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

<table>
<thead>
<tr>
<th>Title</th>
<th>Author</th>
<th>Genre</th>
<th>Year</th>
<th>Pages</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>KILLER IN THE RAIN</strong></td>
<td>Raymond Chandler</td>
<td>Fiction</td>
<td>1899</td>
<td>320</td>
<td>It was in the pulp detective magazines of the 1930s that Raymond Chandler's definitive take on the hard-boiled detective story first appeared. ...</td>
</tr>
<tr>
<td><strong>CULTURE AND SOCIETY 1780-1950 - Page 206</strong></td>
<td>Raymond Williams</td>
<td>Social Sciences</td>
<td>1899</td>
<td>320</td>
<td>This was the positive result of the life of the family in a small house, ... are not separated from personal relationships; and Lawrence knew from this, ...</td>
</tr>
<tr>
<td><strong>CONDITION HUMAINE, LA</strong></td>
<td>Andre Malraux</td>
<td>Fiction</td>
<td>1899</td>
<td>342</td>
<td>En mars 1927, l'armée du Kuomintang dirigée par Chang-Kai-Shek s'approche de Shanghai. Les communistes de la ville organisent le soulèvement des ouvriers ...</td>
</tr>
<tr>
<td><strong>PORTABLE DOROTHY PARKER, THE</strong></td>
<td>Dorothy Parker, Seth</td>
<td>Fiction</td>
<td>1899</td>
<td>360</td>
<td>The second revision in sixty years, the sublime collection ranges over the verse, stories, essays, and journalism of one of the twentieth century's most ...</td>
</tr>
<tr>
<td><strong>Christine</strong></td>
<td>Stephen King</td>
<td>Fiction</td>
<td>1899</td>
<td>394</td>
<td>Page 11 PRÓLOGO Esta é a história de um triângulo amoroso — suporho que este seria o nome — formado por Amie Cunningham, Leigh Cahot e, naturalmente, Christine. ...</td>
</tr>
<tr>
<td><strong>GRAPEs OF WRATH, THE: TEXT AND CRITICISM</strong></td>
<td>John Steinbeck</td>
<td>Literary Criticism</td>
<td>1899</td>
<td>376</td>
<td>This book includes the restored text corrected by scholar Robert DeMott, a map of the Joad family's journey, Steinbeck's declaration of intent in writing ...</td>
</tr>
<tr>
<td><strong>COMPLETE SHORTER FICTION OF VIRGINIA WOOLF, THE</strong></td>
<td>Susan Dick</td>
<td>Fiction</td>
<td>1899</td>
<td>300</td>
<td>This volume brings together all of Virginia Woolf's short stories and sketches. Limited preview</td>
</tr>
<tr>
<td><strong>YELLOW SUBMARINE</strong></td>
<td>The Beatles</td>
<td>Juvenile Fiction</td>
<td>1699</td>
<td>40</td>
<td>Uma aventura fascinante com a mais famosa banda de rock de todos os tempos! Era uma vez - ou duas, talvez - um paraíso chamado Pepperland. ...</td>
</tr>
</tbody>
</table>
Random Dates

1905 - The World According to Drucker
1848 - What Maisie knew
1900 - Letters from Virginia Woolf
1888 - The bonfire of the vanities
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.
"Our reputation precedes us"
The frequency of misdatings

Search on "candy bar" < 1920 yields 66 hits, 46 of them misdated (70%)
Classification Errors
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
Reader, I marketed him.
Other metadata issues

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

Titles linked to unrelated works.
Other metadata issues

Strange bedfellows

The Mosaic navigator: The essential guide to the internet interface
by Sigmund Freud, Katherine Jones - 1939 - 218 pages
Includes glossary, index.
No preview available - About this book - Add to my library
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 BISAC headings like "Antiques and Collectibles" and "Health & Fitness" in the first place…

And publishers didn’t assign BISAC codes to books published before the 1980’s
### 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>
The world according to BISAC

Making shelf 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 scrunching together 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 suburban 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.