The Alan Turing Institute

Text Meets Space

Part 4: Applications

Katie McDonough, Living with Machines

Contents

- A. Why are these methods interesting to humanities scholars?
- B. How can you use them to answer humanistic questions?
- C. What are researchers learning from using these methods?
- D. Asking questions of historical texts
- E. Open Challenges

A. Why are these methods interesting to humanities scholars?

The Scale of Humanities Research

Geographical Scale

In the last 20 years, humanities scholars–especially historians–have been expanding the scale of their analysis to inter/transnational and global topics.

Archive Scale

This often requires reading and interpreting larger numbers of archival or library documents. The typical humanities scholar works independently, and their ability to tackle larger sets of documents is a question of time (and language skills).

Patterns and Cases

Having machine-readable texts can slim down the time needed to identify *patterns* across texts as well as *unique cases* for deeper analysis.

Spatial Data in the Humanities

Tabular Data

In the early 2000s, researchers created datasets that could be used in GIS to answer spatial questions. For historians, this meant painstaking data entry, building tables by hand, or improving some pre-existing historical datasets like national census microdata (esp. to standardize and geocode toponyms). See, for example, projects at the Stanford Spatial History Project:

`https://web.stanford.edu/group/spatialhistory/cgi-bin/site/index.php`

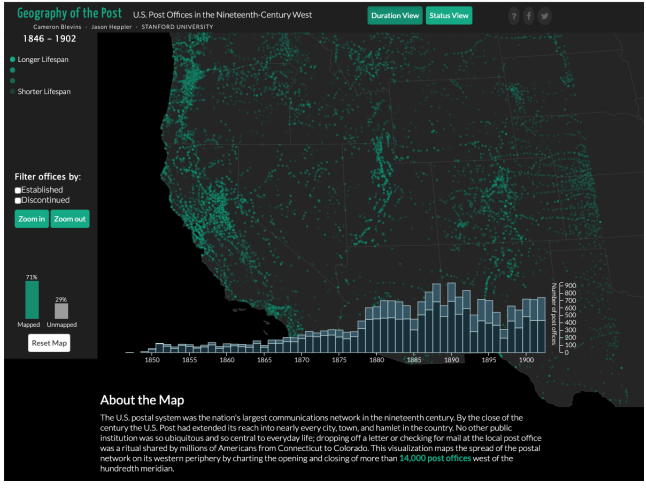


Figure: `http://cameronblevins.org/gotp/`

Thinking Spatially

Tables to texts

Humanities scholars cut their teeth on digital spatial methods with GIS, doing things like examining buffer zones, least-cost path analysis, or simply recording change over time (e.g. along shifting borders)

Post GIS

The ability to quickly create new datasets from digitized texts means we can capture and analyze data in new formats, without the restrictions of GIS. It also means that we can think spatially about new kinds of sources, moving beyond the most explicit information (e.g. names of cities).

Spatial Data in the Humanities

Text Data

Around 2015, humanities scholars were beginning to use Named Entity Recognition and corpus linguistics methods to extract spatial information from prose texts.

Making Data

- Unstructured: Internet Archive OCR'd txt files
- Structured: TEI-encoded xml files from Women Writers Project @ Northeastern (TEI encoding)

Scholars, libraries, and archives are "unlocking" texts everyday for computational analysis, hand keying or annotating texts or using tools like TRANSKRIBUS (to automate transcription of hand-written texts) and Recogito (to semi-automatically annotate place names in historical texts and images).

Working Collaboratively across Disciplines

Learning to work with text

As you will have seen in the earlier slides, information retrieval methods open up vast new territory for spatial analysis. While simple, out of the box tools can help humanities scholars perform simple tasks, the richest research comes from collaboration between experts in different fields.

Collaboration

Pushing the boundaries both of IR and humanities goals is an important criteria for successful collaboration: learning how to learn from another discipline can be one of the most rewarding steps in a project.

B. How can you use them to answer humanistic questions?

Current Projects

Digging into Early Colonial Mexico - `https://www.lancaster.ac.uk/digging-ecm/`

How did the Spanish colonial authorities portray and use information about the newly conquered territories and people? Can we identify, map, and analyse the geographies associated with the colonial period of Mexico, and what was said about them in historical sources, through expedited computational means?

Impresso - `https://impresso-project.ch/`

The project will produce a historical media monitoring tool suite that will bridge the semantic gap between huge volumes of scanned text and humanities scholars willing to understand and interpret its content. This research project aims to draw on new research tools to analyse newspapers and optimise their potential for examining the issue of anti-European trends.

Current Projects 2

GeoDisco - `https://www.msh-lse.fr/projets/geodisco/`

GeoDisco studies geographic discourses in French encyclopedias between 1751 and today. By using text analysis methods that draw on rule-based automatic identification of linguistic information and machine learning classification, we will study how French encyclopedias have expressed geographical information (both in terms of geographical coverage, and in terms of linguistic patterns) over time.

Trading Consequences - `http://tradingconsequences.blogs.edina.ac.uk/`

In the Trading Consequences project, historians, computational linguists, and computer scientists collaborated to develop a text mining system that extracts information from a vast amount of digitized published English-language sources from the "long nineteenth century" (1789 to 1914). The project focused on identifying relationships within the texts between commodities, geographical locations, and dates.

Learning from GIR

Step 1

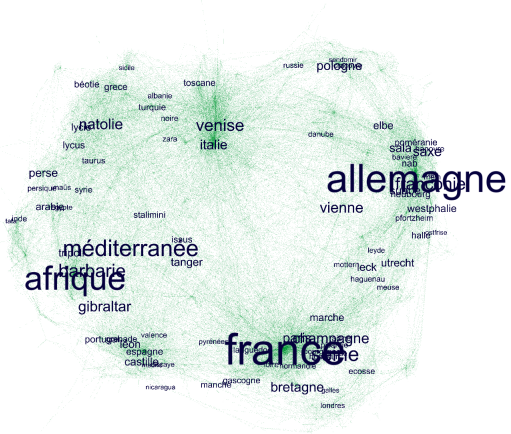
- We can move beyond just searching for and mapping place names.
- We can think about the relationships between places, and the development of a language of expressing "where" a thing is. –> Spatial language has a history.
- We can ask how dis/similar places are based on textual context.

Step 2

How can we use more abstract spatial information to inform humanities research? One blocker: teaching humanities scholars to think spatially.

C. What are researchers learning from using these methods?

Case Study: GeoDisco



Working with early modern French

Figure: Extended Named Entities, Gaio and Moncla 2017, `https://hal.archives-ouvertes.fr/hal-01492994`, McDonough et al 2019, `https://www.tandfonline.com/doi/full/10.1080/13658816.2019.1620235`

Case Study: Living with Machines

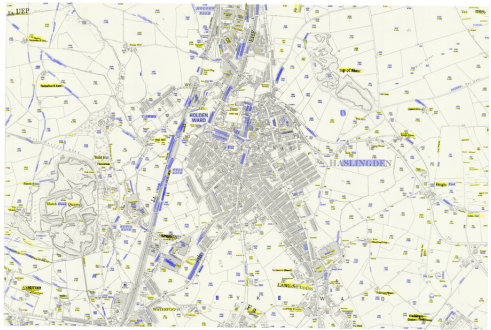


Figure: Text identified using Strabo,USC Spatial Informatics Laboratory. OS Map: Lancashire LXXI.12 (x; Rawtenstall). Survey date: 1890-2, publication date: 1911. Reproduced with the permission of the National Library of Scotland.

Starting with the gazetteer

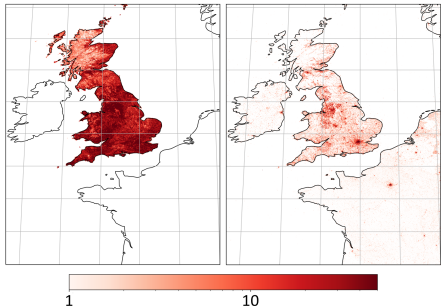


Figure: GB1900 (left) versus WikiGazetteer (right) in GB. The earth's surface is divided in blocks of 0.02 (≈2km), each representing the total number of geotagged articles. NB: colorbar is logarithmic. Coll Ardanuy et al 2019,
`https://bl.iro.bl.uk/work/ff87acd4-f5e0-4870-b8cd-63d82fcc36d8`

Gazetteers and Microtoponyms

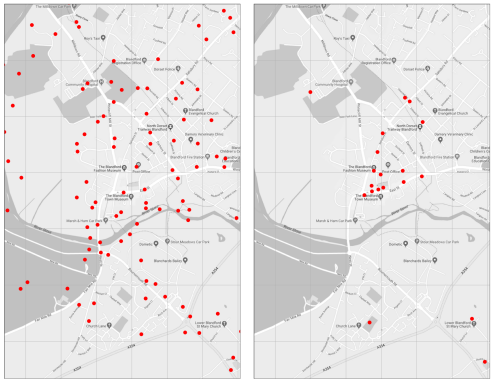


Figure: GB1900 (left) versus WikiGazetteer (right) for Blandford Forum, Dorset. Each red point corresponds to one geotagged entry.

D. Asking questions of historical texts

Sample Texts

Details here: `https://github.com/kmcdono2/ecir_textspace`

Samuel Johnson, *A journey to the western islands of Scotland* (1773)

A new voyage, round the world, in the years 1768, 1769, 1770, and 1771; undertaken by order of His present Majesty, performed by, Captain James Cook, in the ship Endeavour, drawn up from his own journal, and from the papers of Joseph Banks, Esq. F.R.S. (1774)

Samuel Johnson in Scotland

The roads of Scotland afford little diversion to the traveller, who seldom sees himself either encountered or overtaken, and who has nothing to contemplate but grounds that have no visible boundaries, or are separated by walls of loose stone.

Asking questions

In what ways did the public write about roads that had initially been built for the military? –> Where does Johnson discuss transportation infrastructure, and in relation to which places?

Given the brutality of the Highland Clearances, how did English travelers represent the Scottish landscape? –> Where are there beautiful places? Where are there ugly places?

James Cook and Joseph Bank on the Endeavour (1768-71)

On the 26th of January, we took our departure from Cape Horn, which lies in latitude 55 53:S. longitude 68 13:W. The farthest southern latitude that we made was 60 10:, our longitude was then 74 30:W.; and we found the variation of the compass, by the mean of eighteen azimuths, to be 27 9:E. As the weather was frequently calm, Mr. Banks went out in a small boat to shoot birds, among which were some albatrosses and sheerwaters. The albatrosses were observed to be larger than those which had been taken northward of the Streight; one of them measured ten feet two inches from the tip of one wing to that of the other, when they were extended: the sheerwater, on the contrary, is less, and darker coloured on the back.

Cook Example Questions

How does Cook use latitude and longitude in his journal? What is the relationship between quantitative and qualitative expressions of place?

Where does "science" happen during the Endeavour's journey? Is it important to document the location of discoveries? At what scale? Does this change over time?

How geographical are descriptions of individuals and groups of people?

Finally, do modern entity types apply to the 18th century?

E. Open Challenges

Work in Progress

1. What is a place? –> research on "microtoponyms" and other entities, addressing issues of geographic scale in IR, nested entities
2. What are the best methods for selecting toponym candidates for historical texts?
3. How do (*should*) you locate places that have no gazetteer record? –> World Historical Gazetteer Project
4. What is the best way to represent the results of GIR when location data is uncertain? On a map? In a graph?
5. How do we improve GIR for non-English, or pre-modern texts?
6. How do we handle anachronistic entity types?

Resources

See the resources page on our Github repo for this tutorial. `https://github.com/kmcdono2/ecir_textspace/blob/master/resources.md`

Thank you