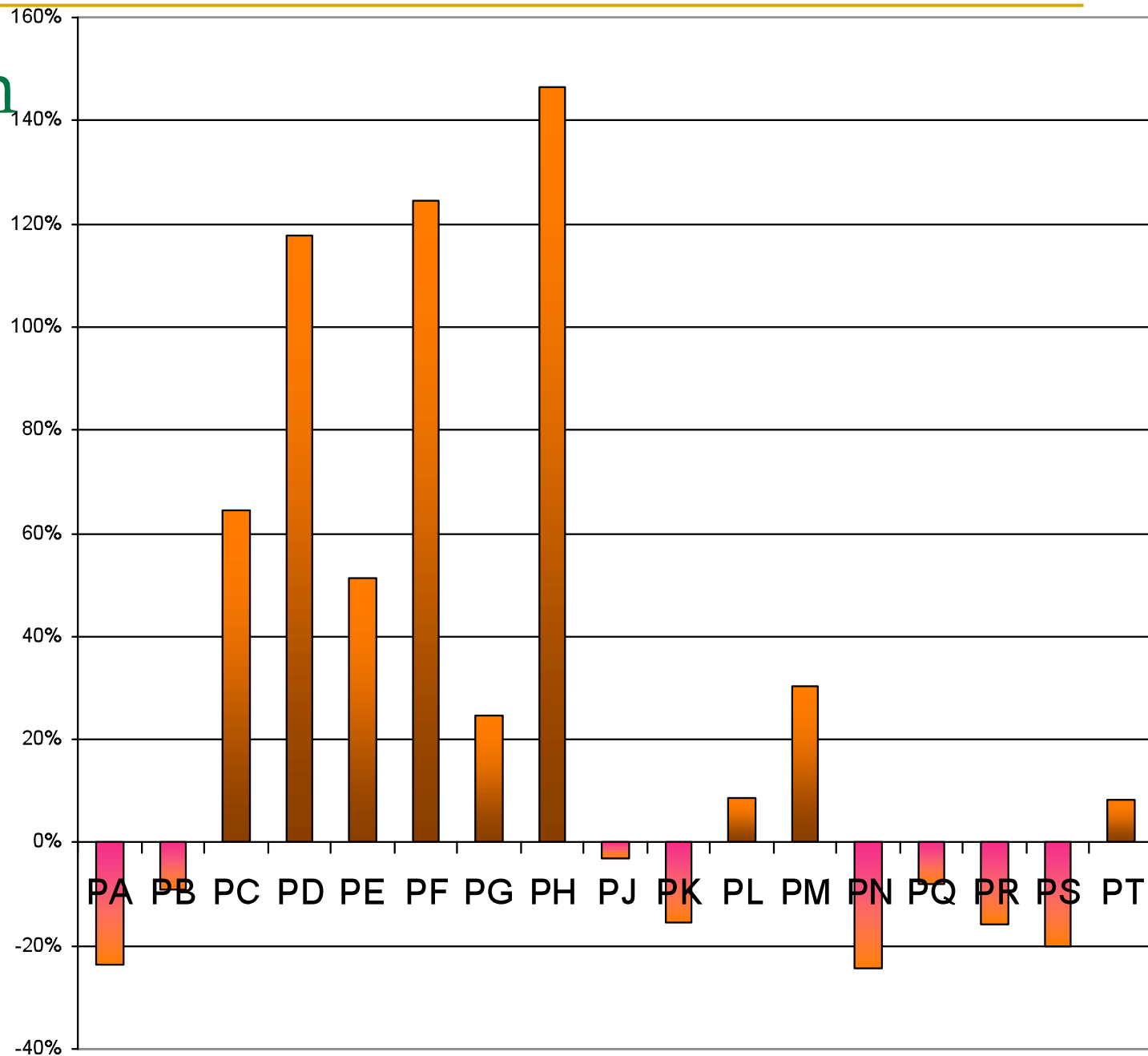


# Comparison

## P class

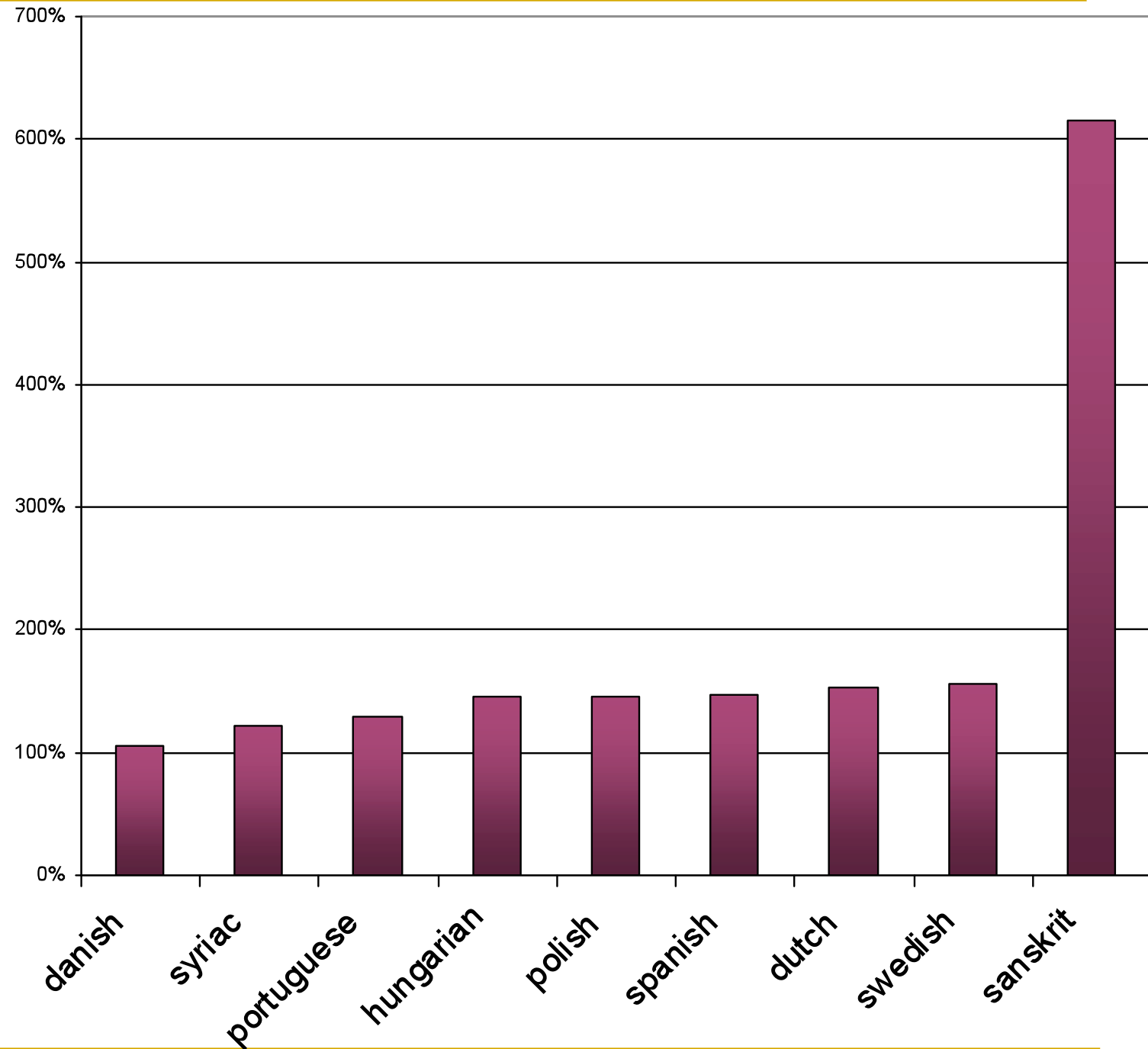
### Dictionaries

- 0.6%  
total scans
- 1.5%  
top titles
- 3.1%  
downloads



# Popular languages

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



# 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

Hillesund, T., & Noring, J. E. (2006)

# Reading online

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

**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...

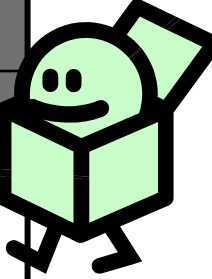

Because an electronic book is functionally closer to a computer than a traditional book [...] it does not provide the external indicators to memory that the classical book does...

Morineau et al, 2005

---

# Delivering the book to the user

User tasks	Printed books	Online books
	Make use copy	
	Make discovery surrogate	Make discovery surrogate
	Search surrogates, choose candidates	Search surrogates, choose candidates
	Examine candidates	Examine candidates
	Browse more candidates	???
	Choose material	Choose material
		Make use copy

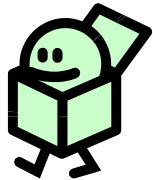


# Delivering the book to the user

## ■ 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, ???)
-

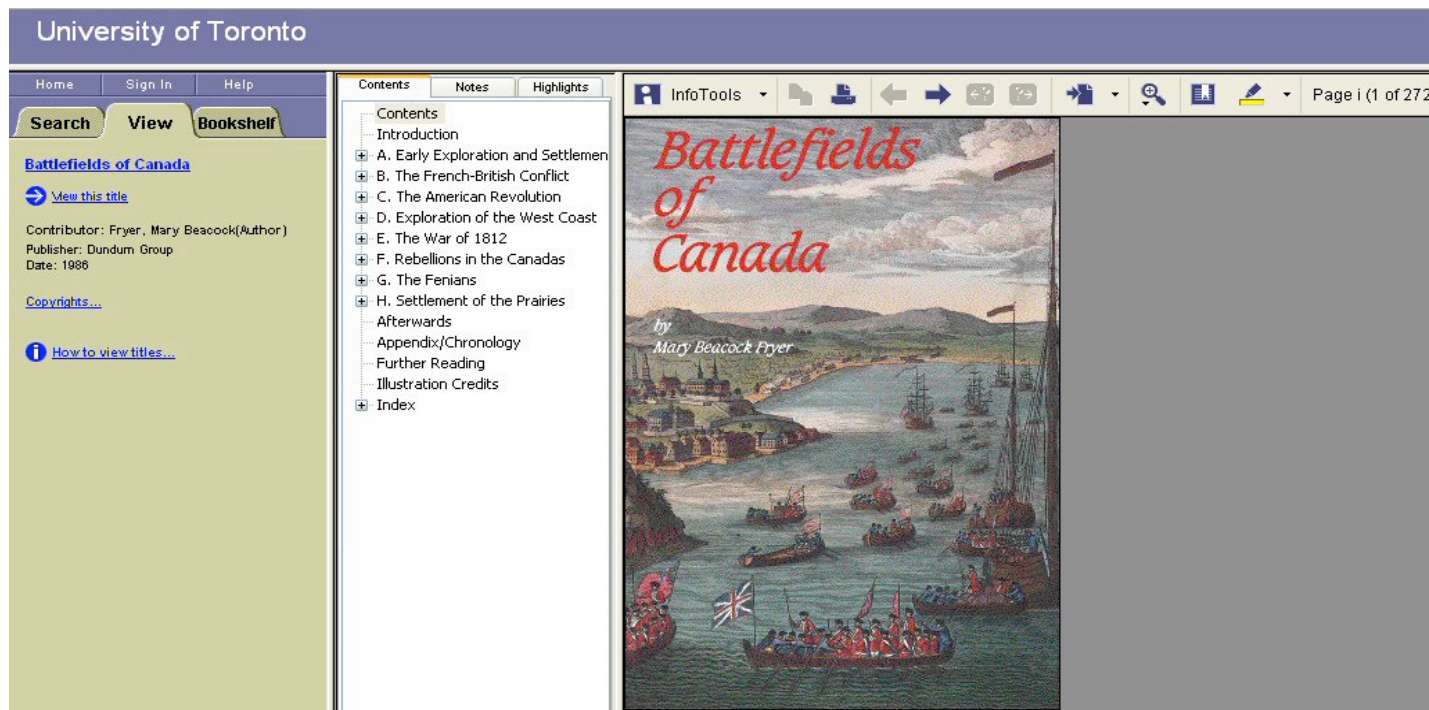
---

## 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:

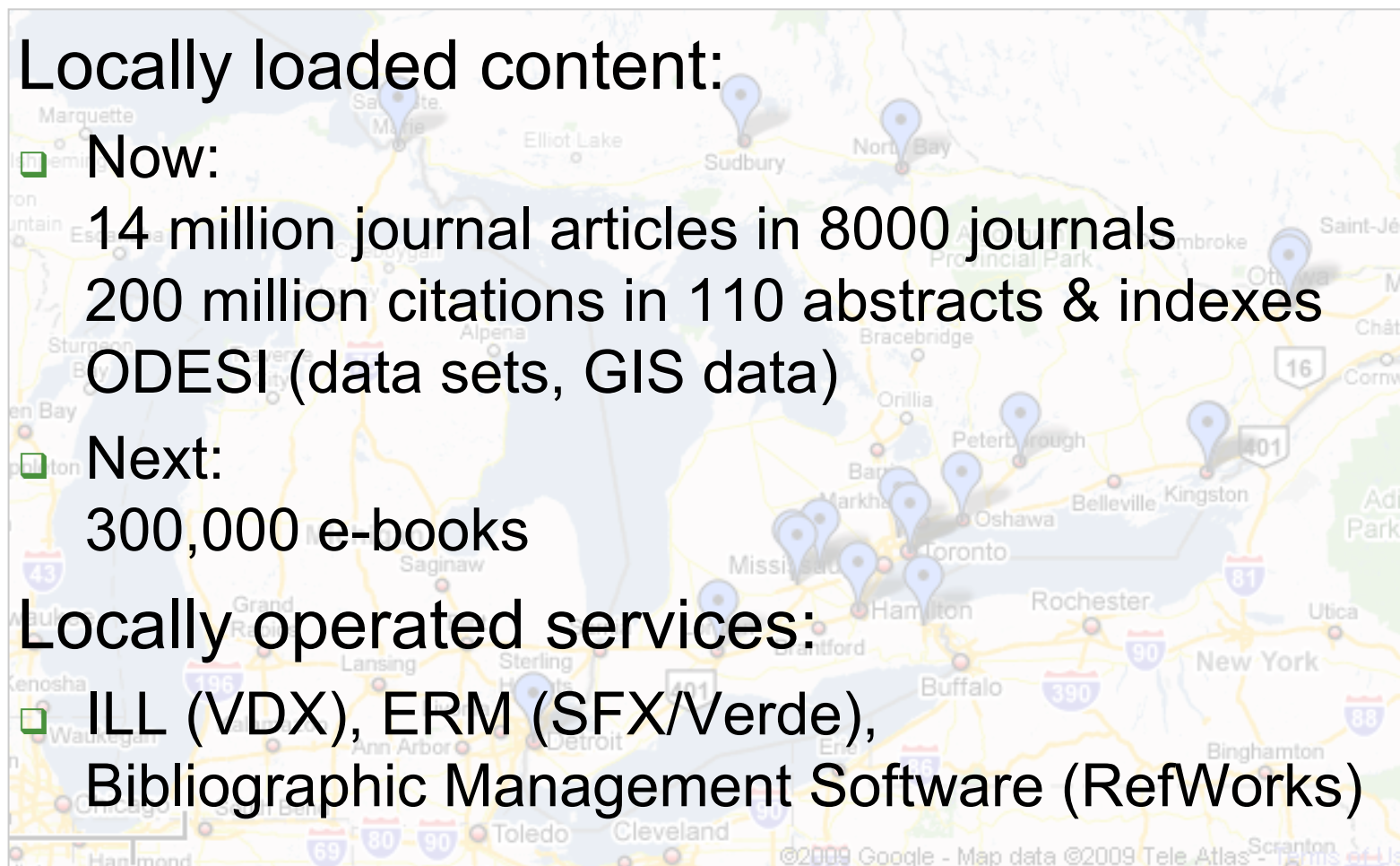
- 14 million journal articles in 8000 journals
    - 200 million citations in 110 abstracts & indexes
    - ODESI (data sets, GIS data)

- 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
-

[SEARCH](#) [FAQ](#) [HELP](#) [EN FRANÇAIS](#)

We found **860** matching items.

**Keywords:** insulin

Did you mean: [Insulin?](#)

Page 1 of 22 [1](#) [2](#) [3](#) ... [21](#) [22](#) [Next →](#)



[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. Insulin Committee, Chart of insulin assays [1923], 1923. 1 p. ; 22 x 34 cm.Item is a hand made form filled in by hand. The chart records the blood sugar of rabbits injected with insulin 4 different insulin samples over a period of 5 hours.University of Toronto Archives. A1980-0027 Box 2Title based on content of chart.This chart records some of the...

**University of Toronto Libraries**



[Insulin, diabetes, and rewards for discoveriesNature](#)






 Bayliss, William Maddock, Sir, 1860-1924, Insulin, diabetes, and rewards for discoveries, 1923. p. 188 - 191 ; 26 x 19 cm.Article gives a general overview of insulin and discusses the problems and merits of patenting methods of manufacturing life-saving drugs.Gerstein Science Information Centre. Stacks.Article. In Nature, vol 3, no. 2780 (Feb. 10, 1923)

**University of Toronto Libraries**

## SEARCH WITHIN THESE RESULTS

 [Go](#)

## MEDIA TYPES

-  Audio (0)
-  Collection (0)
-  Image (44)
-  Text (799)
-  Video (0)

## CONTRIBUTORS

- [University of Toronto Libraries \(821\)](#)
- [Empire Club of Canada \(29\)](#)
- [Oakville Images \(6\)](#)
- [ARCHEION \(2\)](#)
- [Burlington Historical Society \(1\)](#)
- [Centre for Addiction and Mental Health, Archives \(1\)](#)

[\[+\] SEE THE REST](#)

## LOCATIONS

- [Burlington \(1\)](#)
- [Canada - Alberta \(1\)](#)
- [London \(1\)](#)

Location unidentified: 857

## 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

- Bengtson, Jonathan, and Robert Miller. "Canadian Mass Digitization: the University of Toronto Libraries partnership with the Internet Archive – historical overview, recent issues, and future implications". 74<sup>th</sup> IFLA General Conference and Council, July 16, 2008.  
[http://www.ifla.org.sg/IV/ifla74/papers/139-Bengtson\\_Miller-en.pdf](http://www.ifla.org.sg/IV/ifla74/papers/139-Bengtson_Miller-en.pdf)
- Blanche, C., Gueguen, N., Morineau, T., & Tobin, L. (2005). "The emergence of the contextual role of the e-book in cognitive processes through an ecological and functional analysis." *International Journal of Human-Computer Studies*, 62(3), 329-348.
- Christianson, M., & Aucoin, M. (2005). "Electronic or print books: Which are used?" *Library Collections, Acquisitions, and Technical Services*, 29(1), 71-81.
- Hillesund, T., & Noring, J. E. (2006). "Digital libraries and the need for a universal digital publication format." *JEP: the Journal of Electronic Publishing*, vol.9, no.2,

---

# References

- Levine-Clark, M. (2006). "Electronic book usage: A survey at the University of Denver." *portal: Libraries and the Academy*, 6(3), 285-299.
- Liu, Z. (2005) "Reading behavior in the digital environment." *Journal of Documentation* 61 (6), 700-712.
- Liu, Z. (2006), "Print vs. electronic resources: a study of user perceptions, preferences and use", *Information Processing and Management*, Vol. 42 No. 2, pp. 583-92.
- Noyes, J. & Garland, K. (2006) "Explaining students' attitudes toward books and computers." *Computers in Human Behavior* 22, 351-363.
- Su, S. (2005). "Desirable search features of web-based scholarly e-book systems." *Electronic Library*, 23(1), 64-71.
-