

# Doing Things with Text

Lisa Spiro

January, 2013

How do you want to use texts in teaching and/or research?

When do you think digital  
humanities/ humanities computing  
got started?

# 1949: Fr. Busa and the *Index Thomisticus*

CORPUS THOMISTICUM  
**INDEX THOMISTICUS**  
by Roberto Busa SJ and associates  
web edition by Eduardo Bernot and Enrique Alarcón  
English version

Search:

**FOUND 2211 CASES IN 1149 PLACES**

1-10

**Super Sent., lib. 1 q. 1 a. 2 co.** Respondeo. Ad hoc notandum est, quod intellectus Dei, tanto est magis unica et ad plura se extendit: unde intellectus Dei, qui est **lumen** quod est ipse Deus, omnium rerum cognitionem habet distincte. Ita et cur ipse Deus, omnium rerum cognitionem habet distincte. Ita et cur ipse **lumen** inspirationis divinae efficaciam habens, ipsa un-

**Super Sent., lib. 1 q. 1 a. 2 co.** [...] Ad hoc notandum est, quod intellectus Dei, qui est magis unica et ad plura se extendit: unde intellectus Dei, qui est ipse Deus, omnium rerum cognitionem habet distincte. Ita et cur ipse **lumen** inspirationis divinae efficaciam habens, ipsa un-



[http://en.wikipedia.org/wiki/Roberto\\_Busa](http://en.wikipedia.org/wiki/Roberto_Busa)

<http://www.corpusthomisticum.org/it/index.age>

# INTRODUCING TEXT ANALYSIS

# You already do text analysis



<http://www.flickr.com/photos/shareski/3027111111/>

# What is computer assisted text analysis?

---

- “use of computers as an aide to the interpretation of electronic texts” ([Sinclair & Rockwell](#))
- Use text analysis to:
  - Create a concordance
  - Conduct a complex search
  - Categorize works (e.g. by genre, author)
  - Compare/ contrast
  - Track changes in word usage
  - Visualize features of texts, e.g. maps, social networks

See Ted Underwood, “[Where to Start with Text Mining](#)”

# Simple Example: Tag Cloud of Presidential Speeches

---

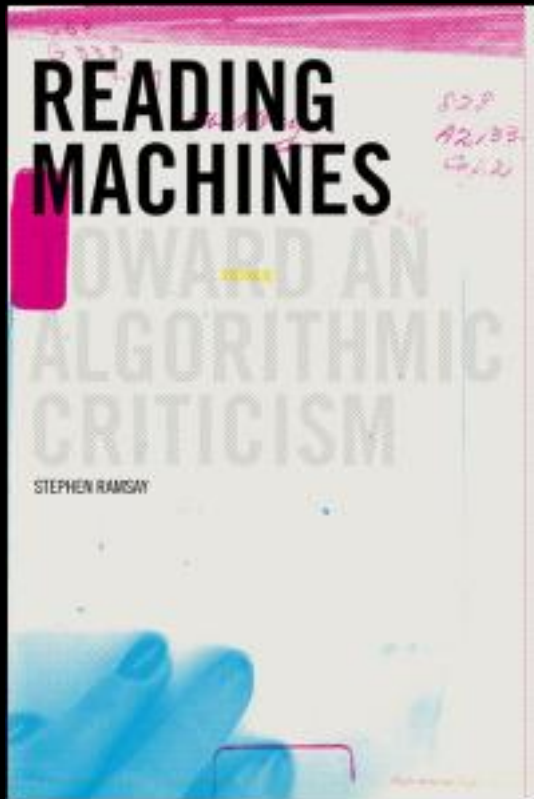
1961-01-20: Inaugural Address

John F. Kennedy (1961-63)

abolish achieving aggression agreements agricultural almighty annually appropriations armaments **armed**  
aspirations ballistic betterment blessings **burdens** **communist** compelled comprehensive constitution convert  
**cooperation** costly dependable devote discovery doubled **economic** education emerging  
endeavor engulf equality explore exports faithful families foe forebears forge **freedom** frustration global  
globe god guaranteed historically hostilities imperative inflation initiated integrity intellectual january labor  
lasting launching love misery **missiles** negotiate nixon nuclear oppose oppressed persistent **pledge** pray  
preserved productivity prosperity quest recognizes renewal reward satellites **science** scientific shield solemn  
soviet space **strength** strive subversion summons tax tempt terror treaty tyranny ultimately  
unemployment veterans victory vitality **war** weapons welfare witnessed wonders

<http://chir.ag/projects/preztags/>





# FRAMEWORKS

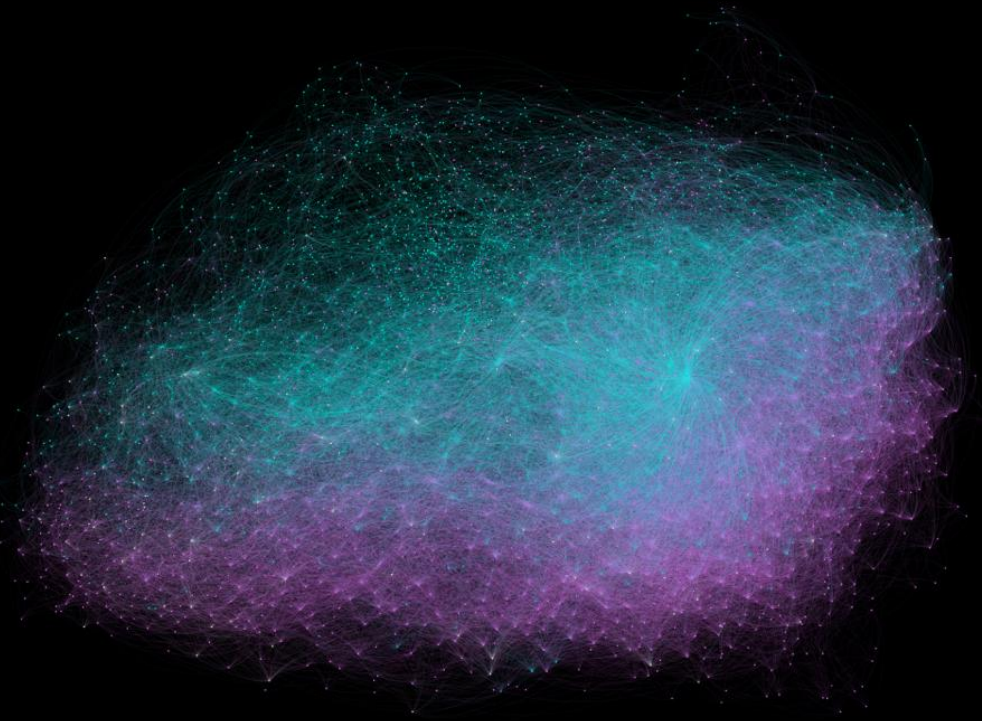
# The Screwmeneutical Imperative (Ramsay)

**Table 26 .2 First twenty-four terms for all characters in The Waves.**

<b>Bernard</b>		<b>Louis</b>		<b>Neville</b>	
thinks	rabbit	mr	clerks	catullus	loads
letter	tick	western	disorder	doomed	mallet
curiosity	tooth	nile	accent	immitigable	marvel
moffat	arrive	australian	beaten	papers	shoots
final	bandaged	beast	bobbing	bookcase	squirting
important	bowled	grained	custard	bored	waits
low	brushed	thou	discord	camel	stair
simple	buzzing	wilt	eating-shop	detect	abject
canopy	complex	pitchers	england	expose	admirable
getting	concrete	steel	eyres	hubbub	ajax
hoot	deeply	attempt	four-thirty	incredible	aloud
hums	detachment	average	ham	lack	bath
<b>Jinny</b>		<b>Rhoda</b>		<b>Susan</b>	
prepared	coach	dips	many-backed	washing	carbolic
melancholy	crag	bunch	minnows	apron	clara
billowing	dazzle	fuller	pond	pear	cow
fiery	deftly	moonlight	structure	seasons	cradle
game	equipped	party	wonder	squirrel	eggs
native	eyebrows	them	tiger	window-pane	ernest
peers	felled	allowed	swallow	kitchen	hams
quicker	frightened	cliffs	africa	baby	hare
victory	gaze	empress	amorous	betty	lettuce
band	jump	fleet	attitude	bitten	locked
banners	loquets	garland	bow	boil	maids

# Macroanalysis (Jockers)

---



[Matthew Jockers](#), 19<sup>th</sup> C novels by gender  
[Presentation at DH 2012](#)

# Use Text Analysis to Identify Simple Themes






---

- Find an electronic text. We'll use Obama's [second inaugural](#).
- Prepare text by removing any unnecessary stuff [we'll skip this]
- Generate a word list (sorted by frequency) using the [TAPoR List Words Tool](#). Under "Subtext limited to," select "all words."
- Examine the list for particular words you expect or don't expect to see.
- Pick a couple words you want to examine. Use [Find Words - Concordance Tool](#) to see them in context. Stick with the default context length of 5 words.

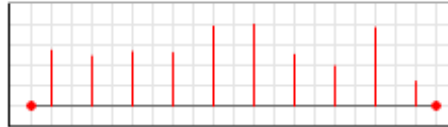
Adapted from [TAPoR Portal Recipes](#)

# TAPoR List Words: *Dracula*

Summary: There are 9752 unique words other than those in the stop list, there are 62589 words other than those in the stop list. There are 163996 words in total including the stop words.

Words	Distribution	Count
said		570
shall		424
know		398
time		384
come		336
van		321
came		305
helsing		299
went		298
like		282
room		230
lucy		225
way		224
man		223
good		222
took		222

# TAPoR Find Words– Concordance (“know” in *Dracula*)



Summary: 397 entries found.

, indeed , I don't	<b>know</b>	how I should be able
language which I did not	<b>know</b>	at all . I was
asked again : Do you	<b>know</b>	what day it is ?
Oh , yes ! I	<b>know</b>	that ! I know that
I know that ! I	<b>know</b>	that , but do you
that , but do you	<b>know</b>	what day it is ?
Day . Do you not	<b>know</b>	that tonight , when the
full sway ? Do you	<b>know</b>	where you are going ,
me . I did not	<b>know</b>	what to do , for
<u>itself</u> , I do not	<b>know</b>	, but I am not
with trees or hills I	<b>know</b>	not , for it is
, my friend . I	<b>know</b>	too much , and my

# TAPoR Collocates

## “know” in *Dracula*

**Summary:** There are 521 unique words other than those in the stop list, there are 1441 words other than those in the stop list. There are 3928 words in total including the stop words.

Words	Counts
	606
Dont	21
Shall	17
Let	12
Lucy	10
Oh	8
Said	8
Things	7
Time	7
Tell	7
Day	6
Better	6
Friend	6
Dear	6

# Exploring Crime in London: Old Bailey Archive

*The Proceedings of the* OLD BAILEY  *London's Central Criminal Court, 1674 to 1913*

[Home](#) | [Search](#) | [About The Proceedings](#) | [Historical Background](#) | [API](#) | [The Project](#) | [Contact](#)

## Old Bailey API Demonstrator

This query page allows you to locate and export the individual trials that make up the **Proceedings**. It is not configured to search the other types of text such as front matter and advertisements. This **Demonstrator** has been created to facilitate the dynamic exploration of trial results, and the export of trial texts and collections of trial URLs both to the bibliographical management system, [Zotero](#), and to the suite of tools for linguistic analysis available through [Voyant Tools](#). It also includes a **More Like This** function that allows you to build new searches based on a **Text Frequency - Inverse Document Frequency** (TF-IDF) methodology. To use the API directly please refer to the [Documentation for Developers](#). For detailed help on the categories of tagged information associated with each trial, please refer to the main help section on the [Guide to Searching](#).

Keyword(s)	<input type="text"/>
Defendant Gender	< All >
Offence Category	< All >
Offence Subcategory	Religious Offences
Victim Gender	< All >
Verdict Category	< All >
Verdict Subcategory	< All >
Punishment Category	< All >
Punishment Subcategory	< All >

[Query URL](#)  
[Zip URL](#)  
Send to Voyant: [10](#) [50](#) [100](#)

Break Down by:  
<nothing>

Offence Subcategory: Religious Offences [Undrill](#)

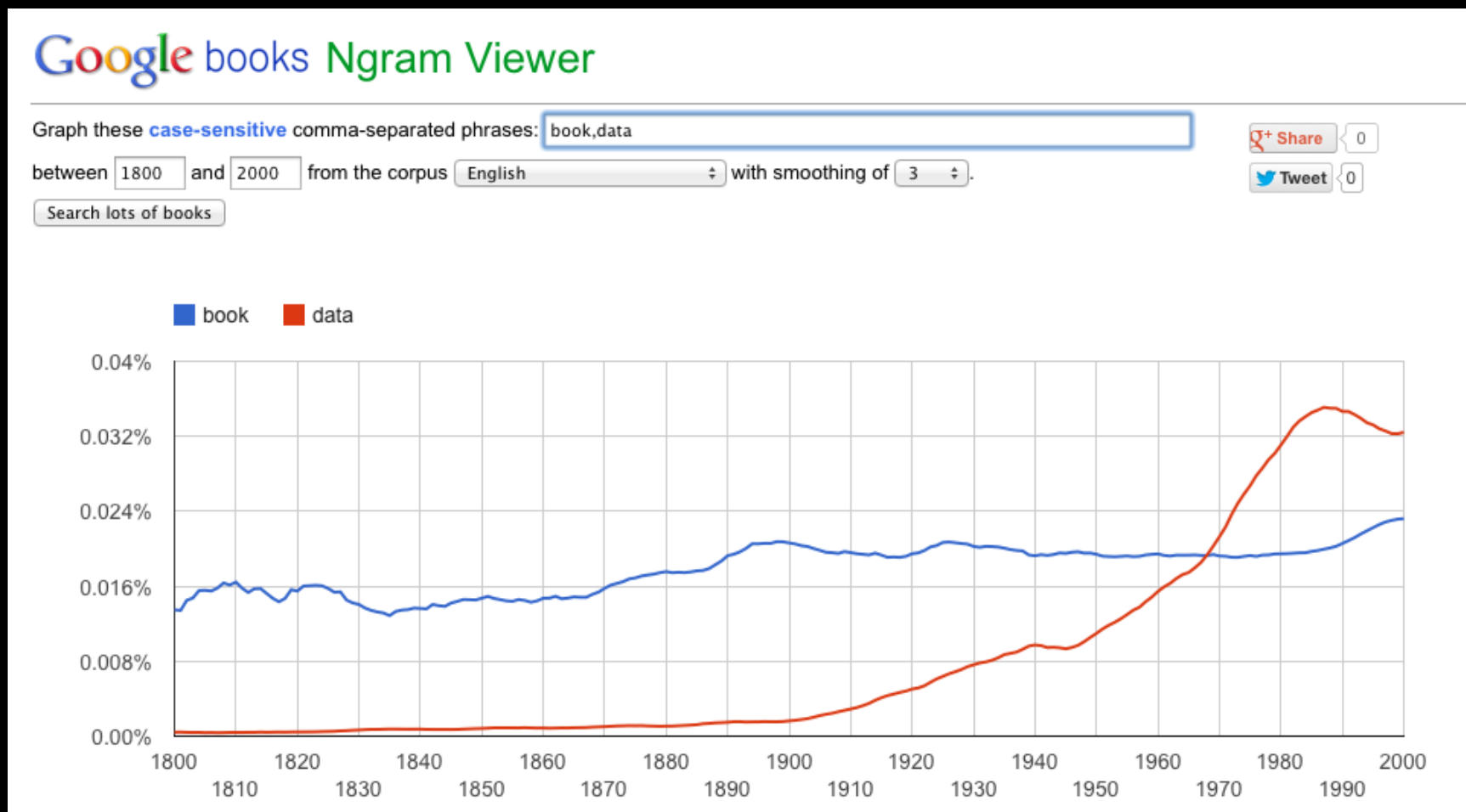
38 hit(s).

<http://www.oldbaileyonline.org/obapi/>





# Exploring Patterns in Texts: Google n-gram viewer



<http://books.google.com/ngrams>

Via Ngrams Tumblr: <http://ngrams.tumblr.com/>

# N-Grams in Action: “Culturomics”

The screenshot shows the Science journal website interface. At the top, the Science logo is followed by the tagline "The World's Leading Journal of Original Scientific Research, Global News, and Commentary." Below this is a navigation bar with links for "Science Home", "Current Issue", "Previous Issues", "Science Express", "Science Products", "My Science", and "About the Journal". The main content area displays the article title "Quantitative Analysis of Culture Using Millions of Digitized Books" by Jean-Baptiste Michel et al., published online on December 16, 2010. The article is categorized as a "RESEARCH ARTICLE". The abstract states: "We constructed a corpus of digitized texts containing about 4% of all books ever printed. Analysis of this corpus enables us to investigate cultural trends quantitatively. We survey the vast terrain of 'culturomics,' focusing on linguistic and cultural phenomena that were reflected in the English language between 1800 and 2000. We show how this corpus can provide insights into fields..."

**Science** The World's Leading Journal of Original Scientific Research, Global News, and Commentary.

Science Home Current Issue Previous Issues Science Express Science Products My Science About the Journal

Home > Science Magazine > 14 January 2011 > Michel *et al.*, 331 (6014): 176–182

Published Online December 16 2010  
Science 14 January 2011:  
Vol. 331 no. 6014 pp. 176–182  
DOI: 10.1126/science.1199644

< Prev | Table of Contents | Next >

Article Views

- > Abstract
- > Full Text
- > Full Text (PDF)
- > Figures Only
- > Supporting Online Material

VERSION HISTORY

- > 331/6014/176 (most recent)
- > science.1199644v1

Article Tools

- > Save to My Folders
- > Download Citation
- > Alert Me When Article is Cited
- > Post to CiteULike

RESEARCH ARTICLE

## Quantitative Analysis of Culture Using Millions of Digitized Books

Jean-Baptiste Michel<sup>1,2,3,4,5,\*†</sup>, Yuan Kui Shen<sup>2,6,7</sup>, Aviva Presser Aiden<sup>2,6,8</sup>, Adrian Veres<sup>2,6,9</sup>, Matthew K. Gray<sup>10</sup>, The Google Books Team<sup>10</sup>, Joseph P. Pickett<sup>11</sup>, Dale Hoiberg<sup>12</sup>, Dan Clancy<sup>10</sup>, Peter Norvig<sup>10</sup>, Jon Orwant<sup>10</sup>, Steven Pinker<sup>5</sup>, Martin A. Nowak<sup>1,13,14</sup>, Erez Lieberman Aiden<sup>1,2,6,14,15,16,17,\*†</sup>

± Author Affiliations

†To whom correspondence should be addressed. E-mail: [jb.michel@gmail.com](mailto:jb.michel@gmail.com) (J.-B.M.); [erez@erez.com](mailto:erez@erez.com) (E.L.A.)

\* These authors contributed equally to this work.

ABSTRACT

We constructed a corpus of digitized texts containing about 4% of all books ever printed. Analysis of this corpus enables us to investigate cultural trends quantitatively. We survey the vast terrain of 'culturomics,' focusing on linguistic and cultural phenomena that were reflected in the English language between 1800 and 2000. We show how this corpus can provide insights into fields...

Get all of Science

Join Now!

ADVERTISEMENT

WEBINAR

## Genetic Biomarkers Revealed

Unraveling the Complexities of Cancer Genomes in Blood Malignancies

<http://www.sciencemag.org/content/331/6014/176.abstract>

[http://www.ted.com/talks/what we learned from 5 million books.html](http://www.ted.com/talks/what_we_learned_from_5_million_books.html)

# What might be some potential issues with culturomics?

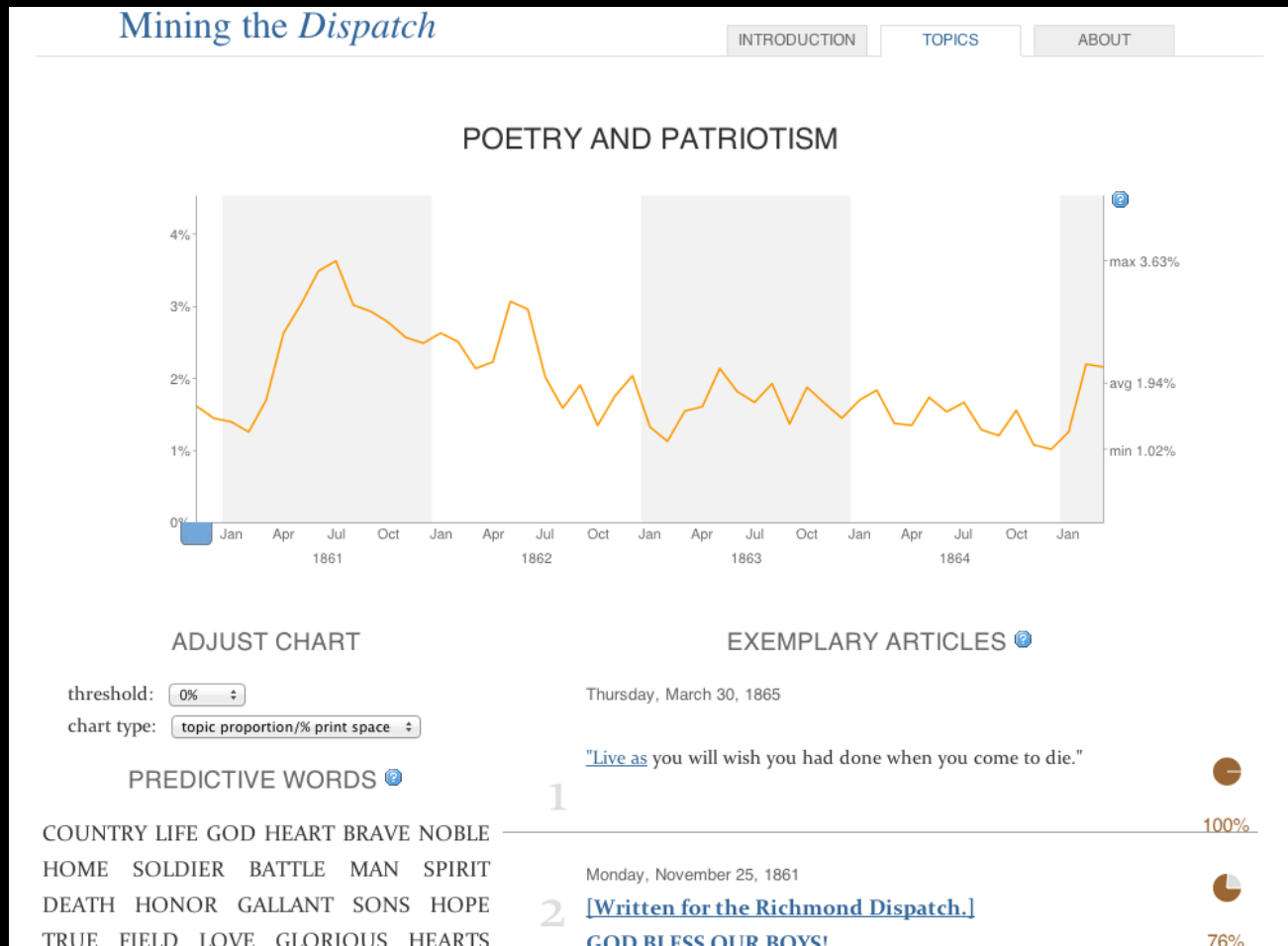
---

- Quality of the data (OCR errors)
- Quality of the metadata (e.g. errors in dates)
- Ambiguity in how words are used,
  - Parts of speech: tear, fence, play, etc.
  - Shifts in meaning of words over time
- Variations in spelling
- Qualms about quantifying culture
- Need for interpretation to make sense of these patterns

NB: Many of the data & metadata issues have been addressed in the most recent release of the n-gram viewer.

See Geoffrey Nunberg, ["Counting on Google Books"](#)

# “cardiogram of the Confederate nation”: Topic Modeling the *Dispatch*



Rob Nelson, [Of Monsters, Men — And Topic Modeling](#) & [Mining the Dispatch](#)

**INTRODUCING <TEXT MARKUP/>**

# Representing Texts via the Text Encoding Initiative (TEI)

---

- Standard for representing features of text
- Used for many scholarly [electronic editions](#)
- Human and machine readable
- Makes explicit features of a text so that they can be processed by computer applications
- Enables exchange of data & preservation
- Supports range of output formats (HTML, PDF, Braille reader, etc.)

# Creating Textual Apparatus with TEI

I have<sup>\*</sup> | I HAVE<sup>\*</sup> | I HAVE<sup>\*</sup> | I HAVE<sup>\*</sup> | I HAVE<sup>\*</sup>

got a quiet

farmhouse<sup>\*</sup> | farm-house<sup>\*</sup> | farmhouse<sup>\*</sup> | farmhouse<sup>\*</sup> | farmhouse<sup>\*</sup>

in the country, a very humble place to be sure, tenanted by a worthy enough man, of the old New-England stamp, where I sometimes go for a day or two in the

winter, to<sup>\*</sup> | winter to<sup>\*</sup> | winter, to<sup>\*</sup> | winter, to<sup>\*</sup> | winter, to<sup>\*</sup>

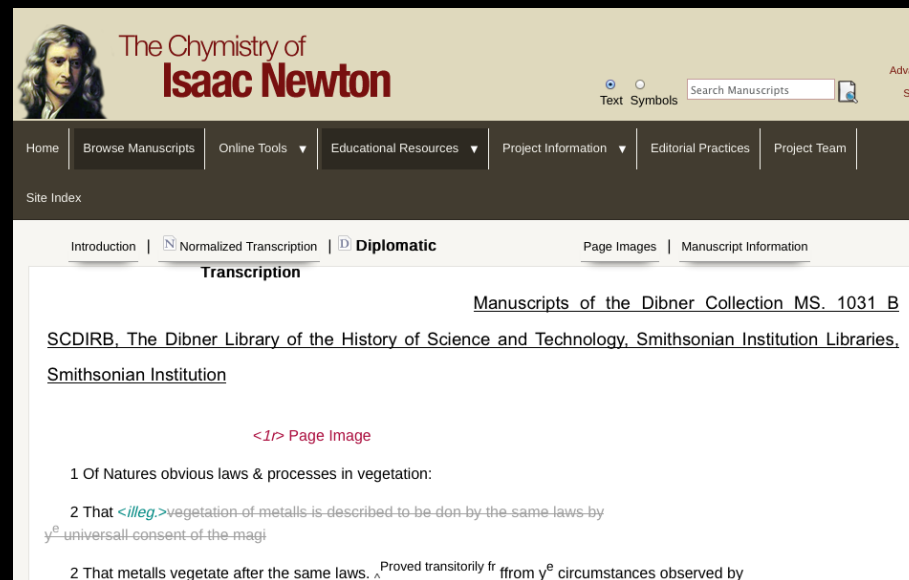
```
<app id="15.1">
<rdg wit=1849SLM type="accidental">I have </rdg>
<rdg wit=1849H type="accidental">I HAVE </rdg>
<rdg wit=1850S type="accidental">I HAVE </rdg>
<rdg wit=1877S type="accidental">I HAVE </rdg>
<rdg wit=1884S type="accidental">I HAVE </rdg>
</app>
```



# What can you do with marked up text?

Explore the [Chymistry of Isaac Newton](#)

- Browse the manuscripts and select one.
- Toggle between the normalized & diplomatic transcription.
- Look at the manuscript information ([metadata](#)).



The screenshot displays the website for 'The Chymistry of Isaac Newton'. At the top left is a portrait of Isaac Newton. The title 'The Chymistry of Isaac Newton' is prominently displayed. A search bar with the text 'Search Manuscripts' and a 'Text Symbols' toggle is visible. Below the header is a navigation menu with links for Home, Browse Manuscripts, Online Tools, Educational Resources, Project Information, Editorial Practices, and Project Team. A 'Site Index' link is also present. The main content area shows the 'Diplomatic' transcription view of a manuscript page. The page title is 'Manuscripts of the Dibner Collection MS. 1031 B'. The text on the page is a list of observations about vegetation and metals, with some words in red indicating markup. The text reads: '1 Of Natures obvious laws & processes in vegetation: 2 That <illeg>vegetation of metallis is described to be don by the same laws by y<sup>e</sup> universall consent of the magi 2 That metallis vegetate after the same laws. Proved transitorily fr from y<sup>e</sup> circumstances observed by

# What does markup look like?

---

Check out an example of a marked-up text:

<http://docsouth.unc.edu/southlit/poe/poe.xml>

What do you notice?


```
- <body>
  <pb id="poe1" n="1"/>
  - <div1>
    <head>TALES</head>
    <byline>BY</byline>
    <docAuthor>EDGAR A. POE.</docAuthor>
  - <div2 type="chapter">
    <head>THE GOLD-BUG.</head>
    - <epigraph>
      - <lg type="poem">
        <l>What ho! what ho! this fellow is dancing mad!</l>
        <l>He hath been bitten by the Tarantula.</l>
      - <l>
        <hi rend="italics">All in the Wrong.</hi>
      </l>
      </lg>
    </epigraph>
  - <p>
    MANY years ago, I contracted an intimacy with a Mr. William Legrand. He was of an ancient Huguenot family,
    reduced him to want: To avoid the mortification consequent upon his disasters, he left New Orleans, the city of h
    Charleston, South Carolina.
  </p>
```

# Compare Metadata Records: Internet Archive

The screenshot displays the Internet Archive interface for a specific book record. The top navigation bar includes 'Web', 'Video', 'Texts', 'Audio', 'Projects', 'About', 'Account', 'TVNews', and 'OpenLibrary'. The main header identifies the site as 'The LIBRARY of CONGRESS'. A search bar is present with a 'GO!' button and an 'Advanced Search' link. The user is identified as 'Anonymous User' with a login option and an 'Upload' button.

The breadcrumb trail reads: [Ebook and Texts Archive](#) > [The Library of Congress](#) > [Reveries of a bachelor. Or a book of the heart](#)

**View the book**



[Read Online](#) (~322 pg)  
[PDF](#) (11.0 M)  
[BW PDF](#) (9.5 M)  
[EPUB](#) (~322 pg)  
[Kindle](#) (~322 pg)  
[Daisy](#) (~322 pg)  
[Full Text](#) (352.1 K)  
[DiVu](#) (5.4 M)

All Files: [HTTPS Torrent](#) (2/0)  
[Help reading texts](#)

**Resources**  
[Bookmark](#)

**Reveries of a bachelor. Or a book of the heart (1851)**

Author: [Marvel, Jk., 1822-1908](#)  
Publisher: [New York, C. Scribner](#)  
Language: [English](#)  
Call number: 6403913  
Digitizing sponsor: [Sloan Foundation](#)  
Book contributor: [The Library of Congress](#)  
Collection: [library\\_of\\_congress: americana](#)

Full catalog record: [MARCXML](#)

[open](#) This book has an [editable web page](#) on [Open Library](#).

**Description**

Added t.-p., illustrated

**Reviews** [Be the first to write a review](#)  
Downloaded 77 times

**Selected metadata**

Page-progression: lr  
Scanningcenter: capitolhill  
Mediatype: texts  
Identifier-bib: 00159715681  
Identifier: reveriesofbachel01marv  
Ppi: 400  
Camera: Canon 5D  
Operator: scanner-ganzorig-purevee@...

<http://archive.org/details/reveriesofbachel01marv>

# Compare Metadata Records: Google Books

## Reveries of a bachelor: or A book of the heart (Google eBook)



+1 1

[Ik Marvel](#)

★★★★★

[2 Reviews](#)

C. Scribner, 1853 - 298 pages

[Search inside](#)

[Preview this book »](#)

[http://books.google.com/books?id=6mURAAAYAAJ&dq=editions:R21e4svUy5kC&source=gbs\\_navlinks\\_s](http://books.google.com/books?id=6mURAAAYAAJ&dq=editions:R21e4svUy5kC&source=gbs_navlinks_s)

# Compare Metadata Records: HATHI Trust

The screenshot shows the Hathi Trust Digital Library interface. At the top left is the logo and name 'HATHI TRUST Digital Library'. To the right are links for 'Help' and 'Feedback'. Below this is a navigation bar with 'Home', 'About', 'Collections', and 'My Collections'. The main section is titled 'Catalog Search' and contains a search input field, a dropdown menu set to 'All Fields', a 'Full view only' checkbox, and a 'Find' button. To the right of the search bar are links for 'Advanced Catalog Search' and 'Search Tips'. On the left side, there is a 'Similar Items' section listing several related records. The main content area displays the record for 'Reveries of a bachelor; or, A book of the heart. By Ik. Marvel [pseud.]'. It includes a small image of the book cover, the title, and a table of metadata. At the top right of the record area are links for 'Cite this' and 'Export to Endnote'. At the bottom of the record area is a 'Viewability' section with a 'Full view' link and a note that the original is from the New York Public Library.

**HATHI TRUST Digital Library** [Help](#) [Feedback](#)

[Home](#) [About](#) [Collections](#) [My Collections](#)

Catalog Search

**Catalog Search**  All Fields  Full view only  [Advanced Catalog Search](#) [Search Tips](#)

**Similar Items**

**Reveries of a bachelor. Or a book of the heart.**  
By: Mitchell, Donald Grant, 1822-1908.  
Published: (1851)

**Reveries of a bachelor : or, A book of the heart /**  
By: Mitchell, Donald Grant, 1822-1908.  
Published: (1852)


**Reveries of a bachelor: or A book of the heart.**  
By: Mitchell, Donald Grant, 1822-1908.  
Published: (1851)

**Reveries of a bachelor, or A book of the heart.**  
By: Mitchell, Donald Grant, 1822-1908.  
Published: (1853)

**Reveries of a bachelor: or A book of the heart.**  
By: Mitchell, Donald Grant, 1822-1908.  
Published: (1851)


**Fresh gleanings; or, A new sheaf from the old fields of continental Europe.**

**Reveries of a bachelor; or, A book of the heart. By Ik. Marvel [pseud.]** [Cite this](#) [Export to Endnote](#)



Main Author:	Mitchell, Donald Grant, 1822-1908.
Language(s):	English
Published:	Philadelphia, David McKay [1850]
Physical Description:	289 p. illus. 20 cm.
Original Format:	Book
Locate a Print Version:	<a href="#">Find in a library</a>

**Viewability:**

 [Full view](#) (original from New York Public Library)

<http://catalog.hathitrust.org/Record/008660628>

# Challenges and Strategies for Doing Text Analysis

---

Challenge	Strategy
Getting access to texts (especially given copyright)	Use <a href="#">digital collections</a> , e.g. <a href="#">HATHI Trust</a> , <a href="#">Open Library</a>
Bad metadata, e.g. errors w/ dates	Inspect metadata; compare to other sources & <a href="#">enhance</a>
Poor data: words breaking across lines, bad OCR, etc	Use <a href="#">pre-processing</a> to clean up OCR, tag parts of speech, etc
Bias in creation of collection	Be explicit about method; recognize limitations
Making sense of the data	Be clear about your process
Acquiring skills to do this work	Use online tutorials; collaborate; attend workshops