



UC Berkeley School of Information

Google Books: The Metadata Mess

Google Book Settlement Conference
UC Berkeley
August 28, 2009

Geoff Nunberg,
School of Information



The Last Library

"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"?

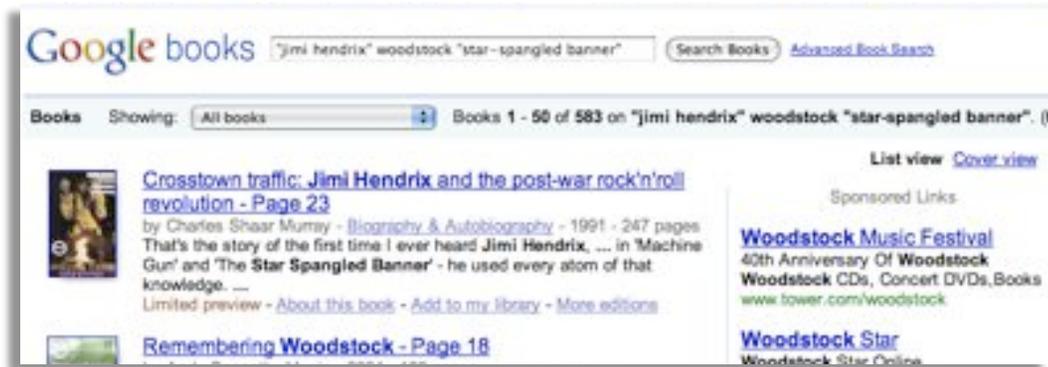
Google Book Search is "a tremendous public good for students, for teachers, for scholars, for everyone." Derek Slater, Google

... 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

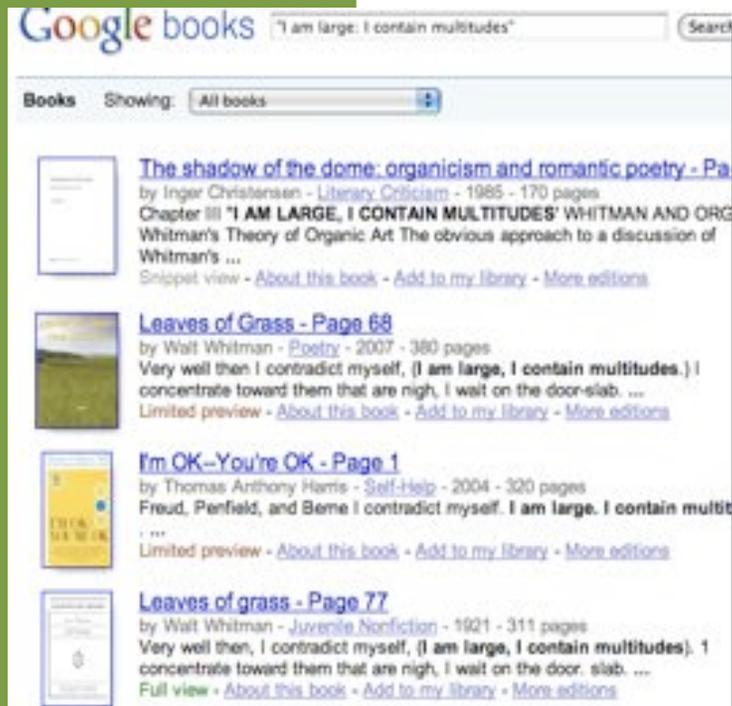
"We just feel this is part of our core mission. There is fantastic information in books. Often when I do a search, what is in a book is miles ahead of what I find on a Web site." Sergey Brin



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.



Web Images Videos Maps News Shopping Mail more

ekansa@alexandriaarchive.org | My library | Web history | My account

Google books statistical nlp Search Books

Foundations of statistical natural language processing By Christopher D. Manning, Heinrich Schütze

statistical nlp Go

4.5 (15) - Write review

Add to my library

Get this book

The MIT Press

Amazon.com - \$62.71

Barnes&Noble.com - \$74.71

Books - \$62.00

Find in a library

All sellers »

Related books

Sponsored Links

Cognitive Computing

Researcher in NLP mining and natural language processing

www.arshaw.com

The MIT Press

Pages included by permission of MIT Press

Result 1 of 100 in this book for statistical nlp - Previous Next - View all

statistical nlp

CHRISTOPHER D. MANNING AND HEINRICH SCHÜTZE

FOUNDATIONS OF STATISTICAL NATURAL LANGUAGE PROCESSING

Find: public domain | Sort: Previous | Highlight all | Match case

Showing 31 results in this book for Z113NZM - Order by: [relevance](#) | [pages](#)

[Clear](#)

[Page 1](#)

indiferen, capaz de ser facilmente arrasado pela maior força militar
a do direito à guerra preventiva, é escolher um alvo completamente
A manica mais simples de estabelecer uma nova norma, tal como
motivos duvidosos, cria-se uma norma. Esse é o significado do poder.
norma. Mas quando os Estados Unidos bombardam a Sérvia por



Metadata Issues: 1899, annus mirabilis



KILLER IN THE RAIN

by RAYMOND CHANDLER - Fiction - 1899

It was in the pulp detective magazines of the 1930s that Raymond Chandler's definitive take on the hard-boiled detective story first appeared. ...
No preview available - [About this book](#) - [Add to my library](#) - [More editions](#)



CULTURE AND SOCIETY 1780-1950 - Page 206

by RAYMOND WILLIAMS - Social Science - 1899

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, ...
[Limited preview](#) - [About this book](#) - [Add to my library](#) - [More editions](#)



CONDITION HUMAINE, LA

by ANDRE MALRAUX - Fiction - 1899 - 352 pages

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 ...
No preview available - [About this book](#) - [Add to my library](#) - [More editions](#)



PORTABLE DOROTHY PARKER, THE

by DOROTHY PARKER, SETH - Fiction - 1899 - 600 pages

The second revision in sixty years, this sublime collection ranges over the verse, stories, essays, and journalism of one of the twentieth century's most ...
No preview available - [About this book](#) - [Add to my library](#) - [More editions](#)



Christine

by Stephen King - Fiction - 1899 - 334 pages

Page 11
PRÓLOGO Esta é a história de um triângulo amoroso — suponho que este seria o nome — formado por Amie Cunningham, Leigh Cahot e, naturalmente, Christine. ...
No preview available - [About this book](#) - [Add to my library](#) - [More editions](#)



GRAPES OF WRATH, THE: TEXT AND CRITICISM

by JOHN STEINBECK - Literary Criticism - 1899 - 806 pages

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 ...
No preview available - [About this book](#) - [Add to my library](#) - [More editions](#)



COMPLETE SHORTER FICTION OF VIRGINIA WOOLF, THE

by SUSAN DICK - Fiction - 1899

This volume brings together all of Virginia Woolf's short stories and sketches.
[Limited preview](#) - [About this book](#) - [Add to my library](#) - [More editions](#)



YELLOW SUBMARINE

by THE BEATLES - Juvenile Fiction - 1899 - 40 pages

Uma aventura fascinante com a maior banda de rock de todos os tempos! Era uma vez - ou duas, talvez - um paraíso chamado Pepperland. ...
No preview available - [About this book](#) - [Add to my library](#) - [More editions](#)



Random Dates

The World According to Drucker



Overview
[Reviews](#) (0)
[Buy](#)

☆☆☆☆ (0) - [Write review](#)
[Add to my library](#)

Book overview

Peter Drucker Is The Man Who The Economist Calls 'The Biography Airs His Controversial Views On Business, Go Future.

No preview available - 1905 - 220 pages

1905

What Maisie knew By Henry James



Overview
[Reviews](#) (3)
[Buy](#)

Book overview

No preview available - 1848

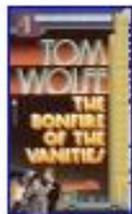
1848

Letters from Virginia Woolf to Jacques and Gwen Raverat

by Virginia Woolf - 1900 - 100 pages

No preview available - [About this book](#) - [Add to my library](#)

1900



The bonfire of the vanities

by Tom Wolfe - Fiction - 1888 - 690 pages

Tom Wolfe's modern American satire tells the story of Sherman McCoy, a Wall Street "Master of the Universe" who has it all — a Park Avenue apartment, a...

No preview available - [About this book](#) - [Add to my library](#) - [More editions](#)

1888



The pervasiveness of misdatings

	Internet security by Dan Farmer, WIETSE VENEMA - Computers - 1899 - 800 pages Computer forensics, the art and science of gathering and analyzing digital evidence, reconstructing data and attacks, and tracking perpetrators, ... No preview available - About this book - Add to my library	1899
	Multicultural Teaching With The Internet by Sunil Chawan - Internet in education - 1905 - 301 pages No preview available - About this book - Add to my library	1905
	Physics by Balfour Stewart - Education - 1878 - 149 pages Full view - About this book - Add to my library - More editions	1878
	Internet For Libraries And Information Centres (book + Floppy) by Jambhekar - Library information networks - 1905 - 174 pages No preview available - About this book - Add to my library	1905
	Emma - Page 55 Fiction - 1946 - 831 pages Die bekannteste dürfte der Internet Relay Chat (IRC) sein, ... Denn: Das Internet ist das, was seine Nutzerinnen und Nutzer daraus machen ... Snippet view - About this book - Add to my library - More editions	1946
	Internet Technologies And Applications 1905 No preview available - About this book - Add to my library - More editions	1905
	Mohit Dictionary Of Internet And Web by Vineet Gupta - Internet - 1905 - 392 pages No preview available - About this book - Add to my library	1905
	Project implementation and management - bridging the gap proc. 5th Internet ... 1876 - 305 pages No preview available - About this book - Add to my library	1905
	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	1939

527 hits returned for
"Internet" before 1950



Famous before their lifetime

The screenshot shows a Google Books search for "charles dickens". The search results list several books, including "Household words: a weekly journal", "A tale of two cities", "Great expectations", "A Christmas carol", and "Our mutual friend". A red circle highlights the "Return content published between" filter in the "Publication Date" section. An arrow points from this filter to a zoomed-in view of the same filter, which shows the search term "charles dickens" and the "Return content published between" filter set to "1812".

Web Images Videos Maps News Shopping Local more ▾ numbergg@gmail.com | My library | My Account |

Google books "charles dickens" Search Books Advanced Book Search

Publication Date Return content published anytime Return content published between [] and []

Books Showing: [All books] Books 1 - 100 of 182 on "charles dickens". (0.22 seconds)

Household words: a weekly journal
by Charles Dickens - Juvenile Fiction - 1742
Full view - About this book - Add to my library - More editions

A tale of two cities
by Charles Dickens - 1800 - 378 pages
No preview available - About this book - Add to my library - More editions

Great expectations
by Anne Marie Mueser, Charles Dickens - 1172 - 30 pages
No preview available - About this book - Add to my library - More editions

A Christmas carol
by Charles Dickens - Juvenile Fiction - 1135 - 294 pages
No preview available - About this book - Add to my library - More editions

Our mutual friend
by Charles Dickens - Drama - 1800
Snippet view - About this book - Add to my library - More editions

CHARLES DICKENS'S
Get Charles Dickens's
Find & Compare Great Options
Charles.Ask.com

Find Charles Dickens
Get current address, phone & more
Easy to use, search for free!
www.usa-people-search.com

"charles dickens" Search Books Advanced Book Search

Publication Date Return content published anytime Return content published between [] and [] 1812

Books 1 - 100 of 182 on "charles dickens". (0.22 seconds)

182 hits reported for "Charles Dickens" before birthdate (1812)

Cf Jimi Hendrix, 81; Led Zeppelin, 59 etc.

Ego-surfing, Edgar Cayce Style



VIRTUAL COMMUNITY: HOMESTEADING ON THE ELECTRONIC FRONTIER

by HOWARD RHEINGOLD - Computers - 1899
[Limited preview](#) - [About this book](#) - [Add to my library](#) - [More editions](#)



Do Que É Feito O Pensamento

by STEVEN PINKER - Psychology - 1899 - 44 pages
Page 411

... no really good? Além disso, como ressaltou o linguista
Geoffrey Nunberg,

embora de cara imaginar o dilema How brilliant was it? View INão



VIDA SOCIAL DA INFORMAÇÃO, A

by JOHN SEELY BROWN, PAUL DUGUID - Computers - 1899 -
300 pages

Com base em suas experiências profissionais como cientista-chefe
e pesquisador-



Annual report of the American historical Association - Page 198

History - 1884

... Berkeley Seized Letters as Legal Evidence in the Paris Revolutionary
Tribunal, 1793-1794. Carla Hesse, University of California, ...

Snippet view - [About this book](#) - [Add to my library](#) - [More editions](#)



New Jersey history

by New Jersey Historical Society - History - 1971

Annalee Saxenian, 'In Search of Power: The Organization of Business
Interests in Silicon Valley and Route 128,' Economy and Society 18 (1989): 25-70

; ...

Snippet view - [About this book](#) - [Add to my library](#) - [More editions](#)

"Our reputation
precedes us"



The frequency of misdatings

The image displays a series of overlapping screenshots from a Google Books search for "candy bar" with a date filter set to "before 1920". The search results show a list of books with snippets of text. Many of these snippets contain dates that are not less than 1920, demonstrating a high frequency of misdatings. For example, one snippet mentions "1910" and another mentions "1911". The search results are displayed in a grid-like format, with the most relevant results appearing on the right side of the collage.

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



Classification Errors



Hamlet

by William Shakespeare - [Antiques & Collectibles](#) - 1994 - 71 pages
très douces manières et de grande mine. En vérité, pour parler de lui avec tact,
il est le calendrier, la carte de la gentry ; vous trouverez en lui le ...
[Limited preview](#) - [About this book](#) - [Add to my library](#) - [More editions](#)



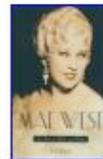
MADAME BOVARY:보봐리 부인(불문학작품 40)

by G.FLAUBERT - [Antiques & Collectibles](#) - 1988 - 430 pages
No preview available - [About this book](#) - [Add to my library](#)



Speculum - Page 283

by Mediaeval Academy of America, JSTOR (Organization) - [Health & Fitness](#) - 1962
WERNER JAEGER, Early Christianity and Greek **Paideia**. Cambridge, Massachusetts:
Belknap Press of Harvard University Press, 1961. Pp. ii, 154. \$8.25. ...
[Snippet view](#) - [About this book](#) - [Add to my library](#) - [More editions](#)



Mae West: An Icon in Black and White - Page 267

by Jill Watts - [Religion](#) - 2003 - 400 pages
Diamond Li's iand **Mae West**'si successful perpetuation of female trans- ... In
the midst of this revival of the **cult** of ...
[Limited preview](#) - [About this book](#) - [Add to my library](#) - [More editions](#)



Australian women: contemporary feminist thought - Page 153

by Norma Grieve, Ailsa Burns - [Foreign Language Study](#) - 1994 - 356 pages
The group moved more or less at random from the writings of Michel **Foucault** to
lesbian-feminist philosophy, from Freud to Freud's commentators and critics. ...
[Snippet view](#) - [About this book](#) - [Add to my library](#) - [More editions](#)



Classification Errors



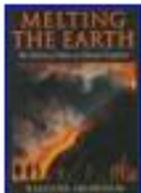
[The merchant of Venice](#)

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](#)

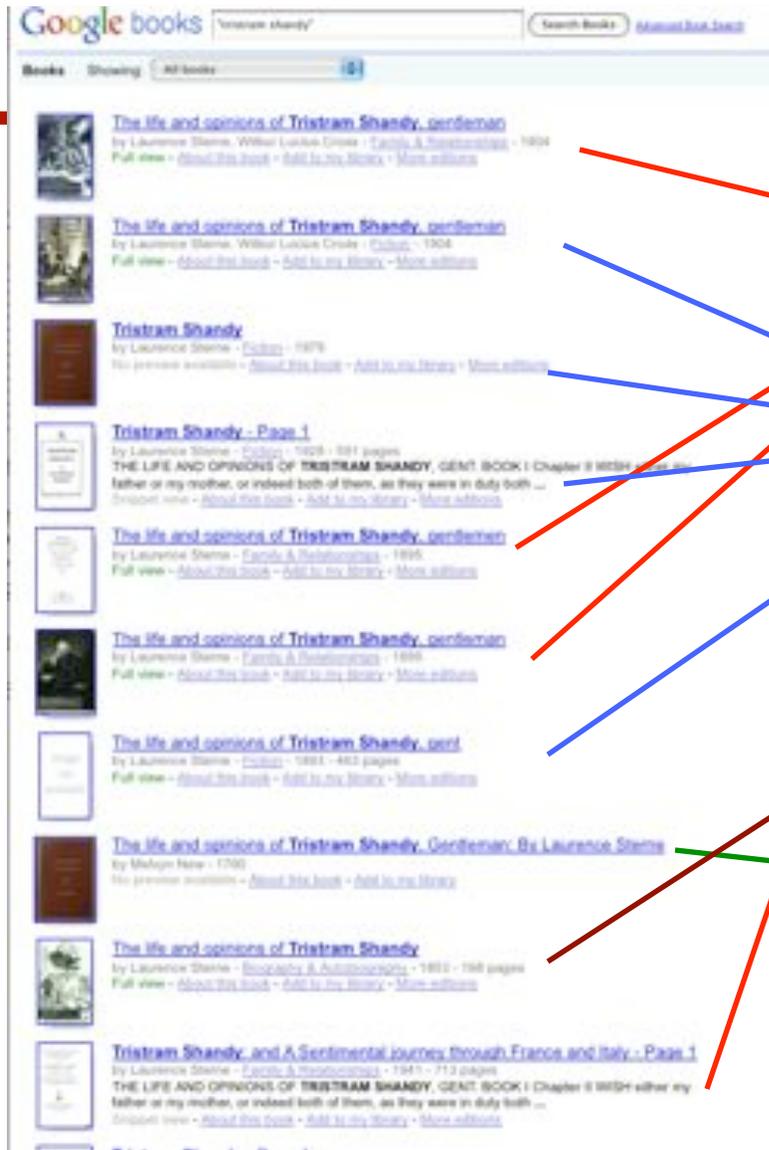


[Catalogue of copyright entries - Page 744](#)

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



family and relationships (4)

fiction (4)

biography and autobiography (1)

Unlabeled (1)
(others classified as "music,"
"history," "literary collections")

**Classifications of first 10 hits for
*Tristram Shandy***



The Pervasiveness of Misclassification

Web Images Videos Maps News Shopping Gmail more ▾ nunbergg@gmail.com | My library | My Account | Sign out

Google books "leaves of grass" Search Books Advanced Book Search

Books Showing: All books Books 1 - 17 of 17 on "leaves of grass". (0.45 seconds) List view Cover view

Leaves of grass
by Walt Whitman - *Juvenile Nonfiction* - 1921 - 311 pages
Unlike many other editions of Leaves of Grass, which reproduce various short, early versions, this Modern Library Paperback Classics "Death-bed" ...
Full view - About this book - Add to my library - More editions

Leaves of grass: first and "death-bed" editions : additional poems
by Walt Whitman, Karen Karbiener, George Stade - *Poetry* - 2004 - 913 pages
Here are some of the remarkable features of Barnes & Noble Classics: New introductions commissioned from today's top writers and scholars Biographies...
Limited preview - About this book - Add to my library - More editions

Leaves of grass
by Walt Whitman - *Fiction* - 1983 - 470 pages
In 1871 he brought out a fifth *Leaves of Grass*, which included "Passage to India," and in 1881 a sixth, in which the poems ...
Limited preview - About this book - Add to my library - More editions

Leaves of grass: including Sands at seventy, first annex, Good-by my fancy...
by Walt Whitman - *Literary Criticism* - 1931 - 580 pages
Snippet view - About this book - Add to my library - More editions

Leaves of grass
by Walt Whitman - *Biography & Autobiography* - 1897 - 455 pages
Copy is in a slip case, book has no covers.
Full view - About this book - Add to my library

Sponsored Links

Ornamental Grass Sale
Grow Ornamental Grasses For Less - \$20 Off Now At Michigan Bulb.
www.MichiganBulb.com

Trees For Every Season
Reinvent Your Living Space Today
Novel Trees Made with You in mind.
www.EarthFlorA.com

Glass Flowers
Handcrafted Glass Flowers & Glass Bouquets from Europe. Unique Gifts!
www.bestartshop.com

Plant Sod
Find Top-Rated Local Sod Pros & Get 4 Free Bids Today!
www.ServiceMagic.com

Patch Perfect Grass Seed
Official Patch Perfect Website
As Seen on TV. Buy 1 Get 1 Free
www.OfficialTVWebsite.com

Leaves Of Plants
Leaves Of Plants Online.
Shop Target.com.
www.Target.com

Plants at fresh&easy™
Fresh and healthy produce choices.
But low prices.

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](#)



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

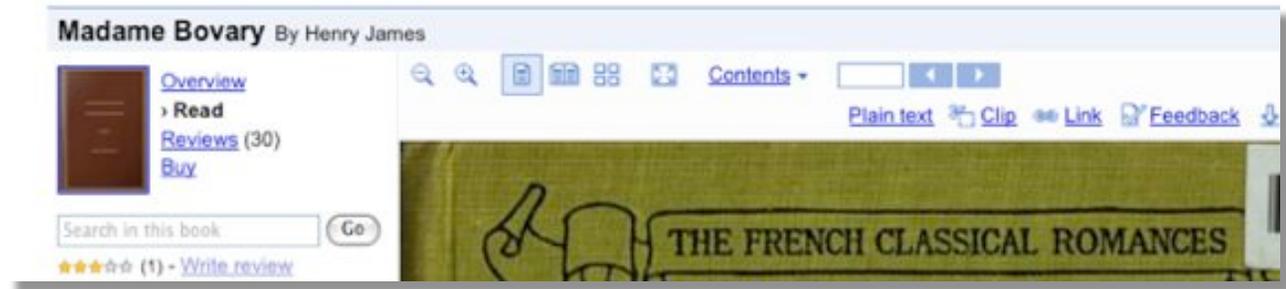
No preview available - [About this book](#) - [Add to my library](#) - [More editions](#)

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

JNF003260	JUVENILE NONFICTION / Animals / Cows *
JNF003230	JUVENILE NONFICTION / Animals / Deer, Moose & Caribou
JNF003050	JUVENILE NONFICTION / Animals / Dinosaurs & Prehistoric Creatures
JNF003060	JUVENILE NONFICTION / Animals / Dogs
JNF003210	JUVENILE NONFICTION / Animals / Ducks, Geese, etc.
JNF003070	JUVENILE NONFICTION / Animals / Elephants

... and Schiller, Petrarch & Verlaine

POE011000	POETRY / Canadian
POE012000	POETRY / Caribbean & Latin American
POE005030	POETRY / Continental European
POE005020	POETRY / English, Irish, Scottish, Welsh
POE014000	POETRY / Epic *



The world according to BISAC

Making space for Bambi & Bullwinkle

JNF003260	JUVENILE NONFICTION / Animals / Cows *
JNF003230	JUVENILE NONFICTION / Animals / Deer, Moose & Caribou
JNF003050	JUVENILE NONFICTION / Animals / Dinosaurs & Prehistoric Creatures
JNF003060	JUVENILE NONFICTION / Animals / Dogs
JNF003210	JUVENILE NONFICTION / Animals / Ducks, Geese, etc.
JNF003070	JUVENILE NONFICTION / Animals / Elephants

... and Schiller, Petrarch & Verlaine

POE011000	POETRY / Canadian
POE012000	POETRY / Caribbean & Latin American
POE005030	POETRY / Continental European
POE005020	POETRY / English, Irish, Scottish, Welsh
POE014000	POETRY / Epic *

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

