Google Books: The Metadata Mess

Google Book Settlement Conference
UC Berkeley
August 28, 2009

Geoff Nunberg,
School of Information
"The cost of creating such a library and Google’s significant lead time advantage suggest that no other entity will create a competing digital library for the foreseeable future."

Directors of ALA, ACRL, ARL in letter to DOJ Antitrust Division, July 29, 2009

There is no Moore's Law for capture…

Hence the urgency of concerns about pricing, access, exclusivity, privacy…and "quality"
Whose interests determine "quality"?

Google Book Search is "a tremendous public good for students, for teachers, for scholars, for everyone." Derek Slater, Google

… but students, scholars and "everyone" may have different purposes for using GBS.
Three ways of using GBS

What "Googling" means: barrelling in sideways

GBS as a borough of Greater Google

"We just feel this is part of our core mission. There is fantastic information in books. Often when I do a search, what is in a book is miles ahead of what I find on a Web site." Sergey Brin
Three ways of using GBS

Seeking out works & editions: the "destination experience"

A particular edition of *Leaves of Grass*
A good edition of *Tristram Shandy*
18th-c. French editions of *Don Quixote*, etc.

The importance of metadata: Who, when, where etc.
Three ways of using GBS

"Batch processing": data mining and "electronic philology"

"It's only reporters and computational linguists who care if [hit-count estimation] is really precise." Peter Norvig, Google

Text databases and the "new philologies":

The importance of language to social, intellectual, and political history & literary study

Coincides emergence of large-scale historical text databases…

- When did *happiness* replace *felicity* in 17th c?
- Plotting the rise & fall of *propaganda*
- How did *liberalism* spread in the early nineteenth-century European context?.
Good enough for scholarship?

Will GBS be an adequate resource for scholarly needs… now and in the future?

Depends on:

- Quality of imaging
- Reliability and robustness of search tools
- Quality and reliability of metadata
  - e.g., date, edition history, author, subject classification, etc.
Good enough for scholarship?

Will GBS be an adequate resource for scholarly needs... now and in the future?

Depends on:

- Quality of imaging
- Reliability and robustness of search tools
- Quality and reliability of metadata
e.g., date, edition history, author, subject classification, etc.

But GBS metadata are awful.
Quality Issues: Botched Scans, OCR, &c.
Metadata Issues:
1899, annus mirabilis
Random Dates

1905
1900
1848
1888
The pervasiveness of misdatings

527 hits returned for "Internet" before 1950
Famous before their lifetime

182 hits reported for "Charles Dickens" before birthdate (1812)
Cf Jimi Hendrix, 81; Led Zeppelin, 59 etc.
Ego-surfing,
Edgar Cayce
Style

"Our reputation precedes us"
The frequency of misdatings

Search on "candy bar" < 1920 yields 66 hits, 46 of them misdated (70%)
Classification Errors
Classification Errors

The merchant of Venice
by Christopher Rice, William Shakespeare - Foreign Language Study - 2007 - 72 pages
No preview available - About this book - Add to my library - More editions

The Century dictionary: an encyclopedic lexicon of the English language - Page 6180
edited by William Dwight Whitney - Family & Relationships - 1891
JD Whitney, The Yosemite Book, p. 24. 2. A device used in dredging, for sweeping the sea-bed in order to obtain delicate forms of marine life, ...
Full view - About this book - Add to my library - More editions

Melting the earth: the history of ideas on volcanic eruptions - Page 76
by Haraldur Sigurdsson - Foreign Language Study - 1999 - 260 pages
Dante also speculated on the powers that had raised the lands above the primeval ocean, opting for an extraterrestrial force. Action was effected "by way of ...
Snippet view - About this book - Add to my library - More editions

Catalogue of copyright entries - Page 744
by Library of Congress. Copyright Office - Drama - 1923
13, Feb., 1922» 0 Feb. 25, 1922 : 2 c. and aff. Mar. 9. ... Volume of production — united States ...
Full view - About this book - Add to my library - More editions
The Pervasiveness of Misclassification

Classifications of first 10 hits for Tristram Shandy

- family and relationships (4)
- fiction (4)
- biography and autobiography (1)
- Unlabeled (1) (others classified as "music," "history," "literary collections")
The Pervasiveness of Misclassification

First 10 hits for *Leaves of Grass* classify it as:

- Juvenile Nonfiction
- Poetry
- Fiction
- Literary Criticism
- Biography & Autobiography, Counterfeits and Counterfeiting
More bad metadata

*Jane Eyre: An Autobiography*
by Currer Bell - History - 2008 - 364 pages
Limited preview - About this book - Add to my library - More editions

*Jane Eyre*
by Charlotte Brontë - Governesses - 1986 - 483 pages
No preview available - About this book - Add to my library - More editions

*Jane Eyre*
by Charlotte Brontë - Love stories - 2000 - 452 pages
No preview available - About this book - Add to my library

*Jane Eyre*
by Charlotte Brontë, Basil Davenport - Architecture - 1946 - 474 pages
Snippet view - About this book - Add to my library - More editions

*Jane Eyre*
by Charlotte Brontë, Clare West - Antiques & Collectibles - 1992 - 105 pages
No preview available - About this book - Add to my library - More editions
Reader, I marketed him.
Other metadata issues

Books ascribed to authors of introductions, or given no author at all.
Other metadata issues

Books ascribed to authors of introductions, or given no author at all.

Titles linked to unrelated works.

"Errors" constitute only a fraction of "bad" metadata
Who is to blame and what is to be done?

"We got the metadata from the libraries":

yes, sometimes… but libraries didn't classify *Hamlet* as "antiques and collectibles" or *Speculum* as "Health & Fitness"

Libraries don't use headings like "Antiques and Collectibles" and "Health & Fitness" in the first place…
cf Google's decision to use BISAC

The world according to BISAC: 3000 subheadings vs. 200,000 for LOC
The world according to BISAC

Making space for Bambi & Bullwinkle

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>JNF003260</td>
<td>JUVENILE NONFICTION / Animals / Cows *</td>
</tr>
<tr>
<td>JNF003230</td>
<td>JUVENILE NONFICTION / Animals / Deer, Moose &amp; Caribou</td>
</tr>
<tr>
<td>JNF003050</td>
<td>JUVENILE NONFICTION / Animals / Dinosaurs &amp; Prehistoric Creatures</td>
</tr>
<tr>
<td>JNF003060</td>
<td>JUVENILE NONFICTION / Animals / Dogs</td>
</tr>
<tr>
<td>JNF003210</td>
<td>JUVENILE NONFICTION / Animals / Ducks, Geese, etc.</td>
</tr>
<tr>
<td>JNF003070</td>
<td>JUVENILE NONFICTION / Animals / Elephants</td>
</tr>
</tbody>
</table>

… and Schiller, Petrarch & Verlaine

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>POE011000</td>
<td>POETRY / Canadian</td>
</tr>
<tr>
<td>POE012000</td>
<td>POETRY / Caribbean &amp; Latin American</td>
</tr>
<tr>
<td>POE005030</td>
<td>POETRY / Continental European</td>
</tr>
<tr>
<td>POE005020</td>
<td>POETRY / English, Irish, Scottish, Welsh</td>
</tr>
<tr>
<td>POE014000</td>
<td>POETRY / Epic *</td>
</tr>
</tbody>
</table>
The world according to BISAC

Making space for Bambi & Bullwinkle

<table>
<thead>
<tr>
<th>Code</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>JNF003260</td>
<td>JUVENILE NONFICTION / Animals / Cows *</td>
</tr>
<tr>
<td>JNF003230</td>
<td>JUVENILE NONFICTION / Animals / Deer, Moose &amp; Caribou</td>
</tr>
<tr>
<td>JNF003050</td>
<td>JUVENILE NONFICTION / Animals / Dinosaurs &amp; Prehistoric Creatures</td>
</tr>
<tr>
<td>JNF003060</td>
<td>JUVENILE NONFICTION / Animals / Dogs</td>
</tr>
<tr>
<td>JNF003210</td>
<td>JUVENILE NONFICTION / Animals / Ducks, Geese, etc.</td>
</tr>
<tr>
<td>JNF003070</td>
<td>JUVENILE NONFICTION / Animals / Elephants</td>
</tr>
</tbody>
</table>

... and Schiller, Petrarch & Verlaine

<table>
<thead>
<tr>
<th>Code</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>POE011000</td>
<td>POETRY / Canadian</td>
</tr>
<tr>
<td>POE012000</td>
<td>POETRY / Caribbean &amp; Latin American</td>
</tr>
<tr>
<td>POE005030</td>
<td>POETRY / Continental European</td>
</tr>
<tr>
<td>POE005020</td>
<td>POETRY / English, Irish, Scottish, Welsh</td>
</tr>
<tr>
<td>POE014000</td>
<td>POETRY / Epic *</td>
</tr>
</tbody>
</table>

Squeezing the universal library into a sububan bookstore
Correcting the Problem

Google: "We're on it (but it isn't a first priority)"

Correcting errors as noticed (like bad scans)?

Crowd Sourcing?

But errors/bad metadata affect 000,000's of records

"Error correction" doesn't address poor & missing metadata, inconsistent/confusing/inappropriate classification schemes

Why should the metadata decisions be left to Google engineers?
Correcting the Problem

HathiTrust to the rescue?

But HathiTrust makes available only out-of-copyright works, has (relatively) limited computational resources.

Why should Google have no obligations to do GBS right?

Google Book Search is "a tremendous public good for students, for teachers, for scholars, for everyone."

Derek Slater, Google

But a public good implies a public trust.