
Digital History: Language and Text Mining

For HIST 511, Topics in Public History: Digital History
Wednesday, March 3rd, 2010

By Roger Bilisoly, Ph.D.
Department of Mathematical Sciences, CCSU

Why study digital text documents?

Text is hard enough: images, video, etc.,
are even more complex.

Scanned type

This aged portion of society were distinguished from

OCR reads as

"niis aged pntkm at society were distinguished from."

PSALM XC. ver. 2.

9

Before the



were brought forth, or ever the Earth

and the



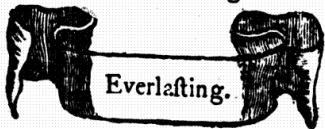
was formed,

thou art



Even

from Everlasting to



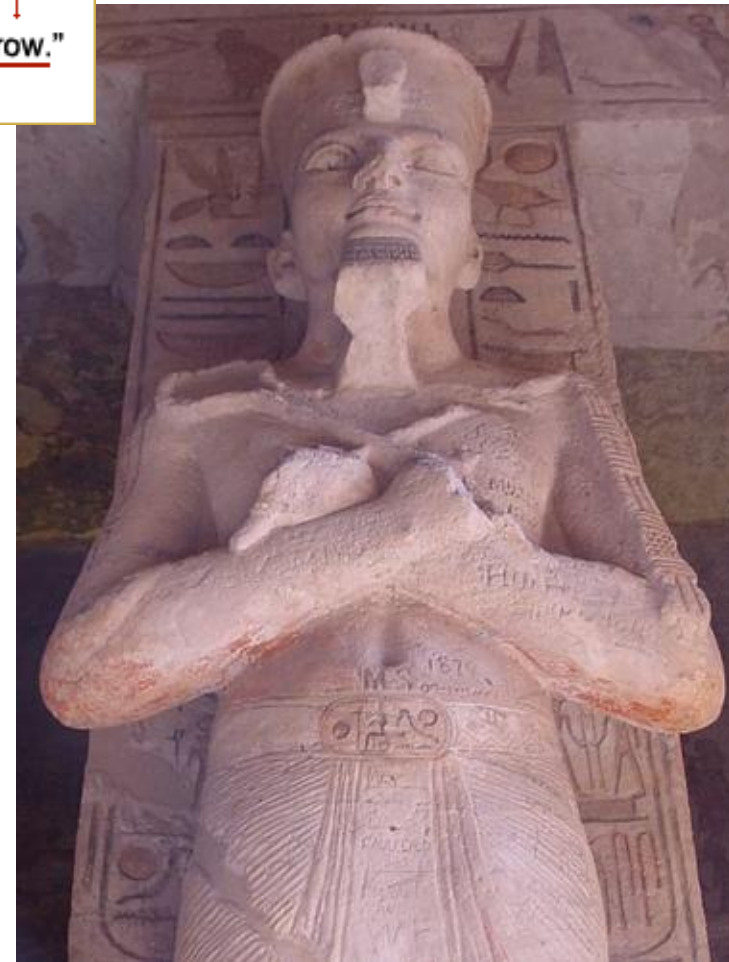
Everlasting.

Before the Mountains were brought forth, or ever
the Earth and the World was formed, thou art God:
Even from Everlasting to Everlasting.

• Top: Example of optical character recognition (OCR) from reCAPTCHA, a project that "digitizes books one word at a time." <http://recaptcha.net/learnmore.html>

• Left: an English Hieroglyphic Bible published by Isaiah Thomas in 1788. This is from the "Early American Imprints" database.

• Right: "Abu Simbel Graffiti Statue." Note the 19th century graffiti on its torso. Taken by Sebastian Niedlich and posted on Flickr at <http://www.flickr.com/photos/42311564@N00/53117067/>



Overview of Talk

1. What is a letter?
2. What is a word?
3. An extended example analyzing the *Dictionary of Canadian Biography*. This example was inspired by Associate Professor William Turkel's discussion posted in his (now inactive) blog "Digital History Hacks."
4. Miscellaneous and References

Main Idea: Although language is complex, computers can profitably analyze text.

Language is complex: For example, What is a letter in English?

- It's tempting to answer that a letter is one of the following: a, b, ...,z, A, B, ..., Z.
- But this is a simplification, which becomes obvious once actual texts are considered.

Beginning of Edgar Allan Poe's "Morella"

MORELLA



Αὐτὸ καθ' αὐτὸ μεθ' αὐτοῦ, μονοειδὲς δει ὄν.

Itself, by itself solely, ONE everlastingly, and single.

PLATO: *Symposium*, 211, B, XXIX.

WITH a feeling of deep yet most singular affection I regarded my friend Morella. Thrown by accident into her society many years ago, my soul, from our first meeting, burned with fires it had never before known; but the fires were not of Eros, and bitter and tormenting to my spirit was the gradual conviction that I could in no manner define their unusual meaning, or regulate their vague intensity. Yet we met; and fate bound us together at the altar; and I never spoke of passion, nor thought of love. She, however, shunned society, and, attaching herself to me alone, rendered me happy. It is a happiness to wonder;— it is a happiness to dream.

Edgar Allan Poe's "The Gold-Bug"

Here Legrand, having reheated the parchment, submitted it to my inspection. The following characters were rudely traced, in a red tint, between the death's head and the goat:

"53++!305))6*;4826)4+)4+).;806*;48!8]60))85;1+8*:+(;+*8!83(88)5*!;
46(;88*96*?;8)*+(;485);5*!2:*+(;4956*2(5*-4)8]8*;4069285);)6!8)4++;
1(+9;48081;8:8+1;48!85;4)485!528806*81(+9;48;(88;4(+?34;48)4+;161;:
188;+?;"

"But," said I, returning him the slip, "I am as much in the dark as ever. Were all the jewels of Golconda awaiting me upon my solution of this enigma, I am quite sure that I should be unable to earn them."

From <http://www.online-literature.com/poe/32/>

“Sir Gawain and the Green Knight” ca 1375-1400

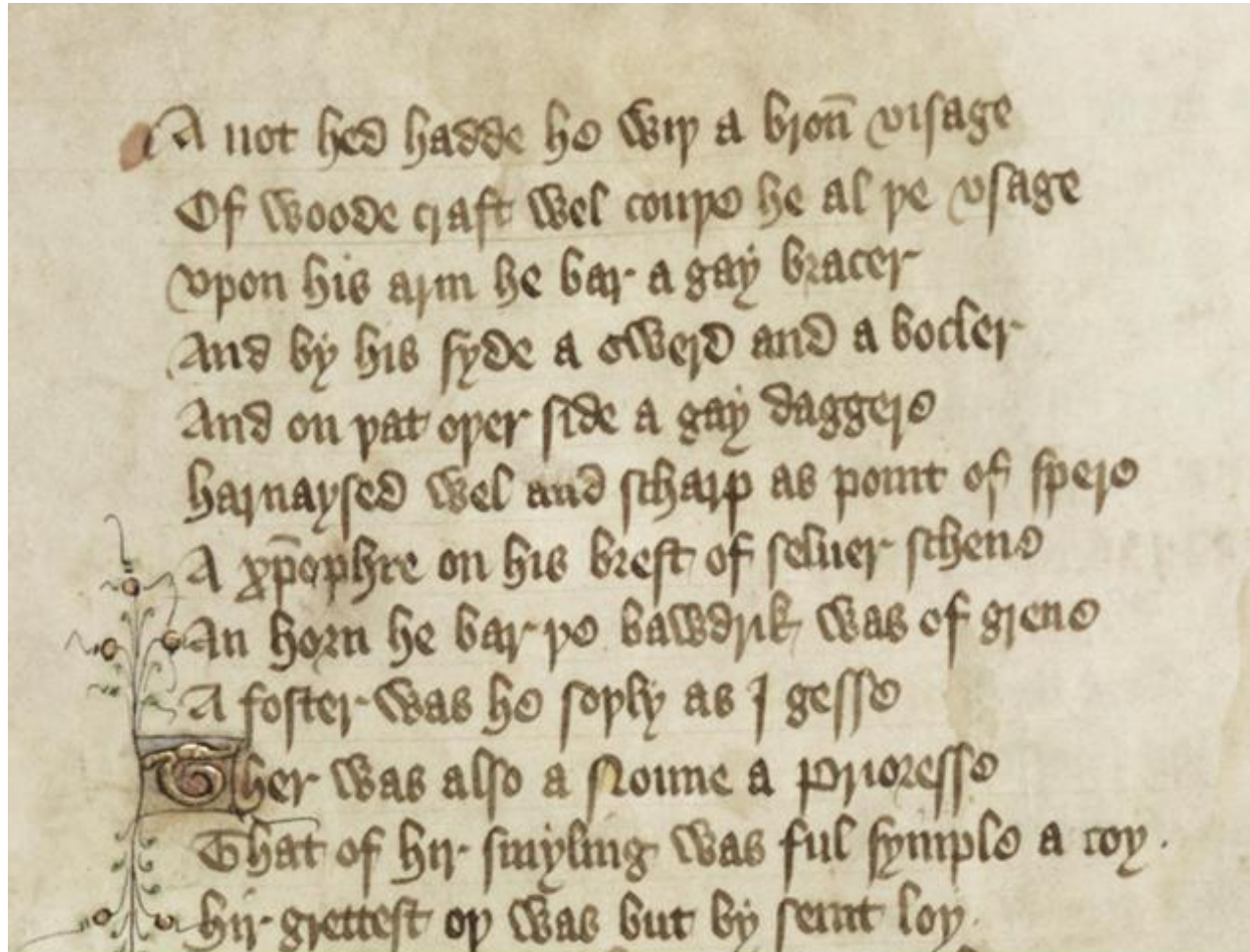
SIPEN þe sege and þe assaut watz sesed at Troye,
þe borȝ brittened and brent to brondeȝ and askez,
þe tulk þat þe trammes of tresoun þer wroȝt
Watz tried for his tricherie, þe trewest on erthe:
Hit watz Ennias þe athel, and his highe kynde,
þat siþen depreced prouinces, and patrounes bicomme
Welneȝe of al þe wele in þe west iles.

Name	Capital	Small	Origin	Description
Ash	U+00C6 Æ	U+00E6 æ	Latin	As in modern English "hat"
Thorn	U+00DE Þ	U+00FE þ	Futharc	þorn: modern "th" (survives in Icelandic)
Eth	U+00D0 Ð	U+00F0 ð	Old Irish	Eð, þæt: modern "th" (still in Icelandic)
Yogh	U+021C ȝ	U+021D ȝ	Old Irish	Y, gh, g, w (not to be confused with Ezh)
Wynn	U+01F7 ƿ	U+01BF ƿ	Futharc	(or Wen): modern "w"

Passage: <http://quod.lib.umich.edu/cgi/t/text/text-idx?c=cme:idno=Gawain:rgn=div1;view=text;cc=cme;node=Gawain%3A1>
Middle English letters with Unicode and descriptions: <http://www.columbiauniversity.org/kermit/st-erkenwald.html>

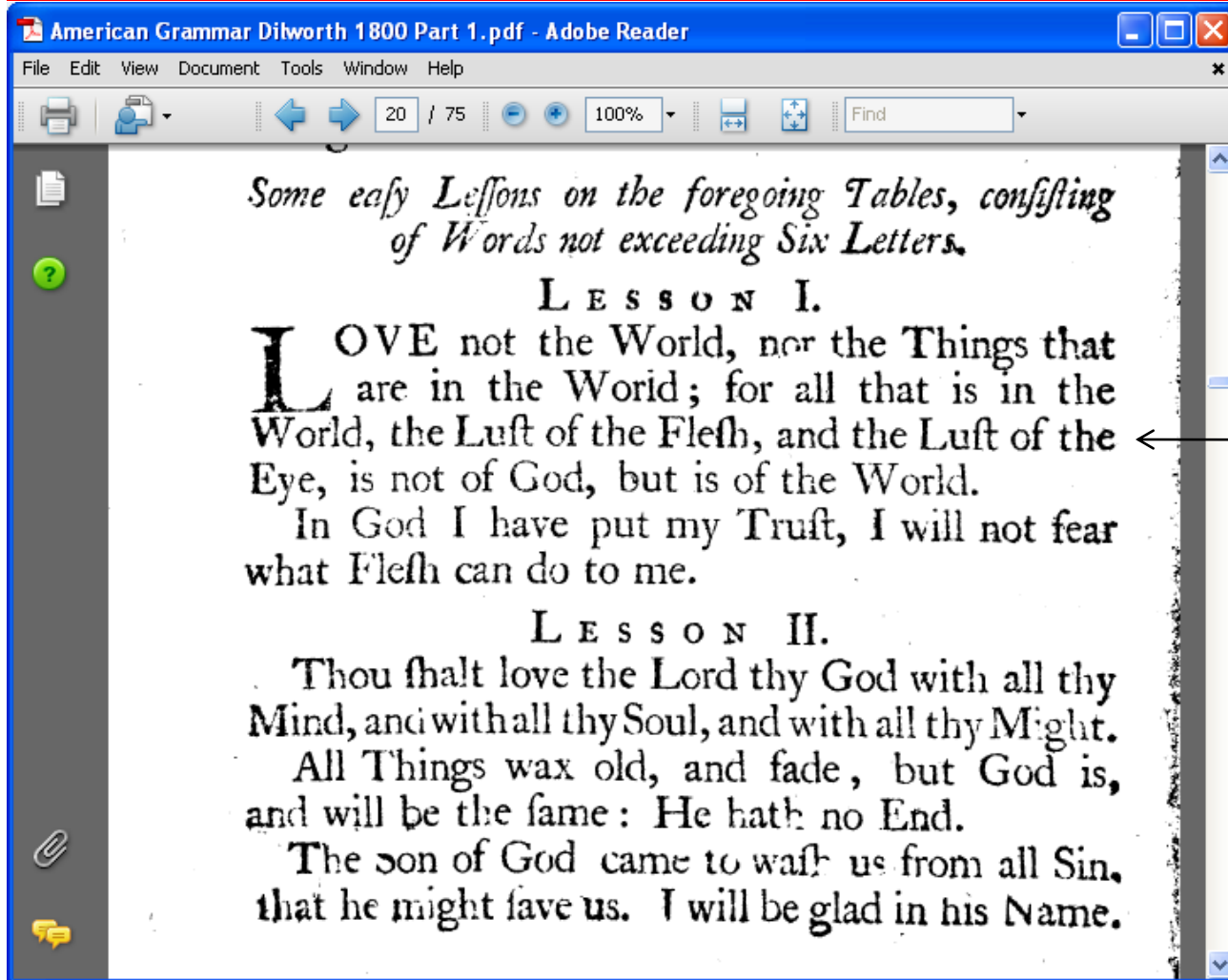
Corpus Christi College MS. 198

Chaucer, late 14th century



<http://image.ox.ac.uk/show?collection=corpus&manuscript=ms198>

A New Guide to the English Tongue by Thomas Dilworth ca. 1800



← Note the long s and the capitalization of the nouns.

Example of 1337

Babel:1337 - Uncyclopedia, the content-free encyclopedia - Windows Internet Explorer

http://uncyclopedia.wikia.com/wiki/Babel:1337

1337 leet

File Edit View Favorites Tools Help

Links Books Copy CCSU DM Dict TFD Digg FB Flickr Search Map Moodle MSNBC Vista EN


McAfee SiteAdvisor

Babel:1337 - Uncyclopedia, the content-free encyclop...

Log in / create account

babel discussion edit history

Babel:1337

 **This article is being considered for deletion in accordance with Uncyclopedia's deletion policy.**
This page may not fit in Uncyclopedia, or may not be funny with little chance for redemption.
Please share your thoughts on the matter at [this article's entry](#) on the [Votes for deletion](#) page.

W3|(0m3 70 |_]n(y(|0p3d14, 73h (0n73n7-fr33 3n(y(|0p3d14 7h47 4ny0n3 (4n 3d17.

637 y0ur 4r53 0u7 7h3r3 4nd (r3473 4n 4r71(|3, 0r h4n6 4r0und 7h3 0n35 w3 h4v3 un71| 4 5h1ny 08j3(7 d157r4(75 y0u.

|]073:1n 73h 3n6|15h 4nd 07h3r |4n6u4635 0n |_]n(y(|0p3d14, w3 h4v3 25,070 4r71(|35 4nd m4ny 1m4635 (73h pr0n!!!!!!1!!!110n3)

R3(3n7 4r71(|35:

[edit]

- h0w 70 83 73h wln 47 z0rk
- Fly!n6 5p46h377! M0n573r
- (0MMUN!5M
- j00 h4v3 7w0 (0w5
- L35B14N5
- K4N'3 VV357
- H3110 K177Y
- ()5(4r VV1!d3
- 1337 5p33k
- y0ur m0m
- 1337:53cr37 N4z1 F0r357 H4x

-4dd m0r3 n0w, 8147(h.-

navigation

- Main Page
- UnNews
- Featured content
- Babel
- Recent changes
- Random page
- Help
- Things to do
- Report a problem

community

- Community portal
- Village Dump
- Chatroom
- Pee Review
- Votes for Highlight
- Votes for Pictures
- Requested Articles

rating

☆☆☆☆☆
rate this article!

Internet 100%

Foreign names need to be transliterated into English, but this might be done in more than one way.

- Gauß
 - This German mathematician's name is spelled *Gauss* in English.
- Чебышёв
 - However, many transliterations exist for this Russian mathematician: Chebyshev, Chebychev, Tchebycheff, Tchebychev, Tchebyshev,...
- 北京
 - Beijing (Hanyu Pinyin), Peking (British Postal), Pei-ching (Wade-Giles), Northern Capitol (Literal translation)
 - There are other Chinese names for this city such as 燕京, so many other transliterations are possible.

The Internet has introduced new characters into text.

- **Example from the EnronSent Email Corpus: now @ is no longer an unusual character.**

- ARSystem@ect.enron.com on 08/07/2000 07:03:23 PM
This request has been pending your approval for 8 days. Please click <http://itcApps.corp.enron.com/srrs/Approve/Detail.asp?ID=000000000000935&Email=jeffrey.a.shankman@enron.com> to approve the request or contact IRM at 713-853-5536 if you have any issues.

Request ID : 000000000000935
Request Create Date : 7/27/00 2:15:23 PM
Requested For : paul.t.lucci@enron.com
Resource Name : EOL US NatGas US GAS PHY FWD FIRM Non-Texas < or = 1
Month
Resource Type : Applications

Corpus available at <http://verbs.colorado.edu/enronsent/>

Language is complex: For example, What is a word in English?

- Dividing a text into words is called *word segmentation*.
 - We ignore here the complexities of linguistic morphology.
- This is easy for a human (who is fluent in the language in which the text is written.)
- However, the details are tricky.

Word segmentation can be difficult.

- *Combination Concrete #2* by Stuart Davis, 1956. On display at the Yale University Art Gallery, New Haven.
- Challenge: Where is the signature of artist?

My photo, which was taken 8-2009.
<http://www.flickr.com/photos/66082566@N00/3881754254/sizes/m/>



Words from Anthony Burgess' *A Clockwork Orange* (1962)

• Nasdat	Meaning	Russian (transliterated)
• Bezoomy	Mad	Byezoomiyi
• Chepooka	Nonsense	Chyepookha
• Droog	Friend	Droog
• Gulliver	Head	Golova
• Koshka	Cat	Koshka
• Moloko	Milk	Moloko
• Ptitsa	Girl	Ptitsa
• Viddy	To see	Vidyet

Words can go out of style.

- What's wrong with the following sentence?

The mission is too important to allow you to jeopardize it.

– HAL in *2001, A Space Odyssey*

- “Jeopardize. This is a modern word which we could easily do without, as it is neither more nor less than its venerable progenitor *to jeopard*, which is greatly preferred by all careful writers.”
 - Alfred Ayres, 1895, *The Verbalist*
- “Jeopardize is a foolish and intolerable word”
 - Richard Grant White, 1870, *Words and Their Uses, Past and Present*
- “I dare not be so bold with my soul as to *jeopard* it in that manner.”
 - 1654 (from the Oxford English Dictionary)

See http://www.bisso.com/ujg_archives/000545.html

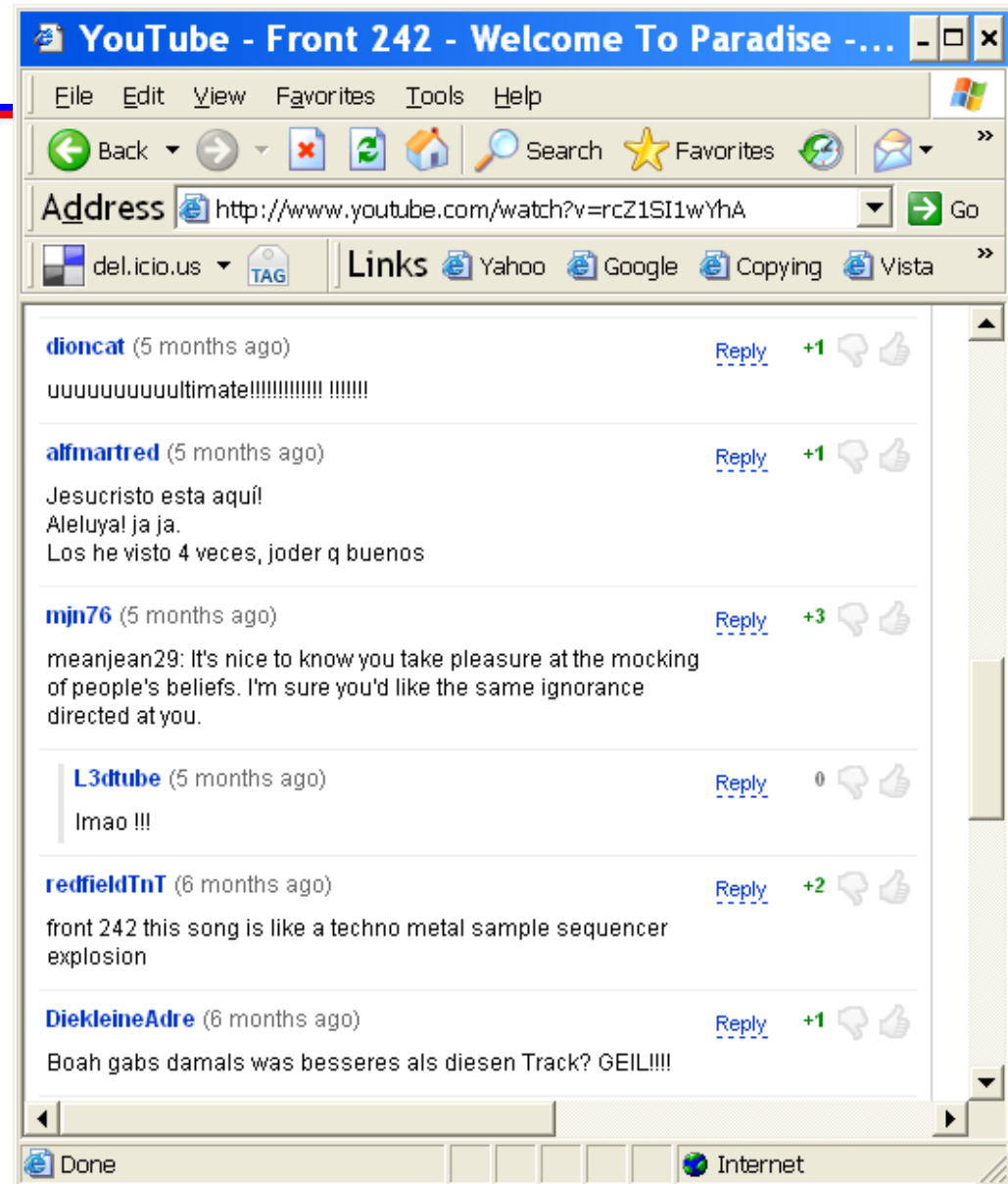
Complication: Multiple spellings

- American vs. British English orthography
 - Color vs. colour (see <http://www2.gsu.edu/~wwwesl/egw/jones/differences.htm>)
- Older forms of English
 - “Sir Gawain and the Green Knight”
- Slang, Dialect
 - Night vs. nite
 - “Well, you see, it 'uz dis way. Ole missus--dat's Miss Watson--she pecks on me all de time, en treats me pooty rough, but she awluz said she wouldn' sell me down to Orleans.” Jim in Mark Twain's *Huckleberry Finn*
- Spam often purposely misspells:
 - Check **This** St0ck Out!
 - **This** JUst In
 - “This” also turns out to be a surname:
 - Patrick **This**' Web page at <http://www.cuyamaca.net/pthiss/>

Complication: Online free-for-all

Online fan sites are not limited to English. This allows slang, abbreviations, typos, etc., in multiple languages for a single document.

User names can also be unusual.

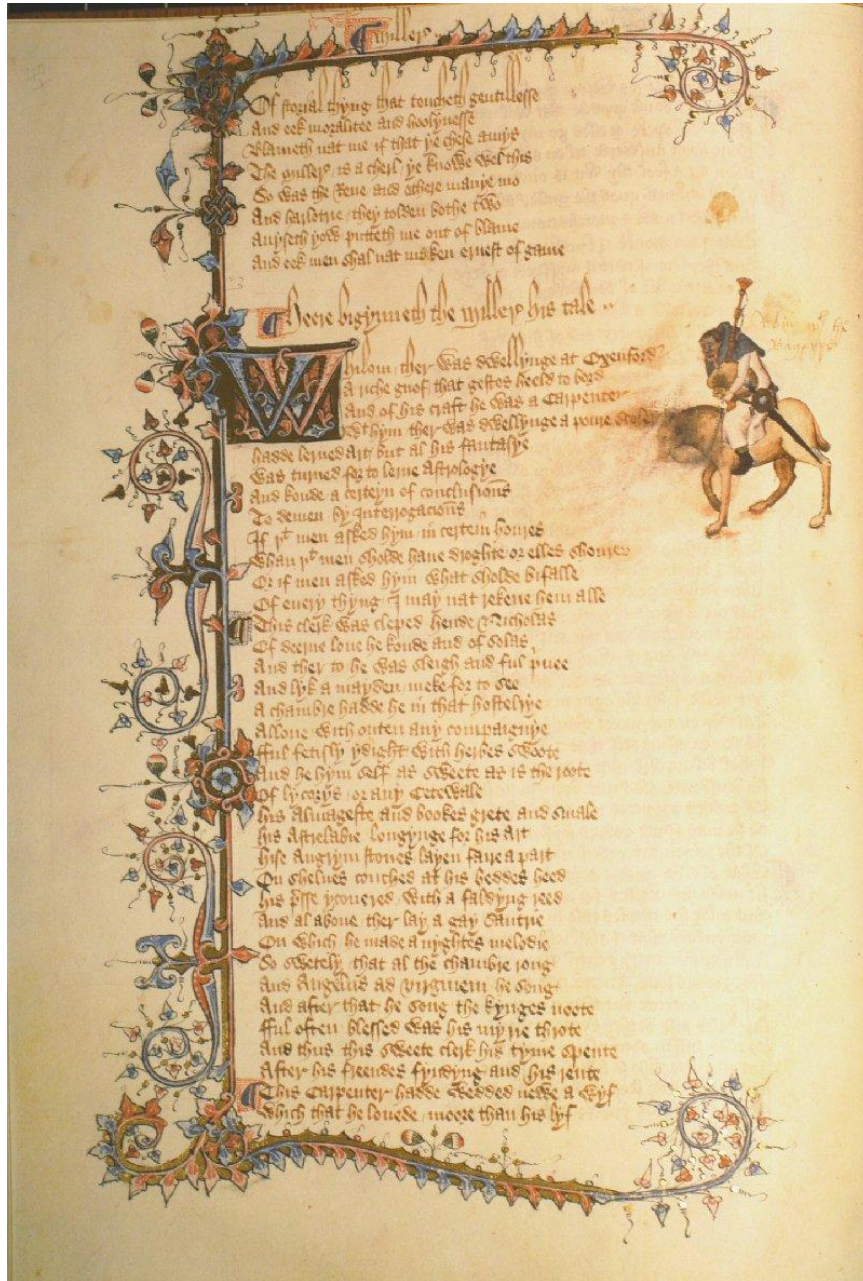


Complication: An entity can have several names.

- Central Connecticut State University, Central, CCSU
- Meriden, Dirty Den
- Sean Combs, Diddy, P. Diddy, Puff, Puff Daddy, Puffy, Sean John
- Yusuf Islam was born Steven Demetre Georgiou; formerly performed under “Cat Stevens” but now performs as “Yusuf.” On 21 September 2004 the plane he was on was forced to land in Bangor, Maine, because “Yousouf Islam” was on the TSA No Fly List.
 - http://en.wikipedia.org/wiki/Cat_Stevens

We have seen that words can be easily modified:

- Example: OCR applied to a text
- Example: Purposely changing spelling as in 1337 or user names from online fan sites
- Example: Transliteration from one language to another as in *Чебышëв* to *Chebyshev*, *Chebychev*, *Tchebycheff*; or Burgess' modifications of Russian in *A Clockwork Orange*
- Example: Change over time as in *jeopard* to *jeopardize*
- In addition, the Chaucer example on next two slides ...



First page of Chaucer's "Miller's Tale" from the Ellesmere Manuscript.

The original was made not long after 1400 (the year Chaucer died), and is roughly 16" x 11".

The characters in the tales are going to the shrine of St. Thomas à Becket in Canterbury Cathedral, and the figure on the horse is the Miller:

He was short-sholdred, brood, a thikke knarre,
Ther was no dore that he nolde heve of harre,
(Lines 551-552 of the "General Prologue")

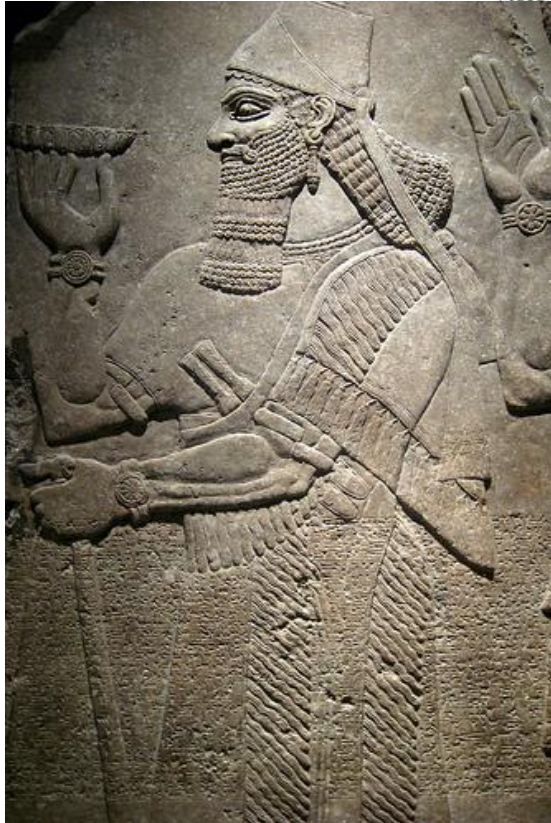
How does this text file version differ from the original Ellesmere Manuscript?

Of storial thyng that toucheth gentillesse,
And eek moralitee and hoolynesse.
Blameth nat me if that ye chese amys.
The millere is a cherl, ye knowe wel this;
* So was the reve eek and othere mo,
And harlotrie they tolden bothe two.
* Avyseth yow, and put me out of blame;
And eek men shal nat maken ernest of game.

From <http://www.luminarium.org/medlit/miller.htm>



Finally, Humans use language in unusual ways ...



Top: My photo of wall decorative pattern in the Miracle Mile Shops, Las Vegas, NV
<http://www.flickr.com/photos/66082566@N00/4281884117/>

Right: My photo of On Kawara's "Oct. 31, 1978" of the Art Institute of Chicago, IL
<http://www.flickr.com/photos/66082566@N00/3619741671/>

Left: King Ashur-nasir-pal at Brooklyn Museum, photo by wallyg
<http://www.flickr.com/photos/wallyg/2440285854/sizes/m/>

A Taste of Text Mining: Analyzing Text with Computers

- **Extracting information from the Web**
 - Power of regular expressions
 - Example used here inspired by William Turkel (Associate Professor of History at the University of Western Ontario)
- **Concordancing**
 - A powerful technique from corpus linguistics
 - Example here uses corpus obtained by Turkel's approach
 - Introduction of some information retrieval (IR) ideas

Extracting Information from the Web

- This is done continuously by *spiders* written by companies like Google to update their search engines.
- Crawling the Web requires sophisticated programming, but *scraping* info from a particular site is not so hard.
- Following example based on ideas given in 6 blog posts at “Digital History Hacks” by William Turkel:
 - <http://digitalhistoryhacks.blogspot.com/2006/01/text-mining-dcb-part-1.html>
 - <http://digitalhistoryhacks.blogspot.com/2006/01/text-mining-dcb-part-2.html>
 - <http://digitalhistoryhacks.blogspot.com/2006/02/text-mining-dcb-part-3.html>
 - <http://digitalhistoryhacks.blogspot.com/2006/02/text-mining-dcb-part-4.html>
 - <http://digitalhistoryhacks.blogspot.com/2006/02/text-mining-dcb-part-5.html>
 - <http://digitalhistoryhacks.blogspot.com/2006/03/text-mining-dcb-part-6.html>

Turkel harvests the online *Dictionary of Canadian Biography (DCB).*

This Web site allows searches using a form.

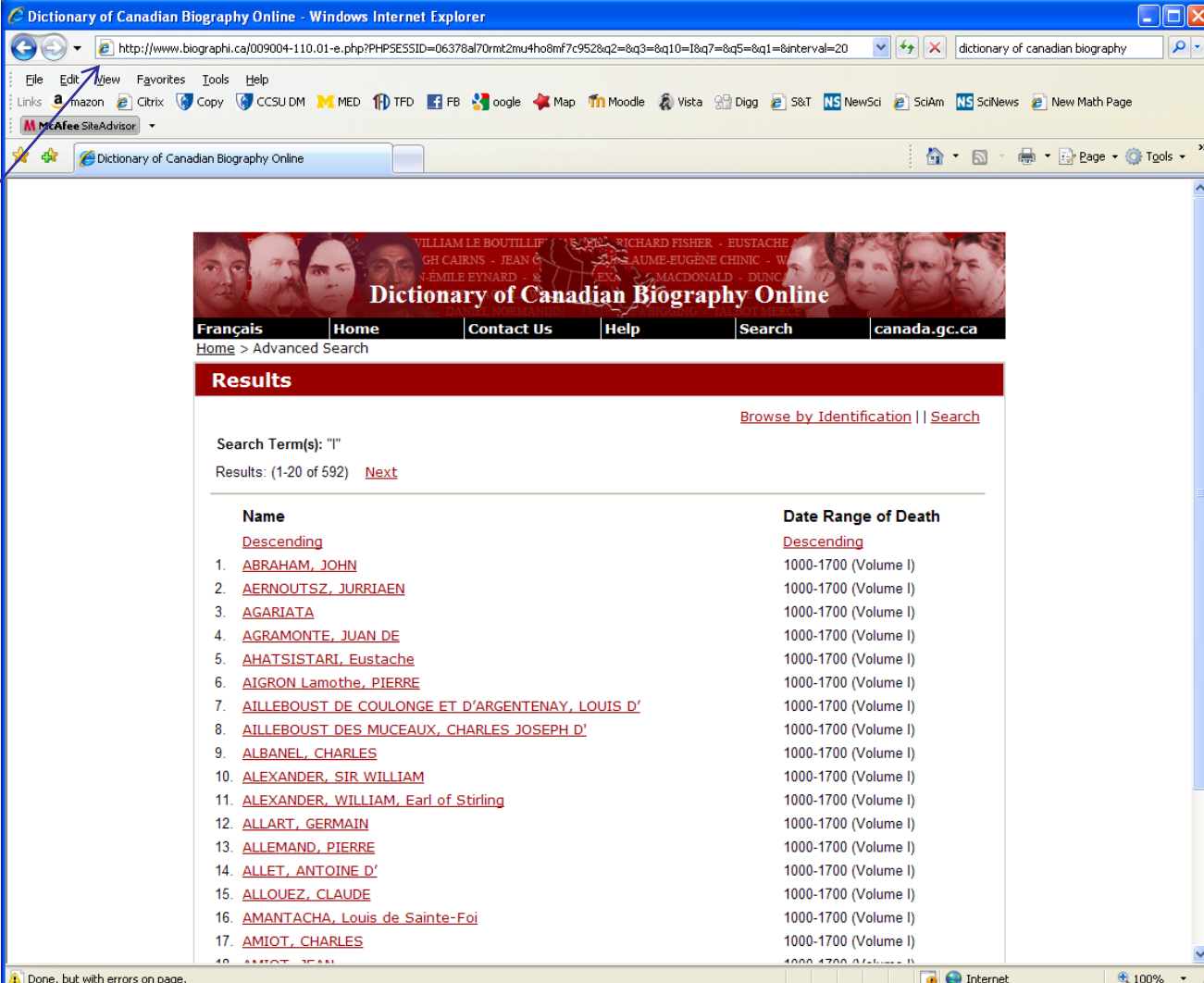
Their terms of use, however, does not forbid downloading all their records.

What is not forbidden must be done!

The screenshot shows a Windows Internet Explorer browser window displaying the Dictionary of Canadian Biography Online website. The address bar shows the URL: <http://www.biographi.ca/009004-100.01-e.php?PHPSESSID=06378a70rmt2mu4ho8mf7c952>. The website features a navigation menu with links for Français, Home, Contact Us, Help, Search, and canada.gc.ca. The main content area is titled "Advanced Search" and includes a sidebar with links to Introduction, Features and Updates, Background, Credits, and Links. The central search form allows users to limit the search by Name, Gender (both), Date Range of Death (All Volumes), Identification, Geographical Location, and Keywords. It also includes a "Number of references per page" dropdown set to 20. A "Keyword Search" section on the right offers a search box and a grid of letters for browsing by first letter of last name. The footer contains "Date Created:", "Date Modified:", "Top of Page", and "Important Notices" links.

A Browser Trick

Form submission can be automated because:
(1) the queries are shown in the URL box
(2) These queries have patterns



Dictionary of Canadian Biography Online - Windows Internet Explorer

http://www.biographi.ca/009004-110.01-e.php?PHPSESSID=06378a170mt2mu4ho8mf7c952&q2=8q3=8q10=1&q7=8q5=8q1=8interval=20 dictionary of canadian biography

File Edit View Favorites Tools Help

Links Amazon Citrix Copy CCSU DM MED TFD FB Google Map Moodle Vista Digg S&T NS NewSci SciAm NS SciNews New Math Page

McAfee SiteAdvisor

Dictionary of Canadian Biography Online

Franglais Home Contact Us Help Search canada.gc.ca

Home > Advanced Search

Results

[Browse by Identification](#) | [Search](#)

Search Term(s): "I"

Results: (1-20 of 592) [Next](#)

Name	Date Range of Death
Descending	Descending
1. ABRAHAM, JOHN	1000-1700 (Volume I)
2. AERNOUTSZ, JURRIAEN	1000-1700 (Volume I)
3. AGARIATA	1000-1700 (Volume I)
4. AGRAMONTE, JUAN DE	1000-1700 (Volume I)
5. AHATSISTARI, Eustache	1000-1700 (Volume I)
6. AIGRON Lamothe, PIERRE	1000-1700 (Volume I)
7. AILLEBOUST DE COULONGE ET D'ARGENTENAY, LOUIS D'	1000-1700 (Volume I)
8. AILLEBOUST DES MUCEAUX, CHARLES JOSEPH D'	1000-1700 (Volume I)
9. ALBANEL, CHARLES	1000-1700 (Volume I)
10. ALEXANDER, SIR WILLIAM	1000-1700 (Volume I)
11. ALEXANDER, WILLIAM, Earl of Stirling	1000-1700 (Volume I)
12. ALLART, GERMAIN	1000-1700 (Volume I)
13. ALLEMAND, PIERRE	1000-1700 (Volume I)
14. ALLET, ANTOINE D'	1000-1700 (Volume I)
15. ALLOUEZ, CLAUDE	1000-1700 (Volume I)
16. AMANTACHA, Louis de Sainte-Foi	1000-1700 (Volume I)
17. AMIOT, CHARLES	1000-1700 (Volume I)
18. AMIOT, JEAN	1000-1700 (Volume I)

Done, but with errors on page.

Below are the URLs for the 1st, 2nd, 3rd, and 4th requests for 20 records.

- <http://www.biographi.ca/009004-110.01-e.php?PHPSESSID=06378al70rmt2mu4ho8mf7c952&q2=&q3=&q10=l&q7=&q5=&q1=&interval=20>
- <http://www.biographi.ca/009004-110.01-e.php?q2=&q3=&q10=l&q7=&q5=&q1=&interval=20&sk=21&&PHPSESSID=06378al70rmt2mu4ho8mf7c952>
- <http://www.biographi.ca/009004-110.01-e.php?q2=&q3=&q10=l&q7=&q5=&q1=&interval=20&sk=41&&&PHPSESSID=06378al70rmt2mu4ho8mf7c952>
- <http://www.biographi.ca/009004-110.01-e.php?q2=&q3=&q10=l&q7=&q5=&q1=&interval=20&sk=61&&&&PHPSESSID=06378al70rmt2mu4ho8mf7c952>

Records

Starting Point

Only eight lines of Perl code downloads all 592 DCB biographies prior to 1700.

```
use LWP::Simple;

open (OUT, ">canadian_bio.txt");

$url_part1 = 'http://www.biographi.ca/009004-110.01-e.php?q2=&q3=&q10=I&q7=&q5=&q1=&interval=100&sk=';
$url_part2 = '&&PHPSESSID=s7drhd5m5ac4dem8vgqq2ppgh7';

for ($i = 1; $i < 601; $i += 100) {
    $doc = get "$url_part1$i$url_part2";
    print OUT "$doc\n\n\n";
}

close(OUT);
```

The reason this is so short is that there is a module LWP that has commands to work with the Web.

```
$doc = get "$url_part1$i$url_part2";
```

This line of code queries the DCB Web page, and the returned HTML is stored in the variable `$doc`.

A Small Sample of the Downloaded DCB HTML

```
<td>
<a href="009004-110.01-e.php?&q10=l&sk=1&s=3&PHPSESSID=s7drhd5m5ac4dem8vgqq2ppgh7">Descending</a></td>
</tr>
<tr>
  <td class="td_data">1.</td>
  <td class="td_data"><a href="009004-119.01-e.php?&id_nbr=1&interval=100&&PHPSESSID=s7drhd5m5ac4dem8vgqq2ppgh7">ABRAHAM, JOHN</a></td>
  <td class="td_data">1000-1700 (Volume I)</td>
</tr>
<tr>
  <td class="td_data">2.</td>
  <td class="td_data"><a href="009004-119.01-e.php?&id_nbr=2&interval=100&&PHPSESSID=s7drhd5m5ac4dem8vgqq2ppgh7">AERNOUTSZ, JURRIAEN</a></td>
  <td class="td_data">1000-1700 (Volume I)</td>
</tr>
<tr>
  <td class="td_data">3.</td>
  <td class="td_data"><a href="009004-119.01-e.php?&id_nbr=3&interval=100&&PHPSESSID=s7drhd5m5ac4dem8vgqq2ppgh7">AGARIATA</a></td>
  <td class="td_data">1000-1700 (Volume I)</td>
</tr>
<tr>
  <td class="td_data">4.</td>
  <td class="td_data"><a href="009004-119.01-e.php?&id_nbr=4&interval=100&&PHPSESSID=s7drhd5m5ac4dem8vgqq2ppgh7">AGRAMONTE, JUAN DE</a></td>
  <td class="td_data">1000-1700 (Volume I)</td>
</tr>
```

We want the URLs to each individual Canadian, which is contained in the `` lines above (the URLs are bolded and in red.) These are extracted (with Perl) and then used to download each biography (again with Perl).

The results are still in HTML, but another program can remove the HTML tags.

1672–89.</P>

<P CLASS="ParagraphFormat">ABRAHAM, JOHN, governor of Port Nelson; fl.

1672–89.</P>

<P CLASS="ParagraphFormat"> He joined the HBC about 1672 and served in James Bay 1672–75 and 1676–78 under Governor Charles Bayly, against whom he brought charges of mismanagement. In 1679 Abraham was appointed second to John Nixon, Bayly's successor and although he absconded with an advance of salary at sailing time, he was engaged in 1681 as mate of the <l>Diligence</l> (Capt. Nehemiah Walker) and wintered in James Bay.</P>

However, all the HTML tags are in <>, which can be removed by a program:

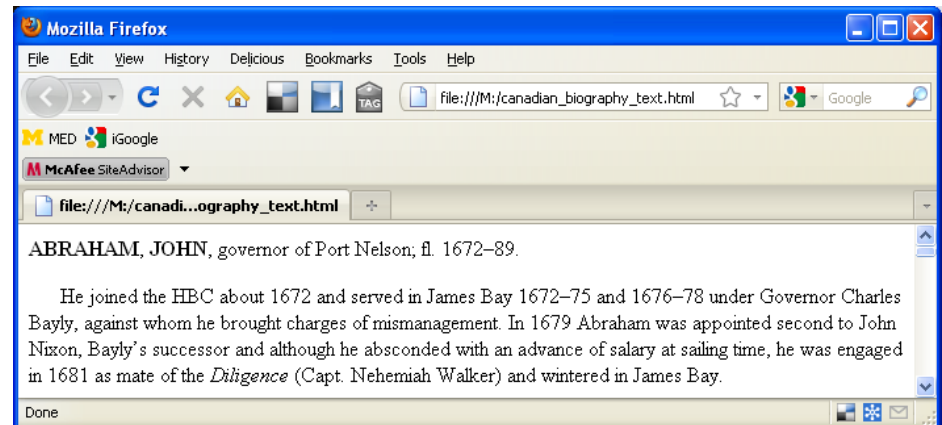
1672–89.

ABRAHAM, JOHN, governor of Port Nelson; fl.

He joined the HBC about 1672 and served in James Bay 1672–75

and 1676–78 under Governor Charles Bayly, against whom he brought charges of mismanagement. In 1679 Abraham was appointed second to John Nixon, Bayly's successor and although he absconded with an advance of salary at sailing time, he was engaged in 1681 as mate of the Diligence (Capt. Nehemiah Walker) and wintered in James Bay.

Note that the top version is still valid HTML:



Now extract dates with a concordancing program.

Key is constructing regular expressions (regexes) to find text a text pattern of interest, which is a 4 digit number starting with 1 in this case.

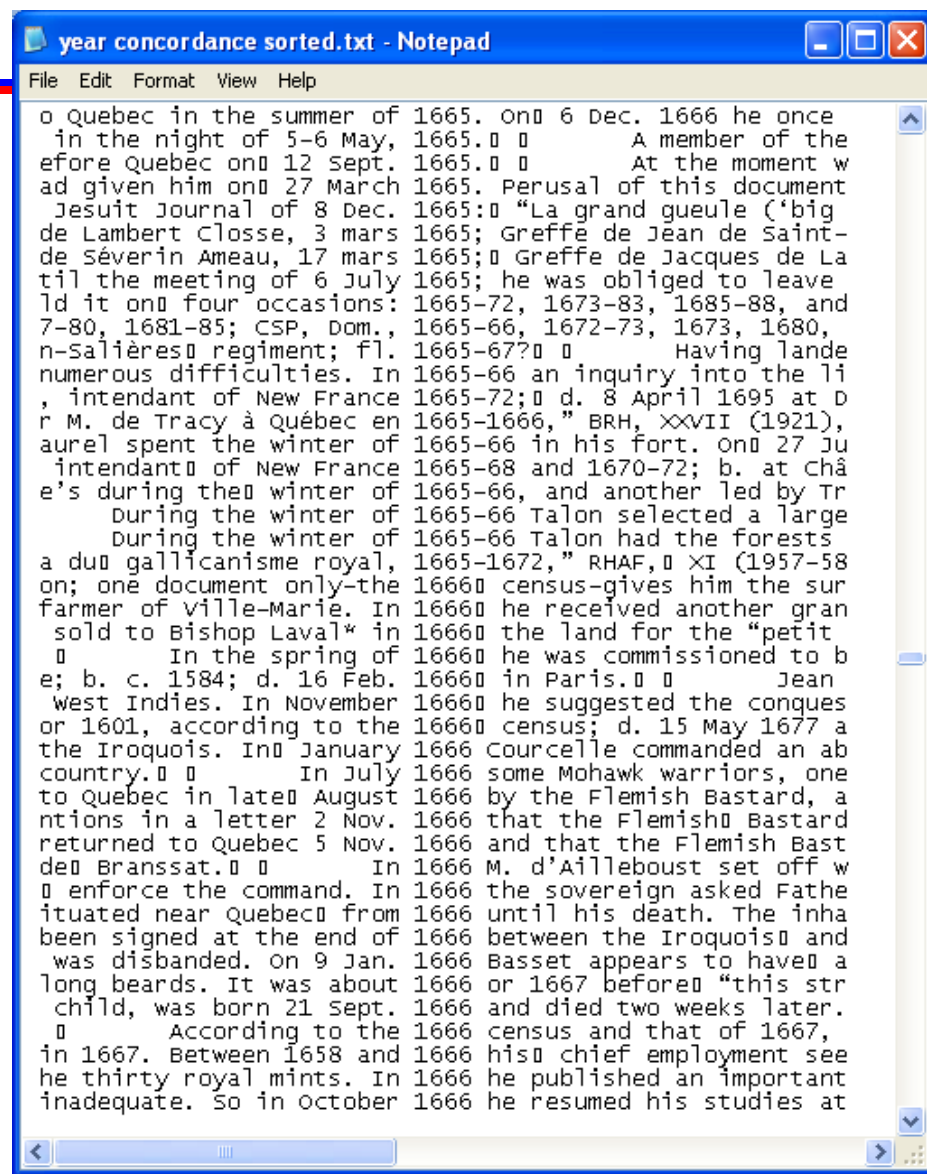
```
$target = '(\D1\d\d\d\D)';
```

\D stands for non-digit

\d stands for digit

At right, all the matches of the regex above are shown after sorting. By looking at this concordance, a variety of patterns emerge.

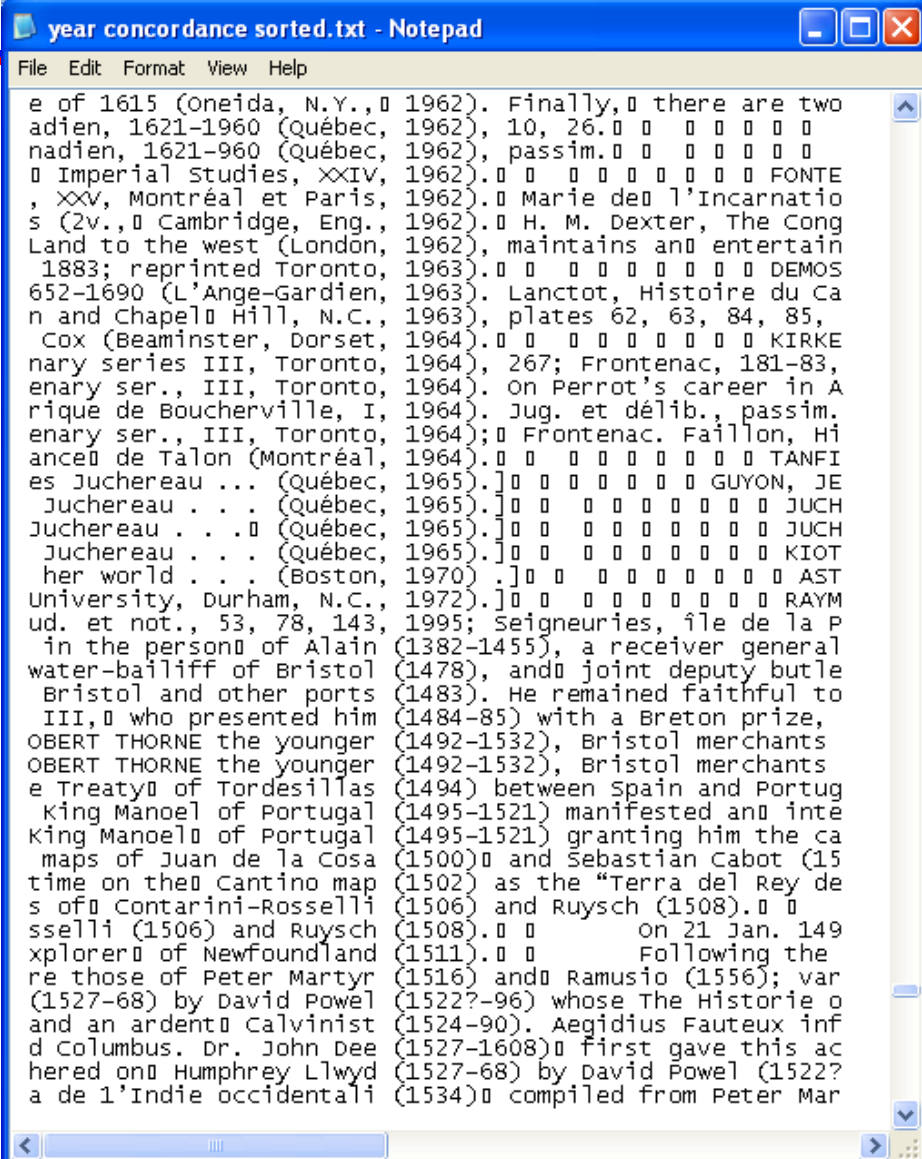
The concordancing program is from Chapter 6 of Bilisoly's PTMP.



Complication: Dates have more than one use.

Dates have many uses and conventions, which complicates their analysis:

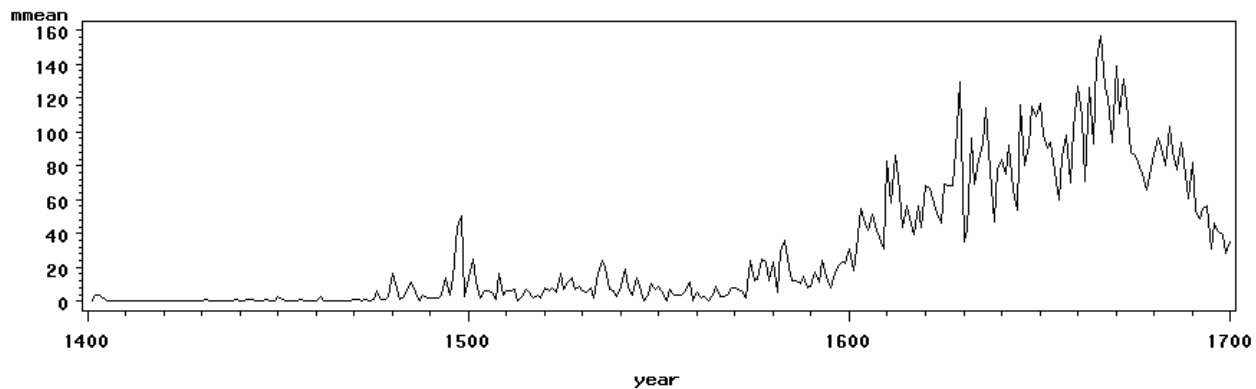
- Although volume 1 of the DCB covers 1000-1700, there are references to modern texts, hence dates in the 20th century appear.
- There are range of dates (e.g., 1495-1521)
- Dates followed by question marks (e.g., 1522?)
- Dates in square brackets (none shown here).
- And so forth ...



```
File Edit Format View Help
e of 1615 (Oneida, N.Y., 1962). Finally, there are two
adien, 1621-1960 (Québec, 1962), 10, 26. 0 0 0 0 0 0 0
nadien, 1621-960 (Québec, 1962), passim. 0 0 0 0 0 0 0
Imperial Studies, XXIV, 1962. 0 0 0 0 0 0 0 FONTE
, XXV, Montréal et Paris, 1962. 0 Marie de l'Incarnatio
s (2v., Cambridge, Eng., 1962). 0 H. M. Dexter, The Cong
Land to the west (London, 1962), maintains and entertain
1883; reprinted Toronto, 1963). 0 0 0 0 0 0 0 DEMOS
652-1690 (L'Ange-Gardien, 1963). Lanctot, Histoire du Ca
n and Chapel Hill, N.C., 1963), plates 62, 63, 84, 85,
Cox (Beaminster, Dorset, 1964). 0 0 0 0 0 0 0 KIRKE
nary series III, Toronto, 1964), 267; Frontenac, 181-83,
enary ser., III, Toronto, 1964). On Perrot's career in A
rique de Boucherville, I, 1964). Jug. et délib., passim.
enary ser., III, Toronto, 1964); Frontenac. Faillon, Hi
ance de Talon (Montréal, 1964). 0 0 0 0 0 0 0 TANFI
es Juchereau ... (Québec, 1965).] 0 0 0 0 0 0 0 GUYON, JE
Juchereau . . . (Québec, 1965).] 0 0 0 0 0 0 0 0 JUCH
Juchereau . . . (Québec, 1965).] 0 0 0 0 0 0 0 0 JUCH
Juchereau . . . (Québec, 1965).] 0 0 0 0 0 0 0 0 KIOT
her world . . . (Boston, 1970).] 0 0 0 0 0 0 0 0 AST
University, Durham, N.C., 1972).] 0 0 0 0 0 0 0 0 RAYM
ud. et not., 53, 78, 143, 1995; Seigneuries, île de la P
in the person of Alain (1382-1455), a receiver general
water-bailiff of Bristol (1478), and joint deputy butle
Bristol and other ports (1483). He remained faithful to
III, who presented him (1484-85) with a Breton prize,
OBERT THORNE the younger (1492-1532), Bristol merchants
OBERT THORNE the younger (1492-1532), Bristol merchants
e Treaty of Tordesillas (1494) between Spain and Portug
King Manoel of Portugal (1495-1521) manifested and inte
King Manoel of Portugal (1495-1521) granting him the ca
maps of Juan de la Cosa (1500) and Sebastian Cabot (15
time on the Cantino map (1502) as the 'Terra del Rey de
s of Contarini-Rosselli (1506) and Ruysch (1508). 0 0
sselli (1506) and Ruysch (1508). 0 0 on 21 Jan. 149
xplorer of Newfoundland (1511). 0 0 Following the
re those of Peter Martyr (1516) and Ramusio (1556); var
(1527-68) by David Powel (1522?-96) whose The Historie o
and an ardent Calvinist (1524-90). Aegidius Fauteux inf
d Columbus. Dr. John Dee (1527-1608) first gave this ac
hered on Humphrey Llwyd (1527-68) by David Powel (1522?
a de l'Indie occidentali (1534) compiled from Peter Mar
```

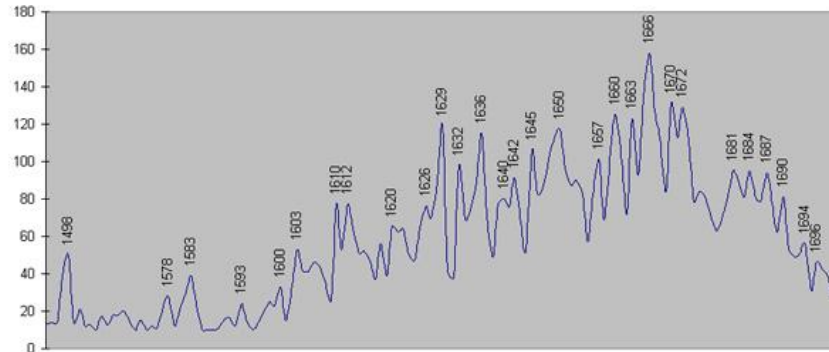
Years Appearing in Volume 1 of the DCB: My Results (top) v. Turkel's (bottom)

Frequency vs. Year based on DCB



Turkel points out that many of the date-peaks do correspond to notable events in early Canadian history. For example, 1498 was Cabot's second voyage, and 1666 was the first census of New France.

Mentions of Each Date in DCB vol 1



Top produced by me using SAS.

Bottom from <http://photos1.blogger.com/blogger/4745/1988/1600/dcbo-vol1-dates.jpg>

The Term-Document Matrix from Information Retrieval (IR)

- One approach to IR is to make a spreadsheet with rows representing terms (e.g., words) and columns representing texts.
- Entries are term counts for each text.
- Example: Such a matrix is given below for 15 of Poe's short stories (texts) and 12 common words (terms). So "The Mystery of Marie Rogêt" contains the word *corpse* 56 times.
- Payoff: Can use mathematical techniques to analyze the resulting matrix. For example, one can compute angles between texts.

Term	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
death	6	0	2	6	9	0	2	0	1	3	2	0	7	7	2
corpse	0	0	0	8	56	0	0	0	0	0	0	0	1	0	4
dead	2	4	0	0	2	2	2	1	0	0	4	0	0	5	1
murder	1	0	0	15	31	0	0	0	1	0	0	0	0	0	3
died	0	0	0	0	3	0	1	0	0	0	0	0	0	0	0
die	1	0	0	1	2	0	0	0	0	0	1	0	2	2	1
deceased	0	0	0	7	9	0	0	0	0	0	0	0	0	0	0
dying	0	1	0	1	1	0	0	0	0	0	0	0	0	5	0
fatal	1	0	0	0	4	0	0	0	0	0	0	0	0	0	1
deadly	1	1	0	0	0	0	0	0	0	0	2	0	0	0	1
decease	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0
murderers	0	0	0	3	9	0	0	0	0	0	0	0	0	0	0

1. The Unparalleled Adventures of One Hans Pfaall
2. The Gold Bug
3. Four Beasts in One
4. The Murders in the Rue Morgue
5. The Mystery of Marie Rogêt
6. The Balloon-Hoax
7. MS. Found in a Bottle
8. The Oval Portrait
9. The Purloined Letter
10. The Thousand-and-Second Tale of Scheherazade
11. A Descent into the Maelström
12. Von Kempelen and his Discovery
13. Mesmeric Revelation
14. The Facts in the Case of M Valdemar
15. The Black Cat

Term-Document Matrix For The DCB

- Here documents are the biographies and terms are years.
- Each person's biography is searched for years.
- Remember that there are complications.
 - Range of years: 1495-1521
 - Years in doubt: 1522?
 - Years for publication references: (London, 1962)

Part of the DCB Name-Year Matrix

	IV	IW	IX	IY	IZ	JA	JB	JC	JD	JE	JF	JG	JH	JI	JJ	JK	JL	JM	JN	JO	JP	JQ
1																						
2	ABRAHAM JOHN	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0
3	AERNOUTSZ (Arentson Aernoutson) JURRIAEN	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	4	0
4	AGARIATA (Agoriata)	0	0	0	0	0	1	0	0	1	0	1	7	0	0	0	0	0	0	0	0	0
5	AGRAMONTE JUAN DE	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	AHATSISTARI Eustache	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	AIGRON Lamothe PIERRE	0	0	0	0	0	2	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
8	AILLEBOUST DE COULONGE	0	1	0	2	2	1	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	AILLEBOUST DES MUCEAUX CHARLES JOSEPH D'	0	0	0	0	0	2	0	1	0	2	0	0	2	0	1	0	1	0	1	0	1
10	ALBANEL CHARLES	1	0	0	2	1	2	2	2	0	0	0	2	1	0	1	0	1	3	1	2	2
11	ALEXANDER SIR WILLIAM	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	ALEXANDER WILLIAM Earl of Stirling	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	ALLART GERMAIN (baptized Théodore)	1	0	0	2	0	1	0	0	1	0	0	1	0	1	1	2	1	0	0	1	1
14	ALLEMAND (Lalemand) PIERRE	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
15	ALLET ANTOINE D'	0	0	0	1	1	2	1	1	0	0	0	0	0	1	0	0	1	0	0	0	0
16	ALLOUEZ CLAUDE	0	1	0	0	1	0	1	0	0	2	1	0	0	4	0	1	0	1	0	0	0
17	AMANTACHA baptized as Louis de Sainte-Foi	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	AMIOT (Amyot) CHARLES	0	0	0	0	0	1	0	1	3	5	2	0	0	0	2	1	0	1	0	0	0
19	AMIOT (Amyot) JEAN	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	AMIOT (Amyot) Villeneuve MATHIEU	0	0	0	0	0	0	1	0	0	1	1	0	0	1	2	0	0	1	0	0	0
21	ANADABIJOU	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	ANDERSON THOMAS	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	ANDIGNÉ DE GRANDFONTAINE	1	0	0	0	0	0	0	0	0	0	2	0	1	1	1	3	2	2	1	0	1
24	ANGIBAUT Champdoré	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	ANNAOTAHA (Annahotaha Anotaha) Étienne	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
26	ANNENRAES	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
27	APRENDESTIGUY (Daprandesteguy Arpentigny)	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0
28	ARGALL (Argoll) SIR SAMUEL	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
29	ASTICOU	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
30	ATIRONTA (Darontal Durantal)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Angles between Canadians in the DCB

```
DCB Angles.nb

angleMatrix = ArcCos[cosMatrix] / Pi * 180 // N;
angleMatrix = Round[angleMatrix * 10 000] / 10 000;

angleMatrix // MatrixForm
```

0.	82.9294	90.	90.	90.	70.2472	90.	86.3556	80.4185	90.	90.	75.9703	59.4581	87.5057	88.4228
82.9294	0.	90.	90.	90.	90.	90.	87.5336	71.9474	90.	90.	80.5538	90.	90.	90.
90.	90.	0.	90.	90.	83.3742	87.7571	66.1029	79.4949	90.	90.	80.2714	90.	88.0132	86.2284
90.	90.	90.	0.	90.	90.	90.	90.	90.	90.	90.	90.	90.	90.	90.
90.	90.	90.	90.	0.	90.	82.3548	86.299	85.1518	90.	88.3933	82.9294	90.	79.8179	77.079
70.2472	90.	83.3742	90.	90.	0.	80.9936	81.2623	78.5275	90.	90.	76.0307	73.1925	77.9941	79.8974
90.	90.	87.7571	90.	82.3548	80.9936	0.	66.7323	66.1229	87.799	88.2899	72.8565	90.	77.7802	76.2323
86.3556	87.5336	66.1029	90.	86.299	81.2623	66.7323	0.	66.8716	87.8642	86.6798	63.5799	87.8183	74.1058	78.3418
80.4185	71.9474	79.4949	90.	85.1518	78.5275	66.1229	66.8716	0.	88.602	88.3705	61.4097	90.	61.4392	76.8925
90.	90.	90.	90.	90.	90.	87.799	87.8642	88.602	0.	58.8005	85.9247	90.	90.	86.299
90.	90.	90.	90.	88.3933	90.	88.2899	86.6798	88.3705	58.8005	0.	84.4545	90.	90.	87.125
75.9703	80.5538	80.2714	90.	82.9294	76.0307	72.8565	63.5799	61.4097	85.9247	84.4545	0.	81.6517	74.8637	83.6791
59.4581	90.	90.	90.	90.	73.1925	90.	87.8183	90.	90.	90.	81.6517	0.	84.0156	86.2193
87.5057	90.	88.0132	90.	79.8179	77.9941	77.7802	74.1058	61.4392	90.	90.	74.8637	84.0156	0.	83.1895
88.4228	90.	86.2284	90.	77.079	79.8974	76.2323	78.3418	76.8925	86.299	87.125	83.6791	86.2193	83.1895	0.
90.	90.	90.	90.	84.1421	90.	90.	86.9788	84.0587	69.2952	65.6643	84.2318	90.	85.8614	90.
85.7215	88.0704	83.1734	90.	90.	78.5737	87.6894	79.4803	81.1602	90.	86.1023	82.8573	88.293	87.9533	71.5651
90.	90.	90.	90.	90.	90.	48.5735	69.2528	79.9835	87.1623	85.5875	77.8132	90.	90.	78.9456
72.8101	86.7228	88.0726	90.	90.	78.3577	79.483	77.2085	74.8767	78.5782	78.9104	67.6675	84.1949	83.0347	76.6977
90.	90.	90.	90.	90.	90.	90.	90.	90.	90.	88.1447	90.	90.	90.	90.
78.6159	90.	90.	90.	90.	83.9827	90.	90.	79.5894	90.	90.	90.	76.3239	79.1066	90.
78.1136	83.1354	87.3129	90.	90.	90.	90.	79.3293	59.6636	86.0431	87.6952	67.5102	87.9802	77.8021	77.6543
90.	90.	90.	90.	90.	90.	90.	90.	90.	90.	77.3025	90.	90.	90.	90.
90.	90.	86.6101	90.	90.	76.3185	57.2129	76.4939	75.2369	90.	90.	79.3058	90.	83.8806	78.3309
90.	90.	90.	90.	90.	90.	64.4348	76.7962	81.4001	90.	90.	77.421	90.	90.	88.1107
58.0188	87.7167	88.6568	90.	90.	84.62	80.8505	88.2315	81.8699	90.	90.	88.3139	60.4334	87.5781	85.4012
90.	90.	90.	90.	90.	90.	87.4425	86.6902	86.7499	90.	71.7309	90.	90.	90.	88.5675
90.	90.	90.	90.	90.	90.	90.	90.	86.77	90.	82.4815	90.	90.	90.	90.
90.	90.	90.	90.	79.4803	90.	88.6082	87.298	88.2315	86.9788	76.9829	90.	90.	90.	87.6603
90.	90.	90.	90.	71.4882	90.	66.0775	86.2397	87.5393	90.	86.7338	86.4149	90.	90.	78.534
79.975	80.9594	71.1207	90.	90.	90.	79.1555	68.5833	71.414	90.	90.	73.1343	90.	85.2198	83.9493
80.5325	90.	90.	90.	90.	86.9955	73.4742	73.9768	78.2706	90.	90.	72.7778	83.2108	87.292	68.9877
90.	90.	90.	90.	90.	90.	90.	90.	90.	90.	90.	90.	90.	90.	90.
90.	90.	90.	90.	90.	90.	90.	90.	90.	90.	84.4812	90.	90.	90.	90.
90.	90.	84.7913	90.	85.5751	79.5389	67.0433	58.8063	84.0115	90.	87.024	81.2602	90.	86.8727	76.0236

Which Canadians are the most alike with respect to years noted in DCB?

- Louis Gaudais-Dupont and Mézy de Saffray
 - 1661, 1662, 1663 (6 times), 1664
 - 1663 (8 times), 1664 (3 times), 1665 (twice)
 - Angle between them is 21.5°
- Thalour du Perron and Sieur de Monts
 - 1662 (twice), 1663 (twice), 1668
 - 1662 (3 times), 1663 (twice)
 - Angle between them is 22.4°

Conclusion of “Extracting Information from the Web”

- Are the pairs of Canadians found on the last slide of interest?
 - Maybe, maybe not
- But the above idea could be done with terms other than years:
 - Canadians and locations
 - Canadians and religious affiliations (Catholic, Protestant)
 - Canadians and a word list (e.g., brave, bravely, bravery, bravest, hero, heroes, heroic, ...)
- Search engines incorporate this approach (along with link analyses)
- Other texts, terms could be used:
 - Artworks and tags (e.g., steve.museum at <http://steve.museum/index.php>)
 - Web pages and tags (e.g., Delicious.com)
 - Books and readers (e.g., Amazon.com’s “Customers who viewed this item also viewed ...”)

Miscellaneous and References

- Following slides address:
 - The Text Encoding Initiative (TEI) and eXtensible Markup Language (XML)
 - Some references for learning to program for non-computer science types
 - Further readings

eXtensible Markup Language (XML) and the Text Encoding Initiative (TEI)

```
<text> <body><div0>
```

```
<head>The following is a Copy of a LETTER sent by the  
Author's Master to the Publisher.  
</head><div1>
```

```
<p n="1"> <name reg="Wheatley, Phillis" type="personal">PHILLIS</name>  
was brought from <name rend="italic" type="geographical">Africa</name> to  
<name type="geographical" key="italic">America</name>, in the Year 1761,  
between Seven and Eight Years of Age. Without any Assistance from School Education,  
and by only what she was taught in the Family, she, in sixteen Months Time from her Arrival,  
attained the English Language, to which she was an utter Stranger before, to such a Degree,  
as to read any, the most difficult Parts of the Sacred Writings, to the great Astonishment of all  
who heard her. </p>
```

No year tags! ☹



```
<p n="2">As to her WRITING, her own Curiosity led her to it;  
and this she learnt in so short a Time, that in the Year 1765, she wrote a Letter to the  
<name reg="Occum, Samson" type="personal">Rev. Mr. OCCOM</name>  
<note resp="editor" type="biographical">Samson Occum (1723-1792) was a converted  
Mohegan Indian who became a Christian minister. He was a friend of Susanna Wheatley,  
Phillis Wheatley's mistress, and a friend and correspondent of Phillis Wheatley.</note>,  
the <name type="ethnological" rend="italic">Indian</name> Minister, while in  
<name type="geographical" rend="italic">England</name>. </p>
```

```
<p n="3">She has a great Inclination to learn the Latin Tongue, and has made some  
Progress in it. This Relation is given by her Master who bought her, and with whom  
she now lives. </p>
```

```
<signed><name reg="Wheatley, John" type="personal">JOHN WHEATLEY</name>. </signed>  
<dateline rend="italic"><name rend="italic" type="geographical">Boston</name>,&br/><date><distinct rend="italic">Nov. </distinct> 14, 1772</date>. </dateline> </div1>
```

```
</div0></body></text> </TEI.2>
```

The DCB used HTML uses tags to inform a Web browser how to display its biographies. It would be useful to have additional tags that encode information for human consumption. A protocol called XML (a form of SGML) was created to do just this. The XML tags are **red** at left.

The TEI Consortium organization (<http://www.tei-c.org/index.xml>) produces standards and encourages the encoding of information in literary and linguistic texts.

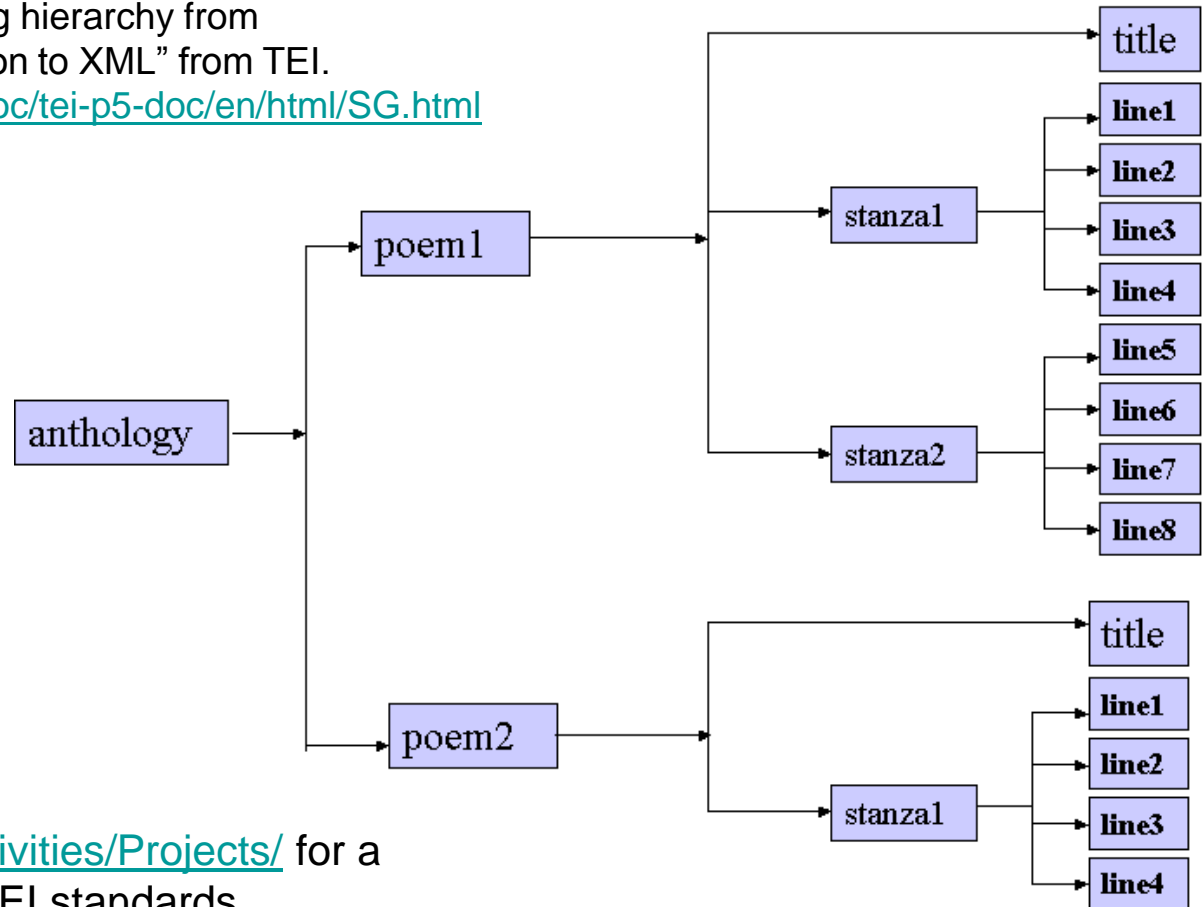
Unfortunately, this kind of tagging is done by humans at present (see example at left), which is labor intensive.

Letter from John Wheatley to the Publisher sent Nov. 14, 1772.
From the Early Americas Digital Archive (<http://www.mith2.umd.edu/eada/>)
Supported by Maryland Institute for Technology in the Humanities (MITH)
http://www.mith2.umd.edu/eada/html/display.php?docs=wheatley_letter.xml&action=show

One of TEI's goals is to specify how to organize information hierarchically.

Example of a tag hierarchy from
"A Gentle Introduction to XML" from TEI.

<http://www.tei-c.org/release/doc/tei-p5-doc/en/html/SG.html>



See <http://www.tei-c.org/Activities/Projects/> for a list of projects adhering to TEI standards.

Learning to Program

- From teaching STAT 527, students vary in their like of programming. However, it's powerful so worth trying if it sounds interesting to you.
- Try “The Programming Historian”
 - Teaches Python
 - By William J. Turkel, Adam Crymble and Alan MacEachern
 - <http://niche-canada.org/programming-historian/>
 - NICHE = Network in Canadian History & Environment
 - NICHE = Nouvelle initiative canadienne en histoire de l'environnement

Also see William Turkel's home page: <http://history.uwo.ca/faculty/turkel/>, which links to his now defunct blog, “Digital History Hacks.”

Learning to Program

- Biologists
 - *Perl for Exploring DNA* by LeBlanc and Dyer
- Linguists
 - *Programming for Linguists: Perl for Language Researchers* by Hammond
 - Beware that some linguists are expert programmers
- Text Mining
 - *Practical Text Mining with Perl* by Bilisoly

Selected Reading

- *Language and Computers: A Practical Introduction to the Computer Analysis of Language*
 - Geoff Barnbrook
- *Practical Text Mining with Perl (PTMP)*
 - Roger Bilisoly
- *Corpus Linguistics: Investigating Language Structure and Use*
 - Biber, Conrad and Reppen
- *Concept Data Analysis: Theory and Applications*
 - Claudio Corpineto and Giovanni Romano (a more technical book)
- *Making the Alphabet Dance: Recreational Wordplay*
 - Ross Eckler
- *Programming for Linguists: Perl for Language Researchers*
 - Michael Hammond
- *Corpora in Applied Linguistics*
 - Susan Hunston
- *Practical English Usage*
 - Michael Swan
- *Beginning Regular Expressions*
 - Andrew Watt
- *Text Mining: Predictive Methods for Analyzing Unstructured Information*
 - Shalom Weiss, Nitin Indurkha, Tong Zhang and Fred Damerau (a more technical book)
- *Geometry and Meaning*
 - Dominic Widdows