Making Books Fully Searchable

EE 380

Daniel Clancy
Engineering Director, Google Print
18-July-05
Making Information Accessible

Arpanet Team
Larry and Sergey’s idea
Google’s Mission

To organize the world’s information and make it universally accessible and useful.

Online Content
Billions of web pages

Offline Content
Billions of items still unindexed
Two Initiatives

**Publisher Program**

Only ~15% of books are in print

**Library Program**

~85% of books are out of print and/or out of copyright – these books are only found in libraries

**GOAL:** Create a comprehensive virtual card catalog of all books in all languages, while respecting publishers’ rights
Google Books Library Project
Really, how many books?

<table>
<thead>
<tr>
<th>Library Holdings</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Library of Congress</td>
<td>24,616,867</td>
</tr>
<tr>
<td>Harvard University</td>
<td>14,437,361</td>
</tr>
<tr>
<td>Chicago Public Library</td>
<td>10,994,943</td>
</tr>
<tr>
<td>New York Public Library</td>
<td>10,608,570</td>
</tr>
<tr>
<td>Yale University</td>
<td>10,492,812</td>
</tr>
<tr>
<td>Queens Borough Public Library</td>
<td>10,357,159</td>
</tr>
<tr>
<td>Oxford University</td>
<td>10,000,000</td>
</tr>
<tr>
<td>….</td>
<td>….</td>
</tr>
<tr>
<td>University of Michigan</td>
<td>7,348,360</td>
</tr>
<tr>
<td>Stanford University</td>
<td>7,286,437</td>
</tr>
</tbody>
</table>
A Typical Library Collection

The value is in the middle

- ~15%
- ~65% or more
- Less than 20%**

**In-Print**
Partner Program

**Unclear copyright status**
Books after 1923 but…
- May be in copyright, but not for sale
- Rights may have reverted to author
- May be in the public domain

**Public Domain**
Books published before 1923

92% of the world's books are neither generating revenue for the copyright holder nor easily accessible to potential readers.*

*Source: Covey, Denise Troll. “Global Cooperation for Global Access: The Million Book Project”
Three User Experiences

Sample Pages View

Snippet View

Full Book View

~15%

65% or more*

20%
A Closer Look at the Snippet View

For books we scan that are still in copyright

• User can view:
  • Bibliographic info
  • A few sentences around the query
  • Restricted searching
  • Same 3 snippets, never more
  • Links to purchase
  • In-print – online bookstores
  • Out of print – used bookstores
User Limitations:

- 10%-20% of pages/user/month
- Scrolling 2 pages left or right
- Full text of book is indexed
- No cut/copy/print
Public Domain Books

- Pre-1923 (in US) books that are no longer in copyright

- **Public Domain UI**
  - Results displayed in onebox
  - Full text of book is indexed
  - Same as original UI with no browsing restrictions
  - No cut/copy/print
  - Additional link to ‘Find it in a library’
  - No attribution to the library that provided the book
Finding Books
It’s Not Just About Our Most Popular Searches…

What are people searching for? Everything

Number of Search Queries/Keyword

Keywords

1. Harry Potter
2. Wireless Networking
3. Peruvian Orchids
4. Jersey City

Google Confidential
Finding Books – “OneBox” Display

- Book search results displayed above search results
- Results appear due to search “triggers”
- Up to 3 books per search

“OneBox” – First of many implementations of Google Print
Finding Books – Search Only Books

Search the full text of books (and discover new ones).
Finding Books – Full Integration (Future)

Books Search results are fully integrated into main Google search results

- Ranking based on relevance, new factors
- Books flagged with “[Book]” for clarity to user
European Constitutionalism Beyond the State
edited by Joseph H Weiler, Marlene Wind - Law - 2003 - 252 pages
Page vi - Court of Justice and the European Economic Constitution (Oxford, Hart Publishing, ... 2001); The Social Construction of Europe (with Th. Christiansen and KE ... [ 69 results from this book ]

Integration in an Expanding European Union: Reassessing the Fundamentals
Page 17 - ... it spelt political death. To speak of a constitution for Europe was to be tainted with the “F” word — to be branded as an old fashioned federalist. ... [ 53 results from this book ]

edited by Robert Blackburn, Jorg Polakiewicz - Law - 2002 - 1051 pages
Page 191 - ... (6) of the Constitution of the Republic of Bulgaria which provides that 'the National Assembly shall ratify all international instruments concerning ... [ 246 results from this book ]

by Max M Edling - Law - 2003 - 352 pages
Page 43 - ... of government in continental Europe, their recurrent talk of the introduction of “Turkish despotism” in ... talked about the Revolution or the Constitution. ...
Ego-searching and discovering your past

“Never before could I have found such an obscure and wonderful gem. Google Book Search prompted me to buy two copies of a book that I never would have known about, otherwise.”

- Bernie Robichau, Columbia, SC
Search: Stanford Tuition

months; evening courses. Enrollment 257. Six weeks. Tuition $19.

SAN FRANCISCO ARCHITECTURAL CLUB, Sharon Building, 426 Post St.
Loring P. Rixford, superintendent. Four architectural ateliers main-
tained in co-operation with the Society of Beaux-Arts Architects. Enrollment 85.

SANTA BARBARA
State Normal School of Manual Arts and Home Economics.
Ednah A. Rich, President; sixteen instructors. Founded 1909. Indus-
trial, normal, applied, domestic and manual arts. Term of 48
weeks. Enrollment: domestic arts 10, manual arts (men only) 45,
industrial arts 23, applied arts 3.

SANTA CLARA
University of Santa Clara, College of Architecture.

STANFORD UNIVERSITY
Leland Stanford Junior University, Division of Graphic Arts, Depart-
ment of Education.
A. B. Clark, head of department; three instructors. Department
founded 1892. Normal art. Tuition $30 for 32 weeks; five year
course.

CANAL ZONE

CULEBRA
Hitt Atelier. Isthmian Canal Commission.
S. M. Hitt, superintendent. Architecture in co-operation with the

COLORADO

BOULDER
Colorado Chautauqua School of Art.
Ida M. Tindall, director; two instructors. Founded 1897. Normal
art. Tuition $10 for six weeks; two year course. Enrollment 20.

DENVER
Fine Arts Academy of Denver, 31 East Eighteenth Avenue.
Abigail Holman, director; five instructors. Founded 1912. Fine,
Themes and Issues

• Selection
• Archival Quality Versus Access
• Accessibility
• Copyright
• Preservation
• Internationalization
• Other Digitization Efforts
• Community Impact
Technical Challenges
Assessing the scale of the problem

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of books</td>
<td>30,000,000</td>
</tr>
<tr>
<td>Years of project</td>
<td>10</td>
</tr>
<tr>
<td>Books per year</td>
<td>3,000,000</td>
</tr>
<tr>
<td>Books per day</td>
<td>12,000</td>
</tr>
<tr>
<td>Pages per book</td>
<td>330</td>
</tr>
<tr>
<td>Pages per day</td>
<td>3,960,000</td>
</tr>
<tr>
<td>Image size per page</td>
<td>5</td>
</tr>
<tr>
<td>TBs a day</td>
<td>20</td>
</tr>
<tr>
<td>PBs per year</td>
<td>5</td>
</tr>
<tr>
<td>Pbs for project</td>
<td>50</td>
</tr>
<tr>
<td>Market cost per book</td>
<td>50</td>
</tr>
<tr>
<td>Cost of project at Market rate</td>
<td>1,500,000,000</td>
</tr>
</tbody>
</table>

Note all numbers here are simply example numbers to demonstrate the scale and challenges of such a project and are not intended to indicate specific plans by Google or our partners or commitments.
Book Flow Process

Some Principles:
• Use smart software when possible.
• Detect problems
• Fix them.
Book Workflow

Retrieve

Checkout

Transport

Return
Scanning

Factors:
- Quality
- Cost
- Speed
- Scalability
- Diversity of content
Google Scanning Technology
Source Problems

Torn Pages
Condition: 4

Text Block Condition: D
la mort ! Elle est pour lui un accident, même un scandale.

— C'est qu'il ne sait pas davantage concevoir la vie, dont la mort est une des propriétés caractéristiques. La vie est, en somme, quelque fourmillement bizarre entièrement confiné dans une pellicule de douze à quinze mille mètres d'épaisseur...
— Ectoderme...
— Naturellement... Tout ce qui est amusant est superficiel. Et accidentel. La vie a quelque chose d'un accident... qui s'est fait des lois.
— Oui... Et dans cette couche mince, vie et mort... Entrées et sorties. La loi fondamentale est statistique. Point de vie sans morts : c'est l'équilibre statistique...
— Et c'est là que vous intervenez, mon cher Docteur.
— J'essaye.
— L'intelligence ne comprend rien à
More recent books can sometimes be harder
What is the appropriate page number of this page?
Some “interesting” cases: what to do with Math?

Indeed, it follows from (3.5’) in view of (4.39) that (4.40) $S_k = \sum_{i=1}^{k} \left(1 - a\right) C_i - \left(1 - a\right) C + \text{From (4.40), } \phi(u)x^a = \phi$ and hence The last inequality means that $TTQ, SF(f, N)$. Further, it follows from (4.38) that (4.41) $P(X^+ > 0) = \phi$ and hence $TTQ, SF(f, N)$. Finally, we get from (4.38) and (4.41) that (4.42) $P(X^- > 0) = \phi$ and hence $TTQ, SF(f, N)$. The relations (4.41) and (4.42) show that the condition (4.35) holds for the strategy $\nu_a$, and hence $TTQ$ is an $a$-$(1 - a)C, f, A^\phi$-hedge. What has been obtained shows that it is possible to hedge a contingent claim with a specified probability $(1 - a)$. Further, the initial funds can be reduced by the amount $a C$, though with a risk $a$ the accepted contingent claim cannot be repaid. PROBLEMS 4.1. Prove that on a no-arbitrage $(B, 5)$-market we have for a standard European option to buy (sell) that $C(N^2) > C(N)$ (respectively, $P(N^2) > P(N)$) when $N > 4.2$. Prove that the fair price $C = C(N, S_0, K)$ of a standard European option to buy, where $N$ is the exercise time, $S_0$ is the initial price of a share, and $K$ is the exercise price, has the following properties: a) $C(S_0, K)$ is monotone in $S_0$ and $K$; b) $C(S_0, K)$ is convex in $S_0$ and $K$; c) $C(S_0, X) = A C(S_0, K)$ for $A > 0.$
Spell Correction?

- no longer. Den she fall down an' Wolf catch her. Wolf look up at Br'er Rabbit, cough-in' an' hangin' on de roof, an' 'e laugh and say, "Ah, ha! ol' fellow, you is hang on still, is you? You done well, but I'll hab you yet. I'll fix you dis night." Well, Rabbit hang on while 'e kin. De smoke da choke um an' de cough da strangle um tel 'e mos' fall; an' at lars' 'e see 'e gots for go. So 'e put 'e ban' in 'e pocket an' full up 'e mout' wid terbacker an' chew um tel 'e mout' full up wid de juice. Den 'e le' go, an' Wolf look up for catch um as 'e fall. But Br'er Rabbit spit out 'e whole mout' ful o'terbacker juice right in Wolf.
RIGHTS OF MAN:

BEING AN

ANSWER TO MR. BURKE’S ATTACK

ON THE

FRENCH REVOLUTION.

BY THOMAS Paine,
SECRETARY FOR FOREIGN AFFAIRS TO CONGRESS IN THE AMERICA*
WAR, Etc., Etc.
Zonton :
PUBLISHED FOR JAMES WATSON.
BY
HOLYOAKE AND CO., 147, FLEET STREET.
1866.
Erroneous metadata

Metadata claims books is Russian - However: 157 pages are in English!

- Automatically detect languages?
  - Better search
  - Better OCR
PROLOGO.

Voy a leerle unos manuscritos, que más desvelos costó á mi padre el sustraerlos á tu curiosidad, que el escribirlos. Sé que cometo una imprudencia satisfaciendo un femenil deseo que te acarrearás muchos dolores; pero contigo más quiero pecar de tolerante que de severo. Profanaré con el secreto la memoria de mi buen padre, mas añadiré quilates á tu cariño: entre los respetos debidos á la memoria de un padre muerto, y el amor
Page Ordering
Quality: Engage the User to Help

- Continually improving the processing software to improve quality.
- Focus on making content accessible to users as soon as possible.
- Biggest Issue: Variation across books
Research Challenges

What will the Digital Library of the future look like?

What are some of the research challenges enabled by the massive amounts of content available via Google Book Search?

How can Google act as a catalyst and enable some of this research through leveraged investment?
Interesting Problems: User Experience

Search and Integration with Web

- No explicit rich link structure.
- How do you approximate page rank, anchors, etc to blend results appropriately?

Landing page

- Given a query, what is the right thing to return to a user?

Browsing a Book

- How do you enable people to “browse” a book on the web while only providing access to a limited portion of the content to any one user?

Navigation Across Books

- How do I find a similar book to the one I am looking at?

Obtaining information about a book?

- Given limited access to view the text a book, what other information can be made available for a user to assess the relevance of a book
Interesting problems: Processing

OCR: pretty good, but not perfect
  • Internationalization
  • Old texts, manuscripts, marginal scribbles

Page Number Detection
Page Ordering
Meta-data and book equivalence
Segmentation and disaggregation
Book summarization
Interesting problems: The Back-end

An abstraction of the “backend” problem:

• Large scale acquisition of possibly imperfect book page images
• Large scale acquisition of meta-information about books
• Crunch it all continuously to produce a high-quality index
• Store the data, archive it, move it around

Luckily, there are mitigating factors…

• GFS, the Google File System
  http://labs.google.com/papers/gfs.html
• MapReduce, a general distributed processing framework
  http://labs.google.com/papers/mapreduce.html
• Large machine clusters
  http://labs.google.com/papers/googlecluster.html
Web of Off-line Content

How do you create an ontology of objects from the off-line world with the myriad of links that connect these objects

Some Relationships that exist:
- FRBR Hierarchy
  - Work, manifestation, hierarchy
- References
- Authorship
- Criticism and review
- Inclusion
- Individuals
- Events
- Temporal relationships
- Different perspectives
- Topical similarity
Finding Stuff

*How do I find what I am looking for?*

Search

- Rich, explicit network structure does not exist
- How to allow the user to be an active participant?

Browsing content

- What are the different dimensions of *similarity* between books and how to you allow people to navigate this space effectively.

Serendipitous Discovery

- How to you enable people to discover stuff they were not looking for?

Assistance

- How can one user’s effort help the experience of another user?
Questions?