

Geoparsing and Spatial Network Analysis in the GAP Projects

Elton Barker, The Open University

Kate Byrne, University of Edinburgh

Leif Isaksen, University of Southampton

Eric Kansa, University of California at Berkeley

Nick Rabinowitz, nickrabinowitz.com

Hestia2 Workshop, 18th July 2013



- 1 Geoparsing ancient texts
 - adapting the Edinburgh Geoparser for GAP
 - online “Unlock Text” service and other projects

- 2 Visualising spatial networks
 - GapVis
 - spatial connection by textual co-occurrence



The GAP projects

- Google Digital Humanities awards, 2010-11 and 2011-12
- Cross-disciplinary and multi-national team
 - humanities, classics, archaeology, natural language processing, graphical interface
- GAP – Google Ancient Places
- GAP2 – the Geographic Annotation Platform
- *GapVis* utility – [nrabinowitz.github.com/gapvis](https://github.com/nrabinowitz/gapvis)
- Prototype “GAP in a box”: *your* text into GapVis



Edinburgh Geoparser

(Claire Grover and Richard Tobin, University of Edinburgh)

1. Geotagging

NLP pipeline to identify placename mentions in text.

2. Georesolution

Select candidate places from suitable gazetteer and assign probability ranking.

Originally designed for modern texts – adapted for GAP.

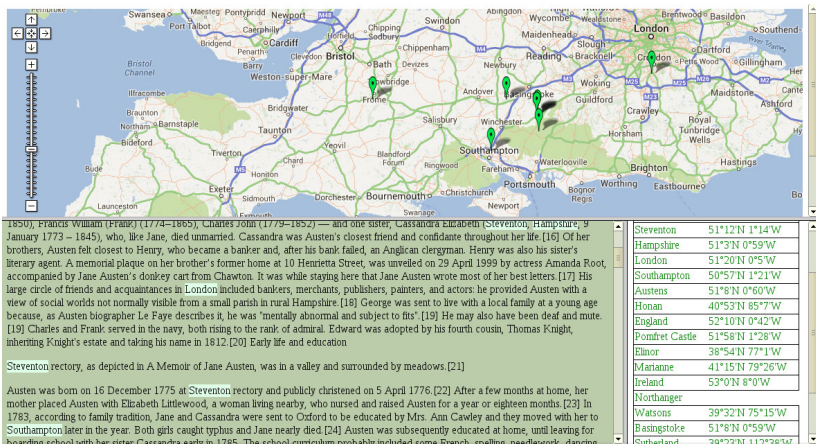


Geoparser demo interface – all candidates



Geoparser interface – only top-ranked location

Today is 196th anniversary of the death of Jane Austen, who spent part of her life in Southampton. 16th Dec 1775 – 18th July 1817.



1850), Francis William (Frank) (1774–1865), Charles John (1779–1852) — and one sister, Cassandra Elizabeth (Steventon, Hampshire, 9 January 1773 – 1845), who, like Jane, died unmarried. Cassandra was Austen's closest friend and confidante throughout her life. [16] Of her brothers, Austen felt closest to Henry, who became a banker and, after his bank failed, an Anglican clergyman. Henry was also his sister's literary agent. A memorial plaque on her brother's former home at 10 Henrietta Street, was unveiled on 29 April 1999 by actress Amanda Root, accompanied by Jane Austen's donkey cart from Chawton. It was while staying here that Jane Austen wrote most of her best letters. [17] His large circle of friends and acquaintances in London included bankers, merchants, publishers, painters, and actors: he provided Austen with a view of social worlds not normally visible from a small parish in rural Hampshire. [18] George was sent to live with a local family at a young age because, as Austen biographer Le Faye describes it, he was 'mentally abnormal and subject to fits'. [19] He may also have been deaf and mute. [19] Charles and Frank served in the navy, both rising to the rank of admiral. Edward was adopted by his fourth cousin, Thomas Knight, inheriting Knight's estate and taking his name in 1812. [20] Early life and education

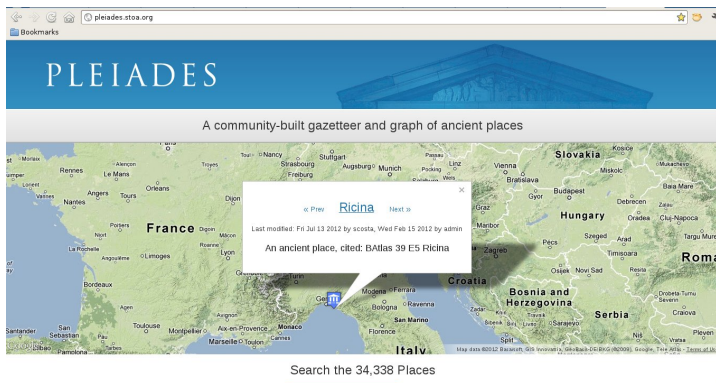
Steventon rectory, as depicted in A Memoir of Jane Austen, was in a valley and surrounded by meadows. [21]

Austen was born on 16 December 1775 at Steventon rectory and publicly christened on 5 April 1776. [22] After a few months at home, her mother placed Austen with Elizabeth Littlewood, a woman living nearby, who nursed and raised Austen for a year or eighteen months. [23] In 1783, according to family tradition, Jane and Cassandra were sent to Oxford to be educated by Mrs. Ann Cawley and they moved with her to Southampton later in the year. Both girls caught typhus and Jane nearly died. [24] Austen was subsequently educated at home, until leaving for boarding school with her sister Cassandra early in 1785. The school curriculum probably included some French, spelling, needlework, dancing,

| | |
|----------------|-----------------|
| Steventon | 51°12'N 1°14'W |
| Hampshire | 51°3'N 0°59'W |
| London | 51°20'N 0°5'W |
| Southampton | 50°57'N 1°21'W |
| Austens | 51°8'N 0°60'W |
| Honan | 40°53'N 85°7'W |
| England | 52°10'N 0°42'W |
| Pomfret Castle | 51°58'N 1°28'W |
| Elinor | 38°54'N 77°1'W |
| Marianne | 41°15'N 79°26'W |
| Ireland | 53°0'N 8°0'W |
| Northanger | |
| Watsons | 39°32'N 75°15'W |
| Basingstoke | 51°8'N 0°59'W |
| Southland | 50°0'N 11°32'W |



Ancient Places Gazetteer – Pleiades



- Pleiades: <http://pleiades.stoa.org/>
- Graph database of ancient places, from Barrington Atlas
- Pleiades+: GAP version linked to Geonames



Adaptations to make Pleiades+

Adding alternative names from Geonames

1. Match Pleiades against Geonames by hand.
2. For matched places, add valid alternative names and extra metadata from Geonames.

eg For *Autricum*, we add *Chartrez*, *Chartres* and *Shartr*



Adaptations to make Pleiades+

Adding alternative names from Geonames

1. Match Pleiades against Geonames by hand.
2. For matched places, add valid alternative names and extra metadata from Geonames.

eg For *Autricum*, we add *Chartrez*, *Chartres* and *Shartr*

Reverse lookup for alternative names at runtime

If geotagger finds candidate placename that's not in Pleiades, we check for alternative names in Geonames.

eg *Egypt* is not in Pleiades; Geonames supplies *Aegyptus* as an alternative, and this **is** in Pleiades.



Distinguishing Places from People

- Geoparser uses various lexicons to help distinguish placenames from other entity categories
 - lexicons are used only at the geotagging step;
lexicon \neq gazetteer
- Ancient texts have different “common” names – *Paris*, *Priam*, *Medea*
- Add lexicon for ancient personal names and drop modern name lexicons





Unlock the potential in your data with our simple web services

Unlock Places

Use [Unlock Places](#) to search for data about place-names. Cross-search different gazetteers and retrieve data in different machine-readable forms for use in web based applications.

The Gazetteers used for searching include Geonames and those available from Ordnance Survey's Open Data

- [Get started](#)
- [API documentation](#)



Find >>



Unlock Text

Use [Unlock Text](#) to extract place-names from documents and find their locations. Unlock Text uses the Places gazetteer cross-search to make ranked best guesses as to the location of place-names.

For use with plain text documents, XML metadata or HTML web pages. The Unlock Text web service works best with short modern texts.

- [Get started](#)
- [API documentation](#)

Lorem ipsum dolor
 sit amet, consectetur
 edinburgh
 adipiscing elit. Vivamus
 eu elit vel urna mattis

Extract >>



Collaboration with Edina – *Unlock Text*

- *Unlock Text* service: <http://unlock.edina.ac.uk/texts/>
- Ongoing enhancements from GAP and other projects
- Online API –
 - ① supply URLs for input text
 - ② collect bundle of output files including full text, tokenised and tagged
 - ③ use output files as desired, eg in graphical interface
- Online documentation not yet fully up to date (see <http://googleancientplaces.wordpress.com/2012/12/21/unlocking-text/>)



Other Adaptations of the Geoparser

- Trading Consequences
<http://tradingconsequences.blogs.edina.ac.uk>
- Digital Exposure of English Place-names (DEEP)
<http://englishplacenames.cerch.kcl.ac.uk>
- GeoDigRef
http://edina.ac.uk/projectGeoDigRef_summary.html
- Botanical records
<http://scargill.inf.ed.ac.uk/clare/botanic.html>
- SYNC3 <http://www.sync3.eu>

Our conclusion

Transfers across genres well but requires adaptation each time.



DEEP – creating historical gazetteer for England

- All 86 volumes of *Survey of English Place-Names*
- Edinburgh's role:
 - convert OCR output to structured data
 - assign lat/long based on EPNS text
 - link to Geonames and to Unlock gazetteer (of UK)
- Gazetteer will be available through *Unlock* services



DEEP – creating historical gazetteer for England

- All 86 volumes of *Survey of English Place-Names*
- Edinburgh's role:
 - convert OCR output to structured data
 - assign lat/long based on EPNS text
 - link to Geonames and to Unlock gazetteer (of UK)
- Gazetteer will be available through *Unlock* services

Application already using it...

Spatial Humanities <http://www.lancaster.ac.uk/spatialhum/> – building Lake District lexicon (modern and historic places) for Edinburgh Geoparser, that will tie to DEEP gazetteer.



- 1 Geoparsing ancient texts
 - adapting the Edinburgh Geoparser for GAP
 - online “Unlock Text” service and other projects

- 2 Visualising spatial networks
 - GapVis
 - spatial connection by textual co-occurrence



User interfaces

- Some users want raw files to use in own systems
- ...but many would prefer graphical interface
- For use in classics teaching, interface is essential
 - clear, usable presentation of content is the goal



GapVis interface

GapVis BETA [Home](#) [About Us](#) [Blog](#)

The Genuine Works of Flavius Josephus, the Jewish historian
By Flavius Josephus

Published 1821 · [View on Google Books](#)

war of that consequence, as to occasion the removal or destruction of six or seven nations of the posterity of Mesraim, with their cities, which Josephus would not have said, if he had not had ancient records to justify his assertions, though those records be now lost.* Gen. ix. 21. t Gen. x. 21. % That the Jews were called Hebrews from their progenitor Heber, our author Josephus here rightly affirms; and not from Abram the Hebrew^a or passenger over Eu- phrates, as many of the moderns suppose. Shem is also called the Father of all the Children of Heber, or of all the city Ninive, and named his subjects Assyrians, who became the most fortunate nation beyond others. Arphaxad named the Arphaxadites, who are now called Chaldeans. Aram had the Aramites, whom the Greeks call Syrians; as Lud founded the Ludites, who are now called Lydians. Of the four sons of Aram, Us founded Trachonitis and Damascus; this country lies between Palestine and Celestria. Ul founded Armenia; and Gather the Bac- trians; and Mesa the Mesaneans: it is now called Charax Spasina. Sala was the son of Arphaxad: and his son was Heber, from whom they originally called the Jews He- brews.*; Heber begat Joctan and Phaleg. He was called Phaleg because he was born at the dispersion of the nations to their several countries, for Phaleg among the He- brews signifies division. Now Joctan, one of the sons of Heber had these sons, Elmo- dad, Saleph, Asermoth, Jera, Adoram, Aizel, Decla, Ebal, Abimeal, Sabeus, Ophir,

Map data ©2012 Basemap, Google, MapIT, MapLink, ORION-AME, SK M&C, Teia Atlas - Terms of Use

Map Satellite

Damascus (Zoom In)

View Details >>

27 references: << prev | next >>

Adam
Phani
Baby
*Asia
Taurus M.
Amanus M.
Babylon
Tarsus
Sala
*Asia
Amath/Ephraim
Damascus
Chana/Chamipolis
Sidon/Col. Aurelia
Hebron
Taris
Asia
Aggyptus
Mambr?
(A) Syria
Damascus
Sidon/Col. Aurelia
Sia
Mambr?
Jerusalem/Hierosolyma Col. Aelia Capitol
Babylon
Mambr?

Shore: Text | Scan

<< previous 12 next >>

Least referenced Most referenced

GapVis (c) 2011 Rick Rabinowitz / Google Ancient Places. Funding by Google

<http://nrabinowitz.github.com/gapvis/>



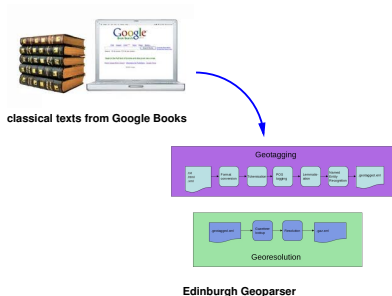
How GapVis Works



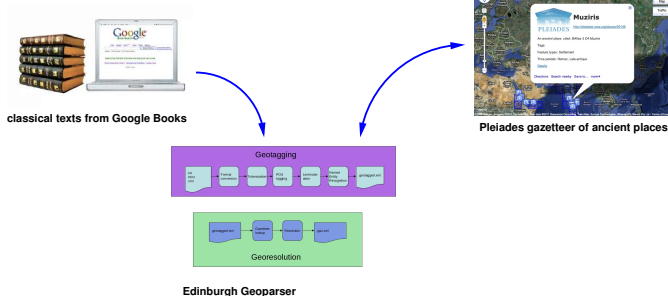
classical texts from Google Books



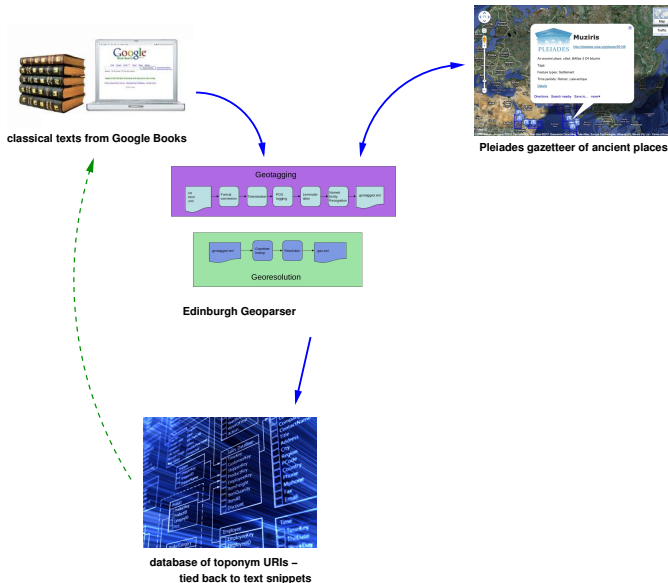
How GapVis Works



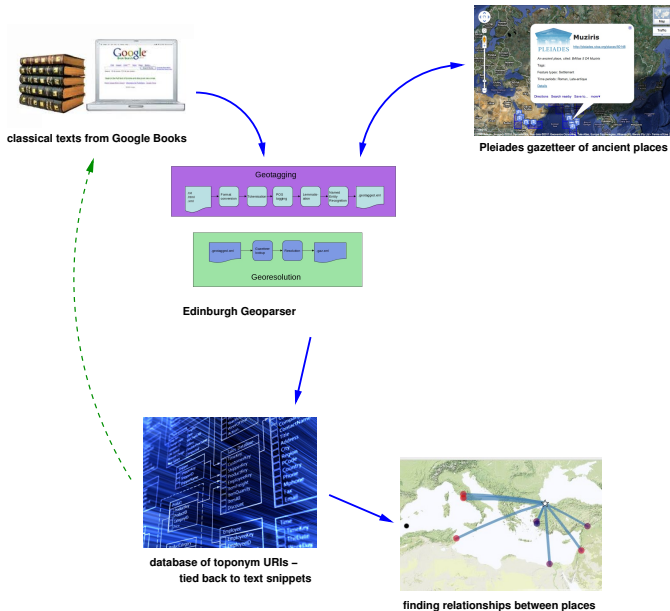
How GapVis Works



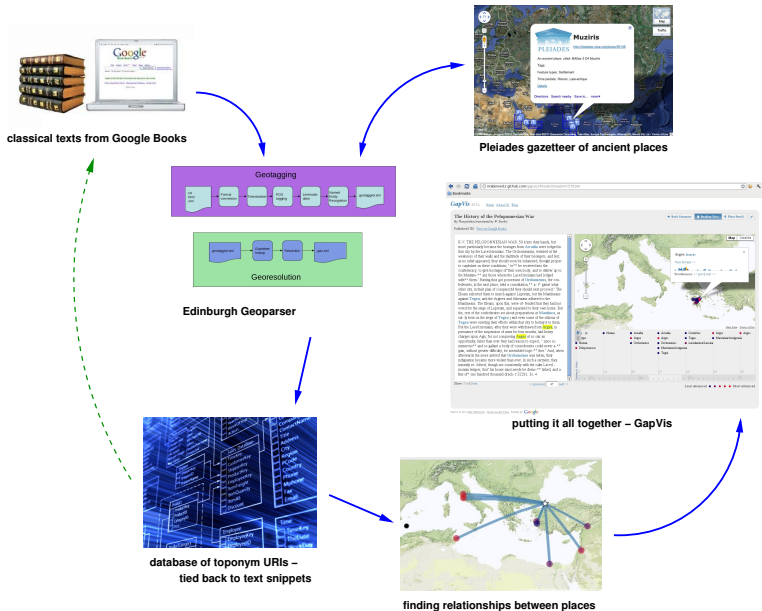
How GapVis Works



How GapVis Works



How GapVis Works



GapVis “views”

- **Book Summary** – overview of places mentioned and where they appear in the narrative
- **Reading View** – for reading the text; with map of places mentioned on current page and “timeline” of places mentioned earlier and later in the book
- **Place Detail** – focus on a particular place, with network map of related places



Thucydides – Book Summary

GapVis BETA [Home](#) · [About Us](#) · [Blog](#)

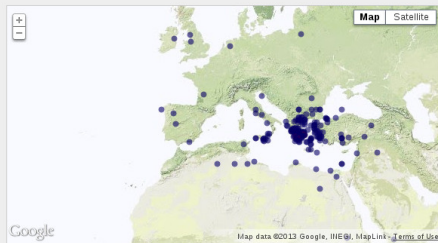
The History of the Peloponnesian War

By Thucydides

Published 1818 · [View on Google Books](#)

The Google Books version of "The History of the Peloponnesian War", by Thucydides, was published in 1818. It references 213 identified ancient places. The place most frequently referenced is [Athens](#), followed by [Hellas](#), [Peloponnesus](#), and [Lacedaemon/Laconia](#).

[Go to Reading View](#)



★ Book Summary

■ Reading View

⌘ Place Detail



Most-Referenced Places



GapVis (c) 2011 Nick Rabinowitz / Google Ancient Places. Funding by Google



19

Thucydides – Place Detail

GapVis BETA [Home](#) · [About Us](#) · [Blog](#)

The History of the Peloponnesian War

By Thucydides

Published 1818 · [View on Google Books](#)

★ Book Summary

▣ Reading View

✦ Place Detail



Attica



42 references

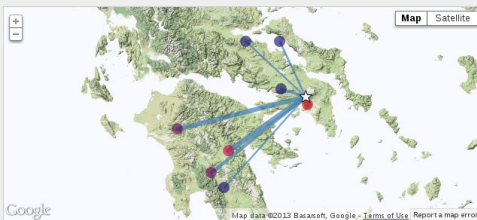
[Report a problem with this record](#)

External Resources

- [Place page on Pleiades](#)
- [Books referencing Attica](#)
- [Pelagios Graph Explorer](#)

Top Related Places

[Athens](#) (22)
[Hellas](#) (15)
[Peloponnesus](#) (13)
[Lacedaemon/Laconia](#) (9)
[Plataea](#) (7)
[Euboea Ins.](#) (5)
[Sparta](#) (4)
[Boeotia](#) (4)



Photos from Flickr · [View on Flickr >>](#)

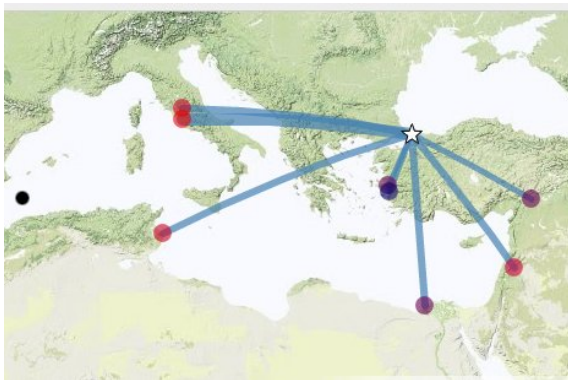


GapVis (c) 2011 Nick Rabinowitz / [Google Ancient Places](#). Funding by Google



Spatial connection by textual co-occurrence

- Frequency counts across entire text for each placename
- Distribution chart of occurrences through text
- Co-occurrence within page (or 500-word window)
- Network is of highest frequency co-mentions



Spatial connection without spatial proximity

- Sometimes nearby places mentioned together – expected?
- Does co-mention with remote locations suggest non-proximity connection – political, military, trading?
- Is systematic analysis feasible?
- “Noise” levels for high/low frequency places?



Low frequency – Thermopylae

Thermopylae

3 references

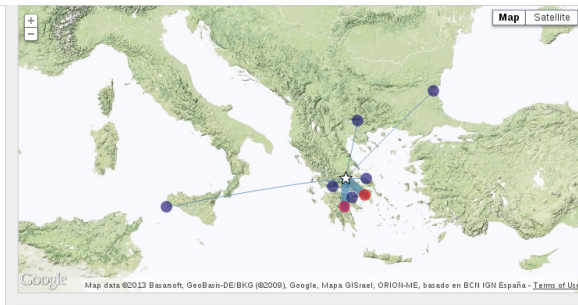
[Report a problem with this record](#)

External Resources

- [Place page on Pleiades](#)
- [Books referencing Thermopylae](#)
- [Pelagios Graph Explorer](#)

Top Related Places

[Athenae](#) (2)
[Hellas](#) (2)
[Corinthus](#) (1)
[Sicilia](#) (1)
[Strymon fl.](#) (1)
[Naupactus](#) (1)
[Thrace](#) (1)
[Dirphys M.](#) (1)



High frequency – Athens

Athenae



189 references

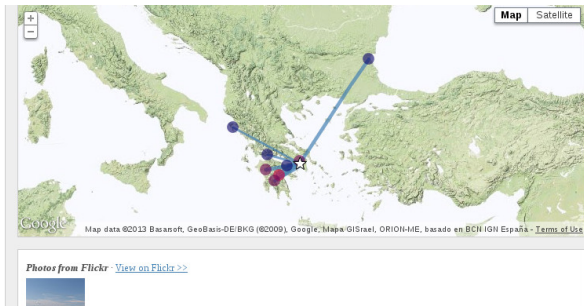
[Report a problem with this record](#)

External Resources

- [Place page on Pleiades](#)
- [Books referencing Athenae](#)
- [Pelagios Graph Explorer](#)

Top Related Places

[Hellas](#) (54)
[Peloponnesus](#) (25)
[Thrace](#) (22)
[Attica](#) (22)
[Lacedaemon/Laconia](#) (21)
[Corinthus](#) (13)
[Naupactus](#) (12)
[Corcyra](#) (12)



Research issues

- Testing interfaces like GapVis in course-work
- Users see interface rather than content
- Difficult to do normative evaluations
- Can we obtain a “gold standard”?
 - places that are independently known to be related in ancient texts like Thucydides?



Questions?

