Computer Science and Digital Humanities

Brian W Kernighan
Department of Computer Science
(and, for a while last year, acting director, Center for Digital Humanities)
Princeton University
What is digital humanities?

• "Digital humanities is an area of scholarly activity at the intersection of computing and the disciplines of the humanities." (Wikipedia)

• collecting, curating, analyzing, and presenting data that comes from traditional humanities fields like history, literature, art, music, …
The Valley of the Shadow

The Valley Project details life in two American communities, one Northern and one Southern, from the time of John Brown’s Raid through the era of Reconstruction. In this digital archive you may explore thousands of original letters and diaries, newspapers and speeches, census and church records, left by men and women in Augusta County, Virginia, and Franklin County, Pennsylvania. Giving voice to hundreds of individual people, the Valley Project tells forgotten stories of life during the era of the Civil War.
ORBIS: Network Model of the Roman World

The Shortest journey from Roma to Londinium in January takes 63.5 days, covering 1851 kilometers. Prices in *denarii*, based on the use of a faster sail ship and a civilian river boat (where applicable), and on these road

Per kilogram of wheat (by donkey): 50.63
Per kilogram of wheat (by wagon): 62.37
Per passenger in a carriage: 2293.26
Browse the Collection

Title: Denarius - Sydenham 1281 - Crawford 5464
Moneyer: Gaius Julius Caesar Octavianus (Augustus); Lucius Pinarius Scarpus
Mint Date: 31 B.C.E.
Denomination: Denarius
Material: Silver
Type-Obverse: Head of Jupiter Ammon to right; behind, AVGVR upwards; before, PONTIF upwards; border of dots
Type-Reverse: Victory on globe to right, holding wreath tied with fillet in right hand and palm branch over left shoulder with left hand; on either side, IMP CAESAR DIVI F (AR ligatured)
Authority: Roman Republic

Title: Denarius - Sydenham 1181 - Crawford 517/2
Moneyer: Marcus Antonius; Marcus Barbatus Philippus; Gaius Julius Caesar Octavianus (Augustus)
Mint Date: 41 B.C.E.
Denomination: Denarius
Material: Silver
Type-Obverse: Head of Marcus Antonius to right; around clockwise, M·ANT·IMP·AVGV·III·VIR·R·P·C·M·BARBAT·Q·P (MP and AV ligatured), border of dots
Authority: Roman Republic
The Proceedings of the Old Bailey, 1674-1913

A fully searchable edition of the largest body of texts detailing the lives of non-elite people ever published, containing 197,745 criminal trials held at London's central criminal court. If you are new to this site, you may find the Getting Started and Guide to Searching videos and tutorials helpful.

To search the Proceedings use the boxes on the right or go to the Search Pages.
824. ELIZABETH WARREN otherwise TERRY was indicted for that she, on the 2d day of October, feloniously did take away, with intent to steal, a linen sheet, value 3 s. from her lodging room, the property of Thomas Rummers, against the statute.
Princeton Prosody Archive
Collecting, Curating, and Coding Prosody

Of man's first disobedience and the fruit Of that forbidden tree whose mortal taste Brought death into the world and all our woe Sing Heavenly Muse.

Welcome to the Princeton Prosody Archive, a full-text searchable database of thousands of digitized books published between 1570 and 1923. The Archive collects historical documents and highlights discourses about the study of language, the study of poetry, and where and how these intersect and diverge.
Sylvia Beach
1887-1962
Mapping Expatriate Paris
The Sylvia Beach Lending Library Project

HEMINGWAY
113 Rue N.D. des Champs.
609 rue Froidevaux
Winter 1927
Hotel Bosqui
9 grad
M.O. 13
Suivre

Oct 8
Mammon art.

" 12
Sentimental Education

" 14
Sailing around the world

" 17
Sea and the jungle

" 22
Fear of the Steppe

Sportsman's Sketches

Oct 22
Oct 27
Nov 16
All of these involve

- digitize data, clean it up
- tag: add metadata
- store it in a suitable data structure
- query, analyze, visualize data
- present results
- make it available to others

- using some combination of manual methods, existing tools and new tools
What social network are these women in?


Children, b. Sullivan, N. Y.:
i. Timothy Stanton Brown, b. Aug. 9, 1828. (See below.)
ii. George Cady Brown, b. Oct. 12, 1833. (See below.)
iii. John Jay Brown, b. Feb. 7, 1834. (See No. 452-i.)
   (See below.)
iv. Anna Gertrude Brown, b. Jan. 30, 1836. (See below.)

The foregoing Timothy Stanton Brown, b. Aug. 9, 1828; m. Nov. 12, 1850, Jane A. Haseltine; she d. 1906. Resided in Sullivan, N. Y.
Children:
i. Cady Haseltine Brown, b. Sept. 18, 1857; d. March 2, 1858.
The foregoing Timothy Stanton Brown, b. Aug. 9, 1828; m. Nov. 12, 1850, Jane A. Haseltine; she d. 1906. Resided in Sullivan, N. Y.

Children:

i. Cady Haseltine Brown, b. Sept. 18, 1857; d. March 2, 1858.


DESCENDANTS OF

NICHOLAS CADY

OF WATERTOWN, MASS.

1645---1910.

BY

ORRIN PEER ALLEN,
PALMER, MASS.

PALMER, MASS.:
Published by the Author.
1910.
Mayflower envy?
Relationships

Nicholas 1 Cady 1620??-1700??
James 2 Cady 1665-1690
John 3 Cady 1680-1751
Ebenezer 4 Cady 1714-1779

Elisha 5 Cady 1750-1821
Asa 6 Cady 1779-1854
Eliza Ann 7 Cady 1807-1880
Timothy Stanton Brown 1828
Sarah Brown Cawley 1861
Robert Cawley 1893
Margaret Cawley

Eleazer 5 Cady 1745-1819
Daniel 6 Cady 1773-1859
Elizabeth 7 Cady Stanton 1815-1902
"We hold these truths to be self-evident: that all men and women are created equal."

Children, b. Tolland, Conn.:

i. John, b. June 29, 1730; killed by a cart running over his head Oct. 25, 1737. (Tolland Record.)


iii. Son (unnamed), b. May 22, 1734; d. soon.


Children, b. Tolland, Conn. :

i. John, b. June 29, 1730; killed by a cart running over his head Oct. 25, 1737. (Tolland Record.) 
Son (unnamed), b. May 22, 1734; d. soon. 
Elizabeth, b. June 6, 1736. 
Nahum, b. March 14, 1743. 
Amos, b. Sept. 3, 1747.
Approach  (just like any other data)

• collect data, clean it up
• tag: add metadata
• organize it (XML, TEI, JSON, SQL, …)
• create interface for access (web site, API, …)
• query, analyze, visualize it with existing tools
• present it
• make new tools
• make all of this available to others
Some existing tools…

- Voyant: textual analysis
- Mallet: classification, topic modeling
- Gephi, Cytoscape: network graph visualization
- Palladio, RAW: visualizing network maps, graphs, timelines
- OpenRefine: cleaning up CSV data
- Juxta: compare textual variants

- D3.js: Javascript graphics
- Leaflet, Mapbox: map displays
- NLTK: natural language processing
- Scikit-Learn: machine learning
- R: statistics, plotting
Voyant word cloud of given names

This corpus has 1 document with 1,775 total words and 795 unique word forms. Created 10 seconds ago.
Cytoscape ancestor map
Alternative ancestor map

Children, b. Tolland, Conn.:

i. John, b. June 29, 1730; killed by a cart running over his head Oct. 25, 1737. (Tolland Record.)


iii. Son (unnamed), b. May 22, 1734; d. soon.


Infant Mortality
(1935 people)

Child by first marriage:


Children by second marriage:

ii. Emma, b. 1854, Stoddard, N. H.; d. Sept. 6, 1858.
iii. Lizzie, b. Stoddard, 1856; d. Sept. 6, 1858.

Children by third marriage:


Lessons I learned the hard way
(don't make the same mistakes – make new ones!)

• Do NOT spend all your time cleaning your data
  – it's easier if someone else has done the hard part

• think clearly about what you want to do
  – explore some data?
  – improve old tools?
  – build new ones?
  – ...

• try a sequence of things, not one big idea
  – do small early experiments to test feasibility

• don't expect existing tools to do the whole job
  – or even much of it
Some student projects

• sentiment analysis of headlines from *The Daily Princetonian*
• inferring book similarities and genres from *Goodreads* ratings
• mapping & analyzing the *Great Tang Records on the Western Region*
• visualization of results of AP examinations
• analysis of student tagging of children's ABC books
• analysis of *Correlates of War* datasets
• visualization of product ingredients of cosmetics
• analysis and visualization of religious trends
• visualizing book borrowing patterns from *Shakespeare and Company*
• machine learning and intelligent filtering of online content
• analysis and evaluation of found Haikus
• OCR and analysis of swim-practice datasets
Xuanzang (玄奘)’s journey to the “Western Regions”
Great Tang Records on the Western Regions (大唐西域记)
Visualizing religious trends
Books within 3 nodes of War and Peace
What should they learn?

• most all data has errors
• tagging is straightforward challenging
• data organization is challenging hard
• existing tools often rarely solve the problems
• building new tools is easy hard
• lots of tradeoffs
  – data quality and completeness vs cost
  – hand curation vs tools
  – make vs buy for tools
  – big plans vs finite time and energy
With many thanks to...

Justin, Sung, Annie, Maia, Jessica, Carole, David, Anton, Will, Matt, Jonathan, Juliana, Meagan Wilson, Jean Bauer, Bill Gleason, Cliff Wulfman