Digital Tools Modules: Semantic Data and Text Mining

Dr Matthew Phillpott
Institute of Historical Research
21 June 2012

What is Semantic Data?

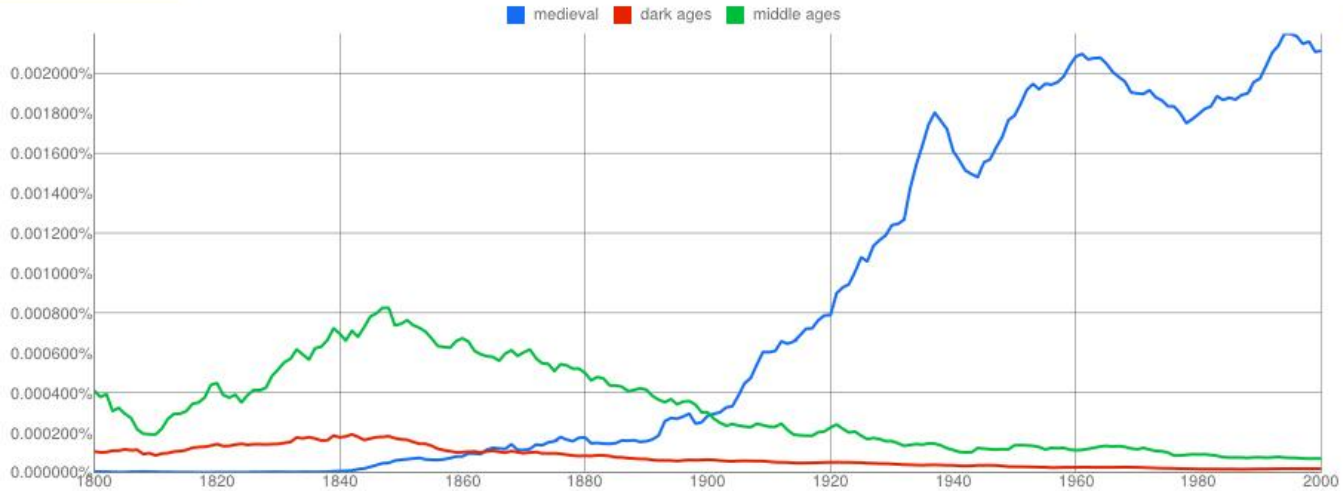
Semantic data is data marked up, however lightly or heavily, in ways which reflect the semantic content of a text, rather than its structure.

What is Text Mining?

An automated method to draw out content based on meaning and context from a large body of text. Enables structure and associations To be revealed in a large body of unstructured data.

Google books Ngram Viewer

Graph these ***case-sensitive*** *comma-separated phrases:* medieval,dark ages,middle ages
between 1800 *and* 2000 *from the corpus* English *with smoothing of* 2.
Search lots of books



+1 0
Tweet 0

Search in Google Books:

1800 - 1814	1815 - 1845	1846 - 1854	1855 - 1942	1943 - 2000	dark ages (English)
1800 - 1916	1917 - 1960	1961 - 1968	1969 - 1994	1995 - 2000	medieval (English)
1800 - 1820	1821 - 1844	1845 - 1852	1853 - 1939	1940 - 2000	middle ages (English)

Run your own experiment! Raw data is available for download here.

History SPOT

Log in Register

Google™ Custom Search

Home About us Resources Documents Help/FAQ Contact us

Latest podcasts

15 November 2011
Does the Digital change anything?

15 November 2011
Great wealth in Argentina, 1810-1930

14 November 2011
Recreational Music-Making and the Fashioning of Political or Diplomati...

11 November 2011
Biology, Brain Theory and History

more

New Podcasts: Does Digital Change anything?

Valerie Johnson and David Thomas discuss

News and Updates from History SPOT

News & Updates

IHRDigProjects Does Digital change anything? https://t.co/88A03FQ9 #News
55 minutes ago · reply · retweet · favorite

IHRDigProjects New podcast on wealth in Argentina between 1810-1930: https://t.co/VxPbdFZ9 #News
4 days ago · reply · retweet · favorite

IHRDigProjects Music-making at the court of Elizabeth I - new podcast from History SPOT: https://t.co/XreNjGgI #News
5 days ago · reply · retweet · favorite

IHRDigProjects What does Biology, Brain Theory and

twitter Join the conversation

Welcome to History SPOT (Seminar Podcasts and Online Training)

Podcasts | Research handbooks | Research training | Collaboration

History SPOT presents podcasts from the internationally renowned research seminar programmes hosted by the Institute of Historical Research (IHR) in an interactive and collaborative space. The 'Seminars' section of the site holds our searchable archive of these podcasts from 2009 to the present, alongside discussion forums and other additional content. This section of the site is designed not simply to provide access to this archive but to allow you to comment on and interact with it. If you have a question or would like to start up or enter a debate or discussion about the topic you can do so here with other historians.

- Archived podcasts from 2009 to the present selected from over 50 history research seminar programmes held at the

Module Index

1. Introduction
2. Creating and gathering data
3. Regular Expressions and Scripting
4. Text Analysis Techniques
5. Named Entry Recognition
6. Taking things Further
7. Further Reading

Digital Tools Home

Return to Digital Tools home page

Settings

- Course administration
  - Question bank
- Switch role to...
  - Return to my normal role
- My profile settings
- Site administration

Topic outline

Text Mining for Historians

3 Regular Expressions and Scripting

Duration: 1.5 hours

Author: **Dr Mark Merry**

Contents and Learning Outcomes (click here)

1.

This section of the course introduces you to regular expressions. Understanding these is essential to being able to use electronic texts for text mining. Reading through the handbook and the exercises that are contained within.

Regular Expressions and Scripting

Previous **Next**

Jump to... Choose...

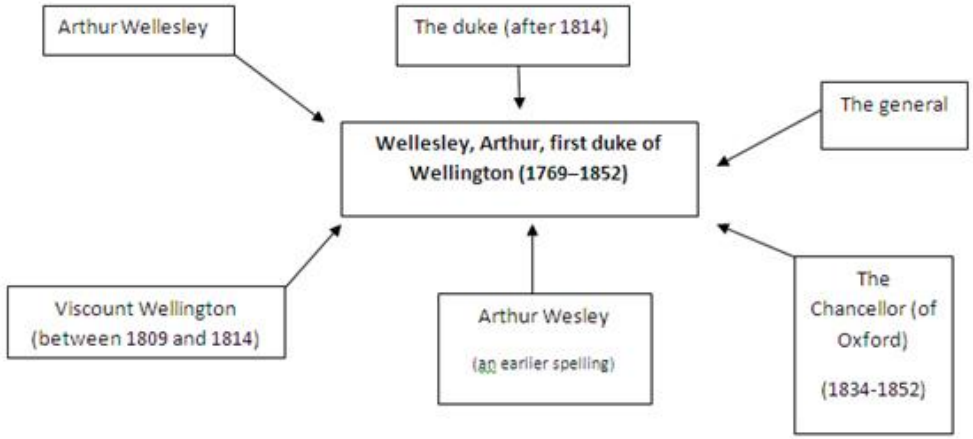
1. Why mark up text?

In a file without markup, the running text will normally be in **plain text** format. To take an extremely simple example, imagine a **text file** that just contains

```
In Wellington I saw a statue of Wellington
```

If we want to search this kind of text file for the string Wellington we can easily find both of the examples by using a find command. But there is no way to distinguish between the person and the place.

Suppose that we want to find all references to the First Duke of Wellington in a text (one that contains, rather than the eight words above, a couple of million words). We don't want references to places called Wellington, or to the duke's son, the Second Duke of Wellington, or any other people called Wellington. But we do want to find all of these references to the same man:



Marking up a text semantically is the most reliable way to be able to return all of the information you want to get from your search. People, places and dates are often the focus of semantic markup, but anything can be marked up - quotations, emotions, economic data or food – anything that a is of interest to the researcher. In this course we'll show you some ways of doing this.

Module Index

1. Introduction
2. XML I
3. XML 2
4. Markup Schemes
5. Project advice

Digital Tools Home

Return to Digital Tools home page

Settings

- Course administration
  - Question bank
- Switch role to...
  - Return to my normal role
- My profile settings
- Site administration

Search

Topic outline

Semantic Data for Historians

2 XML I

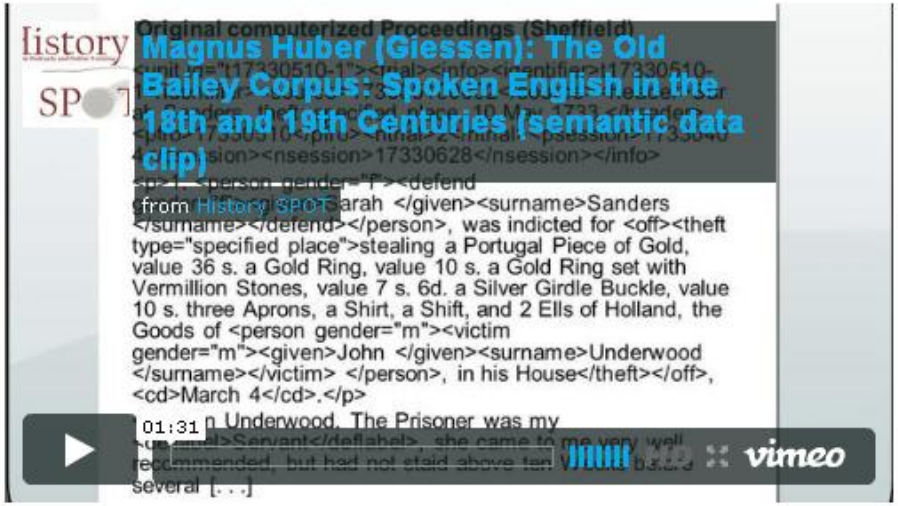
Duration: 2 hours

Author: Jonathan Blaney

Contents and Learning Outcomes (click here)

1.

First listen to this video clip taken from one of the research seminars held at the IHR. In this clip Mangus Huber talks briefly about semantic markup for the Old Bailey criminal trials for 18th and 19th century London.



To listen to the full video or audio podcast click here

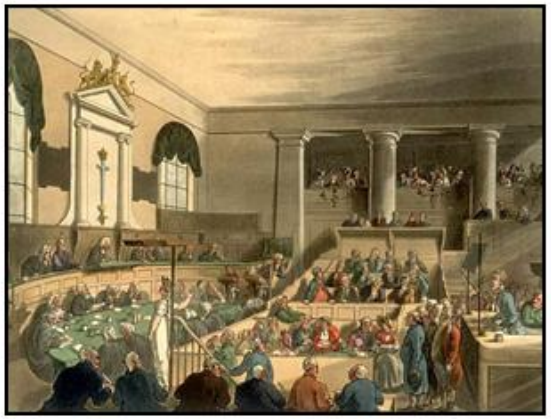
XML Part 1

Table Of Contents

1. What is Text Mining?

1.1 Example: The Old Bailey Proceedings

For example, would you like to know how often the word 'guilty' appears in the Old Bailey trial accounts? The answer is findable using a standard search engine on the Old Bailey Online website (the answer is **182612**). How about how many people were found guilty? The answer is **163261**. What about the number of defendants found guilty of murder? The answer is 1518. These last two figures are not possible to find through the standard search engine as they are an entirely different type of question; we are not looking for how many times the word 'guilty' appears in the proceedings but how many trials resulted in a guilty verdict. We want to discover something meaningful within the body of texts automatically rather than manually checking each and every trial account.



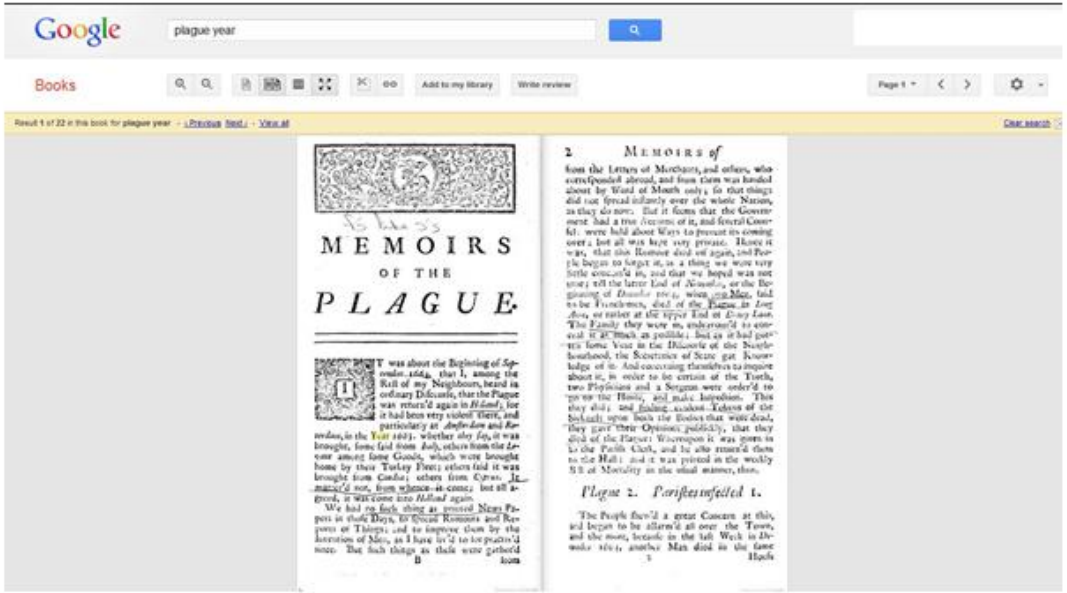
This is a relatively simple example of data mining where the original documents have been marked up and tagged by surname, given name, alias, offence, verdict, and punishment. To calculate those results manually you would have to work your way through 197,745 criminal trial accounts (some 127 million words in total).

This form of text mining, however, is little more than an advanced search engine. Useful, yes but limited. As the creators of the Online Old Bailey themselves admit (and have attempted to redress in a subsequent project):

> 'Analyzing this kind of data by decade, or trial type, or defendant gender etc., can re-enforce the categories, the assumptions, and the prejudices the user brings to each search and those applied by the team that provided the XML markup when the digital archive was first created'.

1. Downloading Data
1.1 Locating books on Google Books

If we continue with the example of the texts available through Google Books, we might try to use the text location tools on offer to find books about the devastating outbreak of plague in London in 1665. Naturally enough we would end up sooner or later at Defoe's (largely fictional) account *A Journal of the Plague Year*.



Daniel Defoe's *A Journal of the Plague Year*, 1722 *(Google Books)*

When viewing a book online in the Google Books collection, assuming that a digital copy is available (and not just a catalogue record, which is sometimes the case) you are presented with a scanned 'facsimile' of the book. This can often tell you quite a lot about the quality of the original copy of the book that has been used in the digitisation – note in this case the scribblings and underlining of several passages. Of particular interest is the warning that appears above the title from the well-meaning owner of the scanned copy, an issue we will be

Donna Baillie talking about her experience on ReScript:
Audio File

Dan Cohen talking about the Ngram Viewer:
Video File

Semantic Data and Text Mining for Historians Training Modules:

http://historyspot.org.uk

Coming Soon