DIGITAL SCHOLARSHIP RESOURCE GUIDE
Samantha Herron
National Digital Initiatives
Junior Fellow
Summer 2017

From Samantha Herron's Digital Scholarship Resource Guide

#TABLE OF CONTENTS

Post 1
#Why Digital Materials Matter

Increasingly, digital archives are emerging and expanding. The Library of Congress' Digital Collections (and therefore its metadata) are always growing, always adding exciting new materials like photographs, newspapers, web archives, audio tracks, maps, and so on. Text, images, and physical objects formerly only available in-person as tangible, hold-able items can now be accessed online as plaintext, digital facsimiles, marc files, .jpgs, .pdfs, hypertext, audio, etc. In addition to making these materials more accessible from all over the world, different digital formats enable exciting, computer-assisted scholarship, projects, and art.

For example:

This is Jane Austen’s *Pride and Prejudice*: https://www.amazon.com/Pride-Prejudice-Jane-Austen/dp/1503290565

This is Jane Austen’s *Pride and Prejudice*: https://books.google.com/books?id=s1gVAAAAYAAJ&printsec=frontcover&dq=pride+and+prejudice&hl=en&sa=X&ved=0ahUKEwiEzPnP4_fUAhWBNT4KHacpAm8Q6AEIKDAA#v=onepage&q=pride%20and%20prejudice&f=false

This is Jane Austen’s *Pride and Prejudice*: http://www.gutenberg.org/files/1342/1342-h/1342-h.htm

This is Jane Austen’s *Pride and Prejudice*: https://www.youtube.com/watch?v=eVHu5-n69qQ

So is this: https://catalog.loc.gov/vwebv/search?searchCode=LCCN&searchArg=00007087&searchType=1&permalink=y

Though all of the above links--a modern day paperback, a digital facsimile, a plaintext copy, an audio recording, and (the catalog record for) a bound copy of the second edition of the book--refer to the same text--Jane Austen’s *Pride and Prejudice*--the kinds of scholarship, arguments, and manipulations we can do using each version depends on its format.

A contemporary paperback copy of *Pride and Prejudice* is likely no help in understanding early 19th century bookbinding practices in London, but the 1813 version of the same may give us some insight. Or, a physical, print copy of the book tells us nothing about word frequency (unless we wanted to count each word up by hand), but a computer could easily return vocabulary density information about a digital text copy. Digital copies do not replace physical texts, but instead open up the text to new kinds of computer-assisted analyses. Digital texts and digital data are the basis for what is broadly termed ‘digital scholarship’, the use of software, code, the Internet, GIS, and so on towards new understandings and visualizations of information.

Example In July 2017, the New York Times covered projects that used data to understand the continued popularity of Jane Austen’s novels, and put forth that the key may have been in the author’s word choice. The authors used a method called “principal components analysis” to graphically represent the presence of naturalism in Austen’s texts. Another study covered by the article found that the author used a higher rate of intensifiers (*very*, *much*, *so*) than her contemporaries and that, in context, this spoke to Austen’s characteristic use of irony.

Computers can be used to see trends and patterns that go unnoticed by the human eye. This is especially helpful for projects like the Jane Austen case study above, where the corpus of interest (the set of texts/other media used for analysis)--in that case, 127 works of early British fiction--would be too labor-intensive, unwieldy, or inappropriate to read one by one for the purposes of the research. Computers can “read” a lot of text very quickly, and tell us information about a corpus that would be impossible to pick up from a close reading of a few books.

Post 2

#Creating Digital Documents

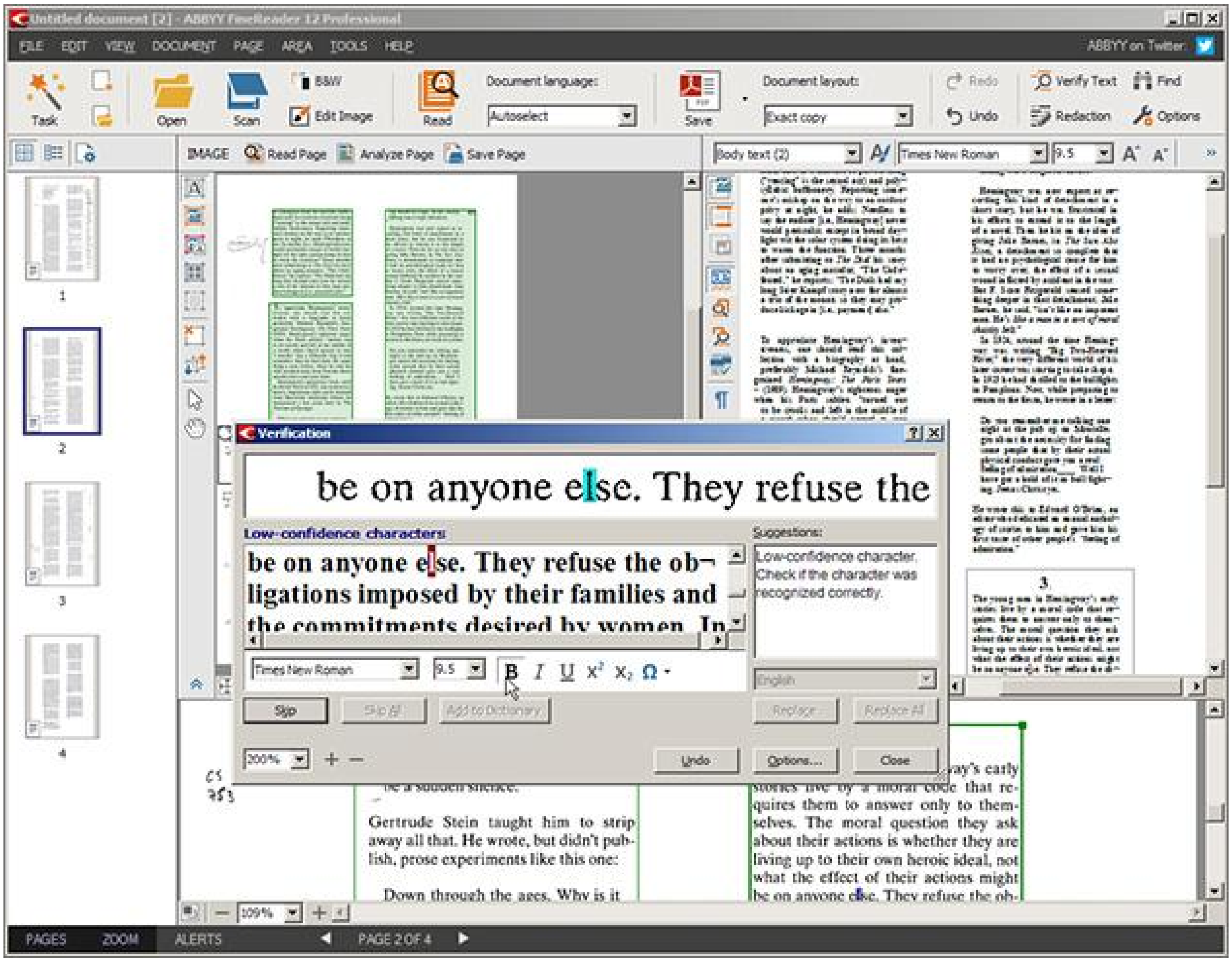
The first step in creating an electronic copy of an analog (non-digital) document is usually scanning it to create a digitized image (for example, a .pdf or a .jpg). Scanning a document is like taking an electronic photograph of it--now it's in a file format that can be saved to a computer, uploaded to the Internet, or shared in an e-mail. In some cases, such as when you are digitizing a film photograph, a high-quality digital image is all you need. But in the case of textual documents, a digital image is often insufficient, or at least inconvenient. In this stage, we only have an image of the text; the text isn't yet in a format that can be searched or manipulated by the computer (think: trying to copy & paste text from a picture you took on your camera--it's not possible).

A book scanner

Optical Character Recognition (OCR) is an automated process that **extracts** text from a digital image of a document to make it readable by a computer. The computer scans through an image of text, attempts to identify the characters (letters, numbers, symbols), and stores them as a separate "layer" of text on the image.

Example Here is a Google Books copy of Alice in Wonderland. Notice that though this ebook is made up of scanned images of a physical copy, you can search the full text contents in the search bar. The OCRed text is "under" this image, and can be accessed if you click on the gear symbol in the upper righthand corner and select "**Plain text**". Notice that you can also download a **.pdf** or an **.epub**.

Though the success of OCR depends on the quality of the software and the quality of the photograph--even sophisticated OCR has trouble navigating images with stray ink blots or faded type--these programs are what allow digital archives users to not only search through catalog metadata, but through the full contents of scanned newspapers (as in Chronicling America) and books (as in Google Books).



Document language: Autoselect

Document layout: Exact copy

Body text (2) Times New Roman 9.5

be on anyone else. They refuse the

be on anyone else. They refuse the ob¬
ligations imposed by their families and
the commitments desired by women. In

Low-confidence character. Check if the character was recognized correctly.

Times New Roman 9.5 English 200% 109%

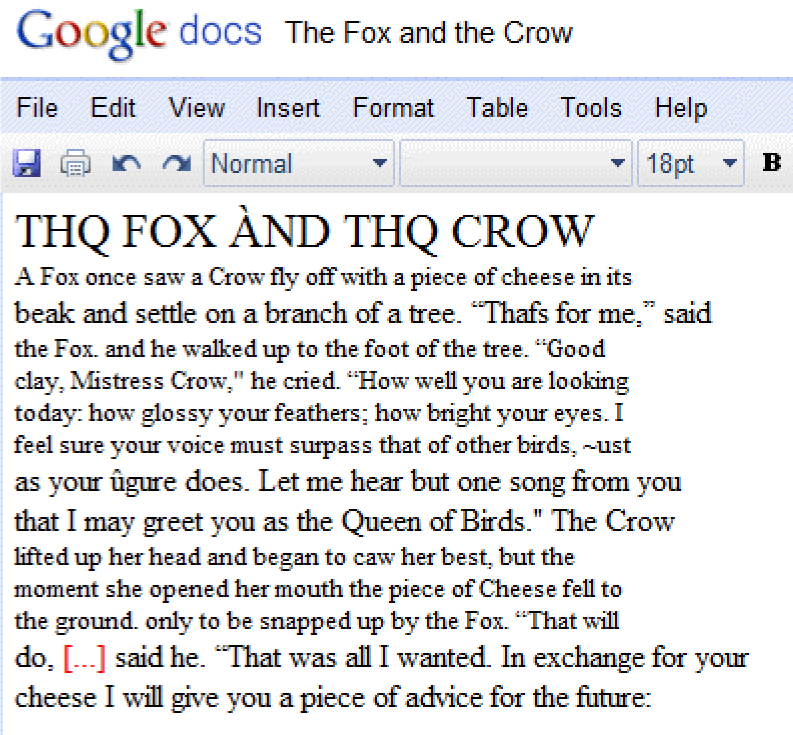
Gertrude Stein taught him to strip away all that. He wrote, but didn't publish, prose experiments like this one:

Down through the ages. Why is it

...ay's early stories live by a moral code that requires them to answer only to themselves. The moral question they ask about their actions is whether they are living up to their own heroic ideal, not what the effect of their actions might be on anyone else. They refuse the ob-

ABBYY FineReader, an OCR software.

As noted, the automated OCR text often needs to be “cleaned” by a human reader. Especially with older, typeset texts that have faded or mildewed or are otherwise irregular, the software may mistake characters or character combinations for others (e.g. the computer might take “rn” to be “m” or “cat” to be “cot” and so on). Though often left “dirty,” OCR that has not been checked through prevents comprehensive searches: if one were searching a set of OCRed texts for every instance of the word “happy,” the computer would not return any of the instances where “happy” had been read as “hoppy” or “hoopy” (and conversely, would inaccurately find where the computer had read “hoppy” to be “happy”). Humans can clean OCR by hand to “train” the computer to interpret characters more accurately (see: machine learning).



THQ FOX ÀND THQ CROW

A Fox once saw a Crow fly off with a piece of cheese in its
beak and settle on a branch of a tree. “Thafs for me,” said
the Fox. and he walked up to the foot of the tree. “Good
clay, Mistress Crow," he cried. “How well you are looking
today: how glossy your feathers; how bright your eyes. I
feel sure your voice must surpass that of other birds, ~ust
as your ûgure does. Let me hear but one song from you
that I may greet you as the Queen of Birds." The Crow
lifted up her head and began to caw her best, but the
moment she opened her mouth the piece of Cheese fell to
the ground. only to be snapped up by the Fox. “That will
do, [...] said he. “That was all I wanted. In exchange for your
cheese I will give you a piece of advice for the future:

In this image of some OCR, we can see some of the errors--the “E”s in the title were interpreted as “Q”s, in the third line, a “t’” was interpreted by the computer as an “f”.

Even with imperfect OCR, digital text is helpful for both close readings and **distant reading**. In addition to more complex computational tasks, digital text allows users to, for instance, find the page number of a quote they remember, or find out if a text ever mentions Christopher Colombus. Text search, enabled by digital text, has changed the way that researchers use database and read documents.

#Metadata + Text Encoding

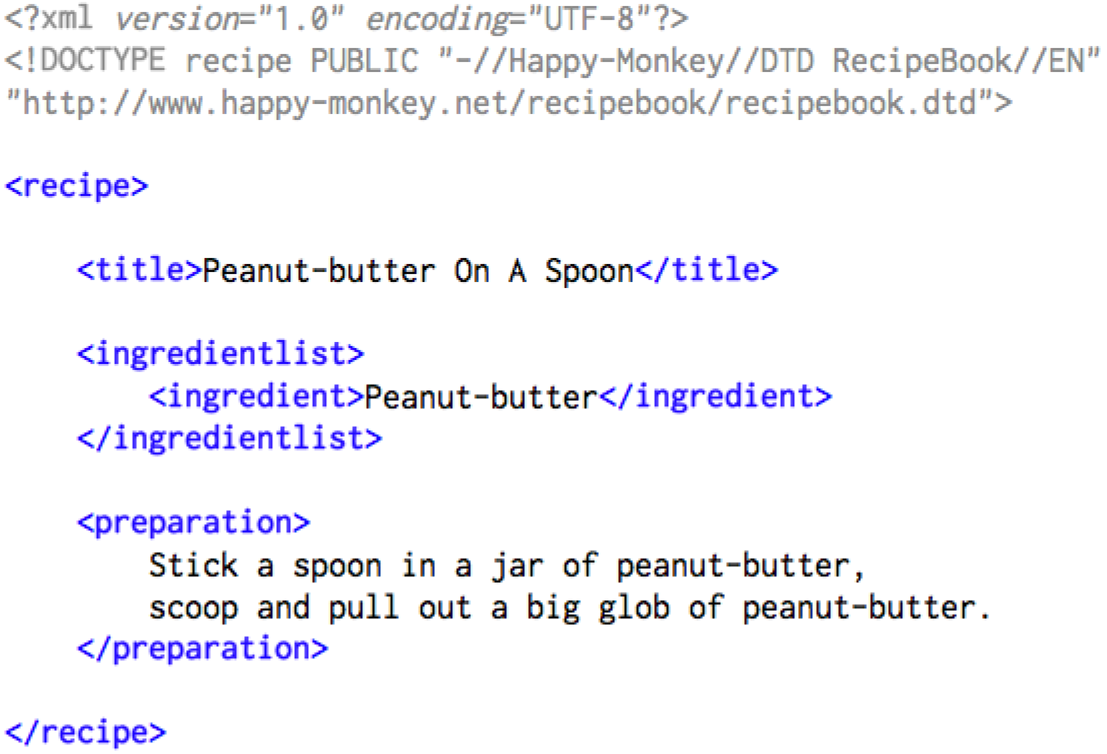
Bibliographic search--locating items in a collections--is one of the foundational tasks of libraries. Computer-searchable library catalogs have revolutionized this task for patrons and staff, enabling users to find more relevant materials more quickly.

Washington, D.C. Jewal Mazique [i.e. Jewel] cataloging in the Library of Congress. Photo by John Collier, Winter 1942. //hdl.loc.gov/loc.pnp/fsa.8d02860

Metadata is “data about data”. Bibliographic metadata is what makes up catalog records, from the time of card catalogs to our present day electronic databases. Every item in a library’s holdings has a bibliographic record made up of this metadata--key descriptors of an item that help users find an item when they need it. For example, metadata about a book might include its title, author, publishing date, ISBN, shelf location, and so on. In a electronic catalog search, this metadata is what allows users to increasingly narrow their results to materials targeted to their needs: Rich, accurate metadata, produced by human catalogers, allow users to find in a library’s holdings, for example, 1. any text material, 2. written in Spanish, 3. about Jorge Luis Borges, 4. between 1990-2000: https://catalog.loc.gov/vwebv/search?searchArg1=Jorge+Luis+Borges&argType1=all&searchCode1=KSUB&searchType=2&combine2=not&searchArg2=&argType2=all&searchCode2=KPNC&combine3=and&searchArg3=&argType3=all&searchCode3=GKEY&year=1517-2017&yearOption=range&fromYear=1990&toYear=2000&location=all&place=all&type=a%3F&language=SPA&recCount=25

Metadata needs to be in a particular format to be read by the computer. A **markup language** is a system for annotating text to give the computer instructions about what each piece of information is. **XML (eXtensible Markup Language)** is one of the most common ways of structuring catalog metadata, because it is legible to both humans and machines.

Here’s an example of some XML:



```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE recipe PUBLIC "-//Happy-Monkey//DTD RecipeBook//EN"
"http://www.happy-monkey.net/recipebook/recipebook.dtd">

<recipe>

    <title>Peanut-butter On A Spoon</title>

    <ingredientlist>
        <ingredient>Peanut-butter</ingredient>
    </ingredientlist>

    <preparation>
        Stick a spoon in a jar of peanut-butter,
        scoop and pull out a big glob of peanut-butter.
    </preparation>

</recipe>
```

XML uses **tags** to label data items. Tags can be embedded inside each other as well. In the above example, <recipe> is the first tag. All of the tags inside between <recipe> and it’s end tag </recipe>, (<title>, <ingredient list>, and <preparation>) are components of <recipe>. Further, <ingredient> is a component of <ingredient list>.

MARC (MAchine Readable Cataloging) standards, developed in the 1960s by Henriette Avram at the Library of Congress, is the international standard data format for the description of items held by libraries. Here are the MARC tags for one of the hits from our Jorge Luis Borges search above: https://catalog.loc.gov/vwebv/staffView?searchId=9361&recPointer=0&recCount=25&bibId=11763921

The three numbers in the left column are “datafields” and the letters are “subfields”. Each field-subfield combination refers to a piece of metadata. For example, 245$a is the title, 245$b is subtitle, 260$ is the place of publication, and so on. The rest of the fields can be found here.

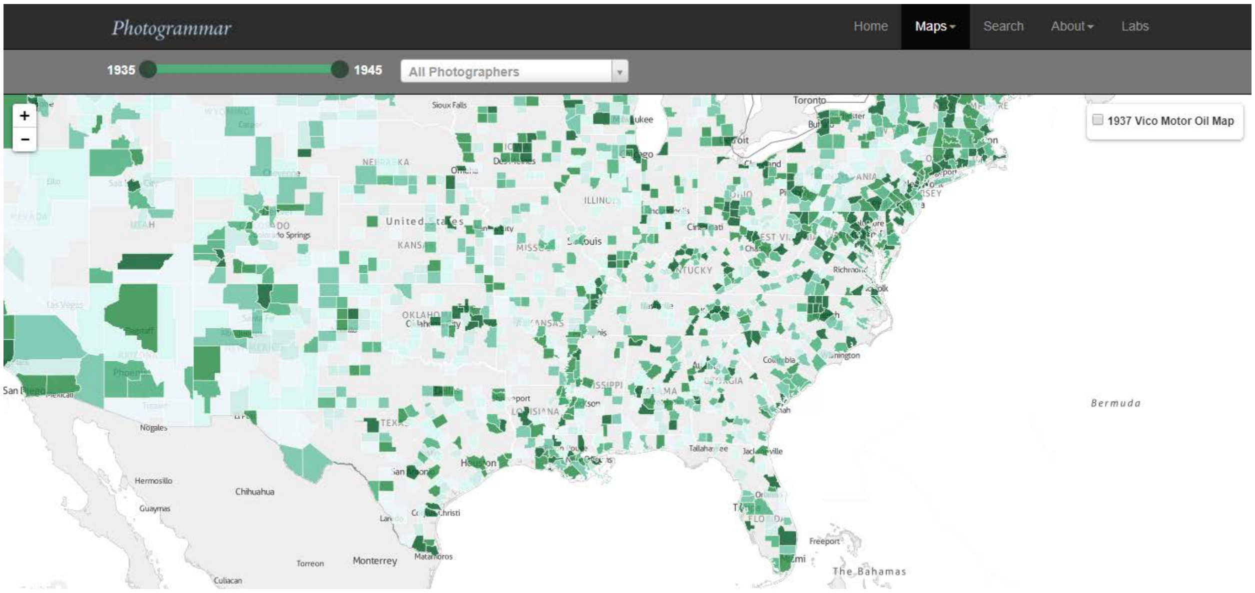
MARCXML is one way of reading and parsing MARC information, popular because it’s an XML schema (and therefore readable by both human and computer). For example, here is the MARCXML file for the same book from above: https://lccn.loc.gov/99228548/marcxml

The datafields and subfields are now XML tags, acting as ‘signposts’ for the computer about what each piece of information means. MARCXML files can be read by humans (provided they know what each datafield means) as well as computers.

The Library of Congress has made available their 2014 Retrospective MARC files for public use: http://www.loc.gov/cds/products/marcDist.php

Examples The Library of Congress’s MARC data could be used for cool visualizations like Ben Schmidt’s visual history of MARC cataloging at the Library of Congress. Matt Miller used the Library’s MARC data to make a dizzying list of every cataloged book in the Library of Congress.

An example of the uses of MARC metadata for non-text materials is Yale University’s Photogrammar, which uses the location information from the Library of Congress’ archive of US Farm Security Administration photos to create an interactive map.



TEI (Text Encoding Initiative) is another important example of xml-style markup. In addition to capturing metadata, TEI guidelines standardize the markup of a text’s contents. Text encoding tells the computer who’s speaking, when a stanza begins and ends, and denotes which parts of text are stage instructions in a play, for example.

Example Here is a TEI file of Shakespeare’s *Macbeth* from the Folger Shakespeare Library. Different tags and attributes (the further specifiers within the tags) describe the speaker, what word they are saying, in what scene, what part of speech the word is, etc. With an encoded text like this, it can easily be manipulated to tell you which character says the most words in the play, which adjective is used most often across all of Shakespeare’s works, and so on. If you were interested in the use of the word ‘lady’ in Macbeth, an un-encoded plaintext version would not allow you to distinguish between references to

“Lady” Macbeth vs. when a character says the word “lady”. TEI versions allow you to do powerful explorations of texts--though good TEI copies take a lot of time to create.

Understanding the various formats in which data is entered and stored allows us to imagine what kinds of digital scholarship is possible with the library data.

Example The Women Writers Project encodes with TEI texts by early modern women writers and includes some text analysis tools.

Post 3

#So now you have digital data...

Great! But what to do?

Regardless of what your data are (sometimes it’s just pictures and documents and notes, sometimes it’s numbers and metadata), storage, organization, and management can get complicated.

Here is an excellent resource list from the CUNY Digital Humanities Resource Guide that covers **cloud** storage, password management, note storage, calendar/contacts, task/to-do lists, citation/reference management, document annotation, backup, conferencing & recording, screencasts, posts, etc.

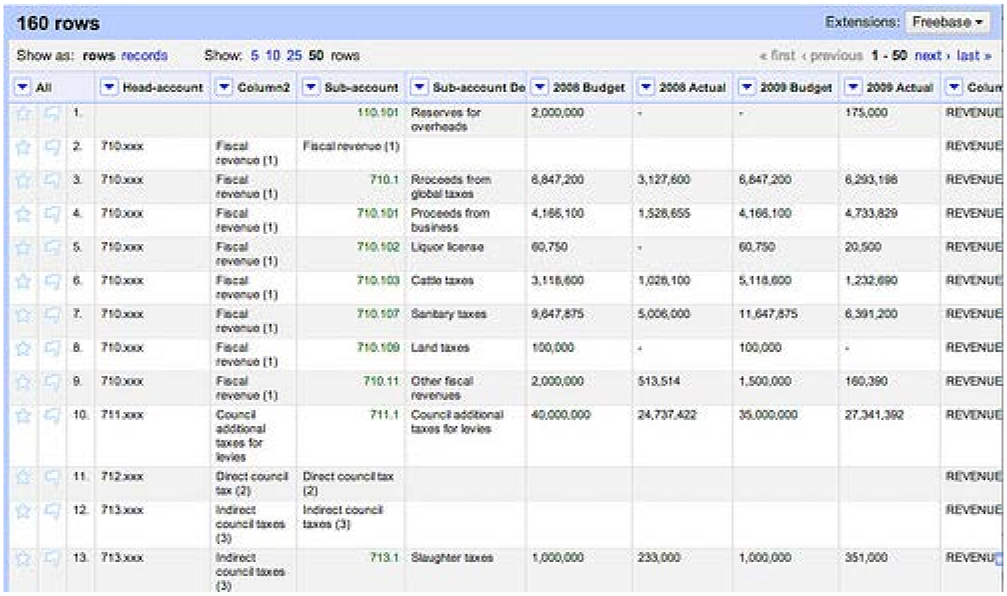
From the above, I will highlight:

- Google Drive and Dropbox, cloud-based secure file storage and sharing services. Both services offer some storage space free, but increased storage costs a monthly fee. With Dropbox, users can save a file to a folder on their computer, and access it on their phone or online. Dropbox folders can be collaborative, shared and synced. Google Drive is a web-based service, available to anyone with a Google account; any file can be uploaded, stored, and shared with others through Drive. Drive will also store Google Documents and Sheets that can be written in browser, and collaborated on in real time.
- Zotero, a citation management service. Zotero allows users to create and organize citations using collections and tags. Zotero can sense bibliographic information in the web browser, and add it to a library with the click of a button. It can generate citations, footnotes, endnotes, and in-text citations in any style, and can integrate with Microsoft Word.

If you have a dataset:

Here are some online courses from School for Data about how to extract, clean, and explore data.

OpenRefine is one popular software for working with and organizing data. It’s like a very fancy Excel sheet.
It looks like this:



160 rows

Extensions: Freebase

Show as: rows records Show: 5 10 25 50 rows

« first ‹ previous 1 - 50 next › last »

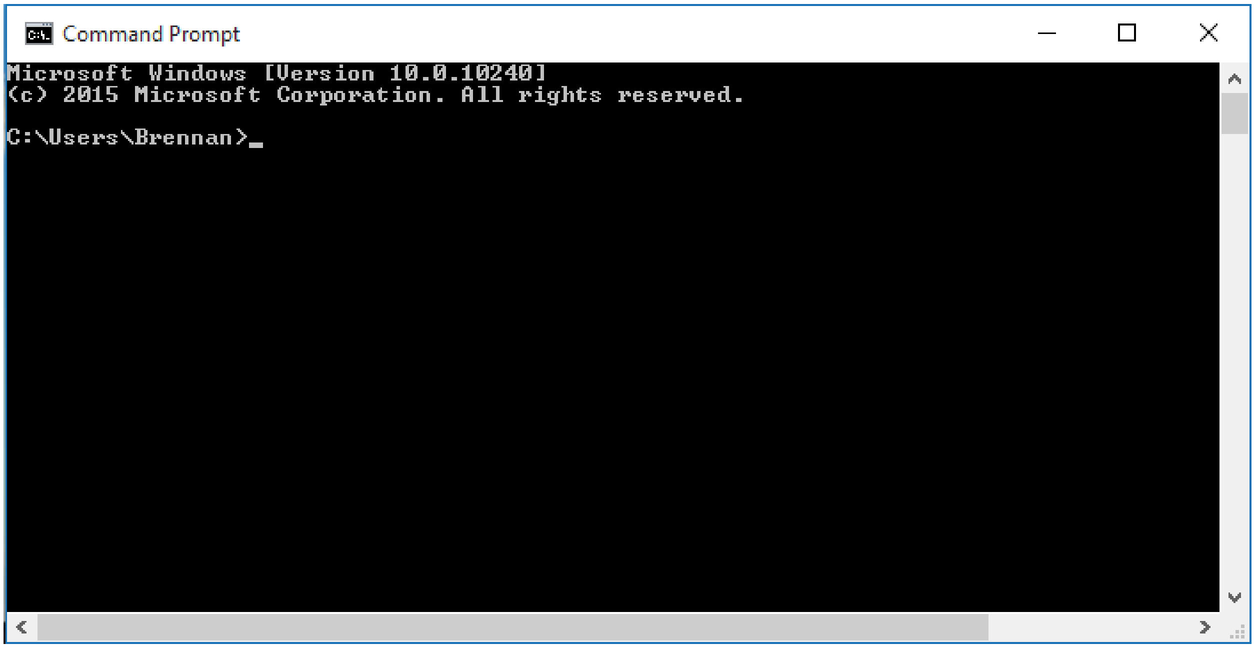
All	Head-account	Column2	Sub-account	Sub-account De	2008 Budget	2008 Actual	2009 Budget	2009 Actual	Colun
1.			110.101	Reserves for overheads	2,000,000	-	-	175,000	REVENUE
2.	710.xxx	Fiscal revenue (1)	Fiscal revenue (1)						REVENUE
3.	710.xxx	Fiscal revenue (1)	710.1	Rroceeds from global taxes	6,847,200	3,127,600	6,847,200	6,293,198	REVENUE
4.	710.xxx	Fiscal revenue (1)	710.101	Proceeds from business	4,166,100	1,528,655	4,166,100	4,733,829	REVENUE
5.	710.xxx	Fiscal revenue (1)	710.102	Liquor license	60,750	-	60,750	20,500	REVENUE
6.	710.xxx	Fiscal revenue (1)	710.103	Cattle taxes	3,118,600	1,028,100	5,118,600	1,232,690	REVENUE
7.	710.xxx	Fiscal revenue (1)	710.107	Sanitary taxes	9,647,875	5,006,000	11,647,875	6,391,200	REVENUE
8.	710.xxx	Fiscal revenue (1)	710.109	Land taxes	100,000	-	100,000	-	REVENUE
9.	710.xxx	Fiscal revenue (1)	710.11	Other fiscal revenues	2,000,000	513,514	1,500,000	160,390	REVENUE
10.	711.xxx	Council additional taxes for levies	711.1	Council additional taxes for levies	40,000,000	24,737,422	35,000,000	27,341,392	REVENUE
11.	712.xxx	Direct council tax (2)	Direct council tax (2)						REVENUE
12.	713.xxx	Indirect council taxes (3)	Indirect council taxes (3)						REVENUE
13.	713.xxx	Indirect council taxes (3)	713.1	Slaughter taxes	1,000,000	233,000	1,000,000	351,000	REVENU

Here is an introduction to OpenRefine from Owen Stephens on behalf of the British Library, 2014. Programming Historian also has a tutorial for cleaning data with OpenRefine.

#Some computer-y basics

A sophisticated **text editing software** is good to have. Unlike a word processor like Microsoft Word, text editors are used to edit **plaintext--**text without other formatting like font, size, page breaks, etc. Text editors are important for writing code and manipulating text. Your computer probably has one preloaded (e.g. Notepad on Windows computers), but there are more robust ones that can be downloaded for free, like Notepad++ for Windows, Text Wrangler for Mac OSX, or Atom for either.

The **command line** is a way of interacting with a computer program with text instructions (commands), instead of point-and-click **GUIs, (graphical user interfaces**). For example, instead of clicking on your Documents folder and scrolling through to find a file, you can type text commands into a **command prompt** to do the same thing. Knowing the basics of the command line helps to understand how a computer thinks, and can be a good introduction to code-ish things for those who have little experience. This Command Line Crash Course from Learn Python the Hard Way gives a quick tutorial on how to use the command line to move through your computer’s file structure.



```
Microsoft Windows [Version 10.0.10240]
(c) 2015 Microsoft Corporation. All rights reserved.

C:\Users\Brennan>_
```

Code Academy has free, interactive lessons in many different coding languages.

Python seems to be the code language of choice for digital scholars (and a lot of other people). It's intuitive to learn and can be used to build a variety of programs.

Post 4

#Text analysis

[Added: In this blog post, Ted Underwood does a way better job than I do at explaining what text mining can do and what some of the obstacles are.]

Clean OCR, good metadata, and richly encoded text open up the possibility for different kinds of computer-assisted text analysis. With instructions from humans ("code"), computers can identify information and patterns across large sets of texts that human researchers would be hard-pressed to discover unaided. For example, computers can find out which words in a corpus are used most and least frequently, which words occur near each other often, what linguistic features are typical of a particular author or genre, or how the mood of a plot changes throughout a novel. Franco Moretti describes this kind of analysis as "**distant reading**", a play on the traditional critical method "close reading". Distant reading implies not the page-by-page study of a few texts, but the aggregation and analysis of large amounts of data.

The strategies of distant reading require scholars to "operationalize" certain concepts--to make abstract ideas quantifiable.

Some important text analysis concepts:

1. **Stylometry -**

Stylometry is the practice of using linguistic study to attribute authorship to an anonymous text. Though some of stylometry's methods and conclusions (both digital and analog) have been disputed, the practice speaks to some of the kinds of evidence researchers hope to surface using text analysis.

One of the early successes of stylometry was in 1964 when Frederick Mosteller and David Wallace used linguistic cues to assign probable authorship to disputed Federalist Papers. Patrick Juola for Scientific American describes it: "[The researchers] showed that the writing style of Alexander Hamilton and James Madison differed in subtle ways. For example, only Madison used the word 'whilst' (Hamilton used 'while' instead). More subtly, while both Hamilton and Madison used the word 'by,' Madison used it much more frequently, enough that you could guess who wrote which papers by looking at how frequently the word was used." Using these methods, they discovered that the disputed papers were likely written by Madison.

Today, computers can perform these kinds of comparative tasks quickly, and keep track of a many different features difficult to track by hand (e.g. not only the relative presence of the word 'by' but the relative presence of 'by' at the beginning vs. the end of a sentence, or the ratio of 'by' to other prepositions, etc.).

Example Stylometry was recently used by researchers to look into the works of Hildegard of Bingen, a female author from the Middle Ages. Because she was not entirely fluent in Latin, she dictated her texts to secretaries who corrected her grammar. Her last collaborator, Guibert of Gembloux, seemed to have made many changes to her dictation while he was secretary. The researchers used digital stylometry methods to display that collaborative works are often styled very differently from works penned by either author individually.

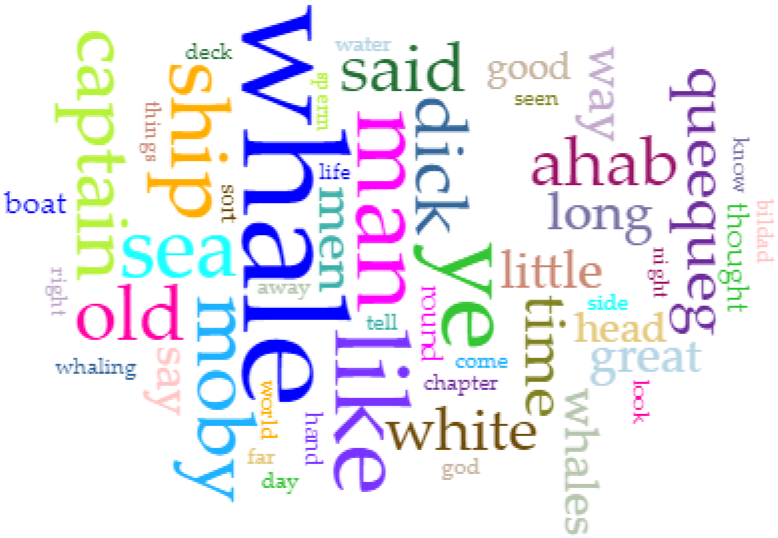
Stylometric methods and assumptions can also be applied beyond author attribution. If stylometry assumes that underlying linguistic features can function as 'fingerprints' for certain authors, linguistic features might also be fingerprints for certain years or genres or national origin of author, and so on. For example, are there linguistically significant identifiers for mystery novels? Can a computer use dialogue to determine if a book was written before 1800? Can computers discover previously unidentified genres? This pamphlet from Stanford Literary Lab gives a good overview of their research into the question of whether computers can determine genre.

2. **Word counts, etc. and topic models -**

Stylometry deals with the attribution and categorization of texts' style. Other distant reading research looks at semantic content, taking into account the meanings of words as opposed to their linguistic role.

Word frequency - One of the simplest kinds of text analysis is word frequency. Computers can count up and rank which words appear most often in a text or set of texts. Though not computationally complicated, term frequency is often an interesting jumping off point for further analysis, and a useful introduction into some of digital humanities' debates. Word frequency is the basis for somewhat more sophisticated analyses like topic modeling, sentiment analysis, and ngrams.

Here’s a **word cloud** for *Moby Dick:*



A word cloud is a simple visualization that uses font size to represent the relative frequency of words--the bigger the font, the more frequently a word is used.

The word cloud is based on this data: (First column is word, second is word count, third is word frequency)

whale	466	0.004317093
like	315	0.0029182069
ye	283	0.002621754
man	251	0.0023253013
ship	227	0.0021029618
sea	216	0.0020010562
old	212	0.0019639996
captain	210	0.0019454713
dick	199	0.0018435656
moby	199	0.0018435656
said	188	0.00174166
ahab	180	0.0016675468
time	169	0.0015656411
little	165	0.0015285845
white	164	0.0015193204
queequeg	162	0.0015007921
long	150	0.0013896223
great	146	0.0013525657
men	138	0.0012784525
way	134	0.001241396
say	132	0.0012228676
whales	132	0.0012228676
head	124	0.0011487544

good	116	0.0010746412
boat	111	0.0010283205
thought	110	0.0010190563
round	106	0.0009819998
sort	101	0.000935679
hand	98	0.0009078866
world	92	0.00085230166
come	90	0.0008337734
sperm	89	0.00082450925
look	88	0.0008152451
whaling	88	0.0008152451
deck	86	0.0007967168
night	84	0.00077818846
chapter	82	0.0007596602
seen	82	0.0007596602
day	78	0.0007226036
know	78	0.0007226036
tell	78	0.0007226036
things	78	0.0007226036
right	77	0.0007133394
water	76	0.0007040753
away	74	0.000685547
bildad	74	0.000685547
far	74	0.000685547
god	74	0.000685547

You’ll notice that this particular word count (completed using **Voyant Tools**) doesn’t include certain **stop words**: ‘fluff’ words like pronouns, articles, conjunctions, and prepositions (e.g. she, that, the, any, but…), keeping only ‘meaning’ words--names, nouns, verbs, adjectives, adverbs.

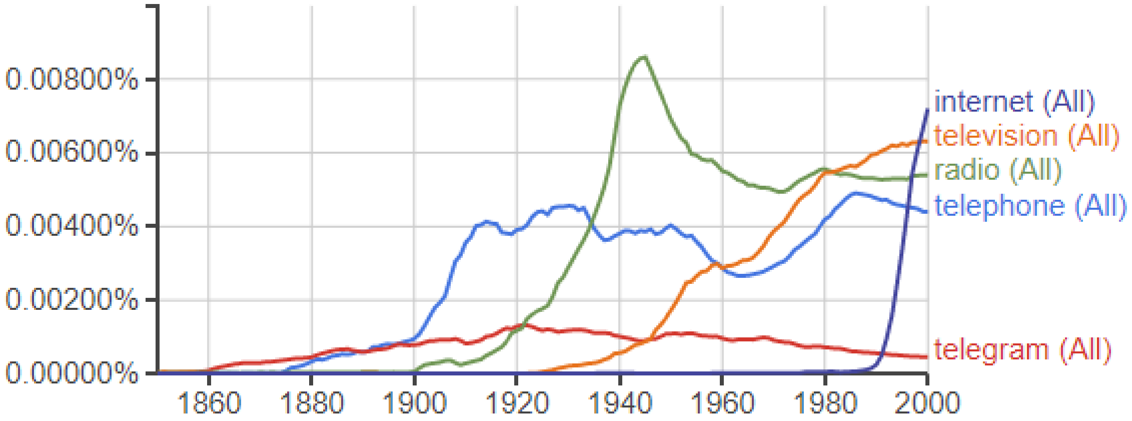
Mostly, this data aligns with what we already know or would assume about *Moby Dick*: that it concerns a whale and an old captain at sea. But with this data, we can ask new questions: Is it significant that ‘whale,’ the most frequent word, is used 150 more times than the runner-up (or even more times if we include the plural ‘whales’ or the verb ‘whaling’)? Why is ‘like’ used so often? Can we safely assume that word count says anything at all about the book’s content or meaning? How does *Moby Dick*’s word frequency compare to Melville’s other works? To the works of his contemporaries?

Voyant is a set of **out-of-the-box tools** that allows you to manipulate and compare texts. Given a corpus (it is preloaded with two corpora: Jane Austen’s novels, and Shakespeare’s plays, but users can also supply their own) Voyant displays word counts and clouds, comparative frequencies over time, concordances, and other visual displays. There are plenty of other more sophisticated and customizable tools available that do similar tasks, but Voyant is one of the most accessible, because it requires no coding by the user.

Here is a link to a list of clean demo corpora to play around with.

Google Books Ngram Viewer is also a powerful example of how word frequencies can be used as a jumping off point for scholarly inquiry. Using Google Books as its massive database, users can track the relative presence of words in books across time.

Here's what a Google ngram looks like:



This ngram compares the (case-insensitive) frequency of 'internet', 'television', 'radio', 'telephone', and 'telegram' across the entire Google Books collection from 1850-2000. This graph (we suppose) reflects a lot of interesting historical information: the birth and quick rise of radio, the birth and quicker rise of the Internet, the birth and steady increase of television, which appears to level out in the 1990s. However, ngrams like this also allow us to ask questions: Does the 1944 peak in frequency of the word 'radio' in books reflect a historical peak in radio popularity? If not, is there some reason why people might be writing more about radios than using them? Or, why was the telegram so infrequently written of in books? Would running this same ngram on a corpus of newspapers rather than books return different results? And so on.

Here are some interesting and silly ngrams from webcomic xkcd.

Word frequency data at both the scale of a single book, and of very many books, asks as many questions as it answers, but can be an interesting jumping off point for beginning to envision texts as data.

Another popular text analysis method is **topic modeling**. A 'topic' is a set of words that frequently colocate in a set of texts (meaning that they occur near each other). In general, topic modeling tool looks through a corpus and spits out clusters of words that are related to each other. So, in a very hypothetical example, if you fed a topic modeling tool the text of every ecology textbook you could find, it might return topics like 'dirt rock soil porous' and 'tree leaf branch root' etc.

The significance of such a tool is more obvious at a large scale. A human can read an article on bananas and state with confidence that the article is about bananas and perhaps that the key words are 'bananas' ‘fruit’ ‘yellow’ ‘potassium’... But when working with a corpus that is say, the text of 100 years of a newspaper, or the text mined from every thread on a subreddit page, the ‘topics’ become more difficult to discern.

Example Robert K. Nelson at the Digital Scholarship Lab at the University of Richmond authored Mining the *Dispatch*, a project that uses topic modeling to look at nearly the full run of a newspaper, the Richmond *Daily Dispatch*, in the early 1860s. For example, one of the topics his model returned was predicted by the words ‘negro years reward boy man named jail delivery give left black paid pay ran color richmond subscriber high apprehension age ranaway free feet delivered.’ Then, by looking at articles where this topic was most prominent, it was determined that this topic most often refers to fugitive slave ads. By tracking the relative presence of this topic through time, one can track the relative presence of fugitive slave ads through time. Other topics identified by the model and named by Nelson include ‘Poetry and Patriotism’, ‘Anti-Northern Diatribes’, ‘Deserters’, ‘Trade’, ‘War Reports’, ‘Death Notices’, ‘Humor’, among others.

Topic models can reveal latent relationships and track hidden trends. Especially for unindexed corpora (like old newspapers, often, that do not have article-level metadata), topic modeling can be used to identify the incidence of certain kinds of content that would take years to tag by hand, if it were possible at all.

A popular topic modeling tool is **MALLET**, for those comfortable working in the **command line**. Programming Historian has a tutorial for getting started using MALLET for topic modeling. If you’re not comfortable in the command line, there is a **GUI (graphical user interface)** tool for implementing MALLET here (meaning you can input files and output topics without entering code yourself), and a blog post from Miriam Posner on how to interpret the output.

SOME TEXT ANALYSIS TOOLS:

- AntConc: Concordance tool.
- DiRT Directory: Digital Research Tools directory.
- From the Page: Crowdsourcing manuscript transcription software
- Google Ngram Viewer: Explore ngrams in Google books corpus.
- Juxta: For textual criticism (identification of textual variants). Look at base and witness texts side by side, locate variations easily. Offers analytic visualizations like heat maps.
- Natural Language Toolkit: Computational linguistics platform for building Python programs that work with language data. “It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active discussion forum.”
- MALLET: (Machine Learning for Language Toolkit) - Java-based package for sophisticated text analysis. Often used for topic modeling. See description above.
- Programming Historian: Peer-reviewed, novice-friendly tutorials for digital tools and techniques.
- R: Statistical software. Often used for text manipulation, but the language is less user-friendly than other coding languages (say, Python).
- Stanford Natural Language Processing Group: Set of Natural Language Processing tools

- Stylo: Suite of stylometric tools for R.
- Voyant: Web-based text explorer (See above).
- WordHoard: WordHoard contains the entire canon of Early Greek epics, as well as all of Chaucer, Shakespeare, and Spenser. The texts are annotated or tagged by 'morphological, lexical, prosodic, and narratological crieteria'. Has a user-interface that allows non-technical users to explore textual date.

WHERE TO GET TEXTS:

- Project Gutenberg
- Hathi Trust
- Internet Archive
- Google Books

Post 5

#Spatial Humanities / GIS / Mapping / Timelines

[Added: George Mason University's Lincoln Mullen has made available the syllabus/lesson plans for a Spatial Humanities workshop here. It's excellent.]

In what has been often termed the '**spatial turn**,' quantitative humanities and social sciences have come to emphasize place and space in their analyses. The mass amounts of geographical and temporal data available has lent itself to new ways of imagining and visualizing global networks. Increasingly sophisticated maps and timelines are increasingly simple to make and use.

Geographic Information Systems (GIS) is the field of techniques and scholarship that combines tabular data with geographical features to query, map, and visualize information. GIS technologies developed in the natural sciences to track things like weather, traffic, and disease patterns, but have moved into the humanities, enabling the spatial mapping of literature and history.

At its core, GIS involves the layering of data onto maps.

Example Professor Anne Kelly Knowles at Middlebury College, researcher Dan Miller, and cartographer Alex Tait developed an interactive map of the civil war battle at Gettysburg. Understanding the technological complications of surveillance during the Civil War, these scholars sought to represent the map of the battle narratively, through time and space, and with generated panoramas of what commanders were likely able to see on the battlefield.

Example ORBIS: The Stanford Geospatial Network Model of the Roman World, a project by Walter Scheidel and Elijah Meeks, is an interactive map that allows users to, for example, input two locations from ancient Rome, select the season, a travel priority (fastest, cheapest, shortest), the modes of travel, etc. and returns the most efficient route.
For more examples, see the list included here under "From the standard map to a varieties of maps" in a section of Lincoln Mullen's Spatial Humanities workshop.

The Library of Congress' digital collections often have location metadata. For example, this collection of African American Photographs Assembled for 1900 Paris Exposition has city metadata for the photographs. Collections like this can lend themselves to mapping projects, such as in the aforementioned Photogrammar.

ArcGIS is one of the most popular and sophisticated GIS tools for spatial analyses, often powering mapping projects like those above. Unfortunately, ArcGIS can be quite complicated or else too expensive for independent scholars/scholars at institutions without a license. Free, open-source options exist, however, like QGIS. Other open-source options have been indexed here.

Google My Maps and Google Earth are both popular free options that allow users to upload geographic data, annotate maps, calculate distances, and display networks. Tutorials are available for both of these, as well as for accessing their APIs. Google Fusion Tables allow users to upload, view, chart, and map data--test it out with one of their example data sets, or look through the gallery.

Sample geographic data sets can be found here on the Resources for Spatial Humanities page of Lincoln Mullen's workshop. This page also includes data repositories, historical maps, syllabi, tutorials, and some further reading.

To transform addresses into latitude/longitude data, you will need a geocoder tool. Geocode is a Google Sheets add-on. Here is a geocoder from GPSVisualizer.

The Spatial Humanities project from UVA's Scholars' Lab also collects 'geospatial scholarship' resources and gives an overview of the 'spatial turn' in different disciplines.

TOOLS FOR SPATIAL ANALYSIS

- Carto: Mapping software, free for some services.
- StoryMapJS: Free. Make maps and timelines enhanced by narrative and visual content. Media like YouTube videos and tweets can be attached to certain times and locations. Example: Washington Post, How the Islamic State is carving out a new country.
- Palladio: Free data viz and mapping.
  Help: Getting Started with Palladio - Miriam Posner
- Google Fusion Tables: Free, see description above.
- Google Earth: Free, see description above.
- Google My Maps: Free, see description above.
- QGIS: Free, see description above.

Data across time can also be visualized with new digital tools.

TOOLS FOR TIMELINES, ETC.

- TimeMapper: Transform Google Spreadsheet into an interactive timeline and map.
- Timeline.js: Use a given Google spreadsheet template to make a sophisticated timeline.

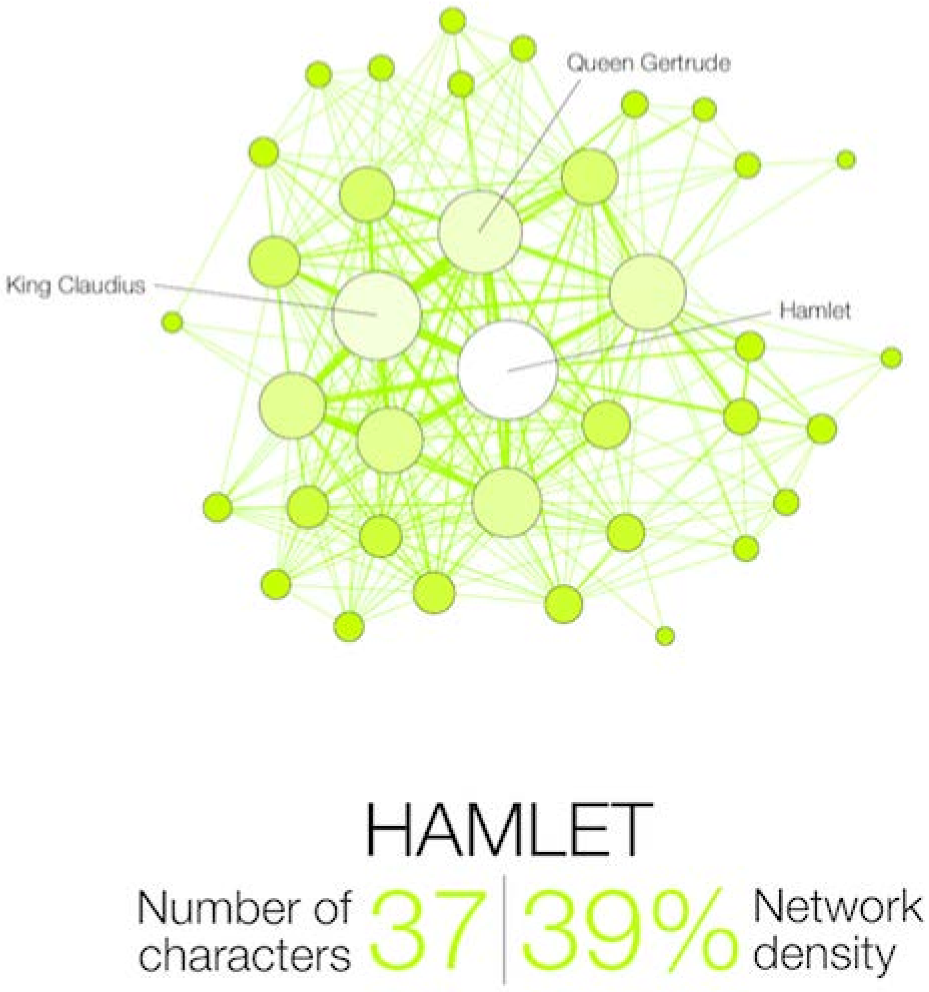
- Neatline: A suite of add-on tools for Omeka, an online exhibit program.
- myHistro: Timeline tool advertised for archiving personal or organizational histories.

Post 6

#Network analysis

Network analysis looks at relationships within a dataset. In the humanities, network analysis can look at kinship ties, social media connections, or conversations between characters in a novel. In network analysis, one looks at vertices (called 'nodes') connected by lines (called 'edges').

Example Kindred Britain networks nearly 30,000 important figures from British culture connected by kinship, marriage, etc. Users can select two individuals and see how they are related and through who and across how much time. Martin Grandjean's network visualizations of Shakespeare's tragedies represents each character as a node connected to each other by an edge if they appear in a scene together.



Gephi is the tool most often used for network analysis projects. It's free, open-source, and well documented. An introduction/tutorial to Gephi by Martin Grandjean can be found here.

This blog post by Elijah Meeks introduces network analysis and representation.

Scott Weingart's article Demystifying Networks, Parts I & II in the Journal of Digital Humanities covers some of the conceptual issues of network analysis.

For more by Elijah Meeks and Scott Weingart, here is a round-up of their posts on network analysis.

NETWORK ANALYSIS TOOLS

Gephi - Free, open-source graph visualization tool.
Here is an introduce to Gephi from Martin Grandjean, creator of the above Shakespeare tragedy visualizations.
Palladio - Free, web-based tool from Stanford. Copy and paste spreadsheet or upload tabular data to quickly make network graphs and geospatial maps.

Post 7

#DIGITAL SCHOLARSHIP PEOPLE + BLOGS

Miriam Posner, UCLA
Bethany Nowviskie, UVA + DLF at CLIR
Ted Underwood, University of Illinois
Dan Cohen, Northeastern
Ben Schmidt, Northeastern
Sapping Attention
Matthew Jockers, University of Nebraska
Matthew Kirschenbaum, University of Maryland
Mark Sample, Davidson College

And more...This is the list of feeds subscribed to by DHNow**. This is the list of blogs from the CUNY Digital Humanities Research Guide.**

#DIGITAL SCHOLARSHIP LABS

Explore these websites for more examples of completed and in-progress digital scholarship projects. [*These are incomplete lists but provide a starting point for learning more.*]

Digital Scholarship Lab, University of Richmond
Digital Scholarship Lab, Brown University
CESTA (Center for Spatial and Textual Analysis), Stanford University
Spatial History group, CESTA, Stanford University

Literary Lab, Stanford University
Text Technologies, Stanford University
Center for Interdisciplinary Research, Stanford University
Roy Rosenzweig Center for History and New Media, George Mason University - Creators of THATCamp and DHNow blog + software Omeka and Zotero.
Scholars' Lab, University of Virginia
Institute for Advanced Technology in the Humanities, University of Virginia
Maryland Institute for Technology in the Humanities, University of Maryland
MIT Hyperstudio, Massachusetts Institute of Technology
Matrix, Michigan State University

DIGITAL SCHOLARSHIP SYLLABI

The CUNY Digital Humanities Resource Guide has compiled many available digital scholarship syllabi and related tools here.

Miriam Posner's Fall 2015 Introduction to Digital Humanities syllabus is online here, and she has also collected other intro syllabi here.

Miriam Posner has also made a Digital Humanities and the Library bibliography.

Digital Art History 101 - Johanna Drucker, Steven Nelson, Todd Presner, Miriam Posner

SOME DIGITAL SCHOLARSHIP BOOKS

Available online:

Burdick, Anne, and Johanna Drucker, Peter Lunenfeld, Todd Presner, Jeffrey Schnapp. *Digital_Humanities*. Cambridge: MIT Press, 2012. Available online: https://mitpress.mit.edu/sites/default/files/titles/content/9780262018470_Open_Access_Edition.pdf

Gold, Matthew K, ed. *Debates in the Digital Humanities 2016*. Minneapolis: University of Minnesota Press, 2016. Available online: http://dhdebates.gc.cuny.edu/

Gold, Matthew K, ed. *Debates in the Digital Humanities 2012*. Minneapolis: University of Minnesota Press, 2012. Available online: http://dhdebates.gc.cuny.edu/

Schreibman, S. Siemens, R., Unsworth, J., eds. *A Companion to Digital Humanities*. Blackwell Companions to Literature and Culture, 2007. Available online: http://www.digitalhumanities.org/companion/

Schreibman, S., Siemens, R., eds. *A Companion to Digital Literary Studies.* Blackwell Companions to Literature and Culture, 2008. Available online: http://www.digitalhumanities.org/companionDLS/