Geoparsing Historical Documents

Claire Grover

THE UNIVERSITY of EDINBURGH informatics
Overview

• Geoparsing
  – Geotagging: automatic identification of place names in text (NER)
  – Geo-resolution: automatic grounding of place names to lat/long (ambiguity resolution)
  – The Edinburgh Geoparser
  – Gazetteers

• Two projects
  – Trading Consequences (Digging into Data II: historical documents)
  – Palimpsest (AHRC Big Data: literature)
The Edinburgh Geoparser

• Result of many years of collaboration with EDINA
• Use NER to identify place names in text
• Find matching records in a gazetteer
• Resolve ambiguity using contextual information from the document
• The Edinburgh Geoparser is embedded in EDINA’s Unlock Text service
• Also available as stand-alone toolkit
Leicester prepares for Richard III reburial

Deddington usually divides its sympathies between the two sides, but for this they are coming out in favour of the deposed king in the Battle of Bosworth Field.

Leicester is filling up with yellow road signs, warning of the return of the once and future king: “King Richard III Route Sunday 22 – expect delays”.

Many have been wondering how to get hold of them to keep as souvenirs of an extraordinary episode in the history of the city and county, as the final preparations get underway for the 21st century reburial of a medieval king.

It is more than two years since the bones rediscovered by archaeologists under a council car park in the heart of Leicester, were formally identified as those of the last Plantagenet king.

In 1485 the burial by Franciscan friars was unceremonious. In the heat of August, several days after his death at Bosworth, he was buried hastily in a roughly-dug grave slightly too small for his naked body. This time a week of events will surround the reinterment on Thursday 26 in the cathedral – delayed for a year by a legal challenge claiming that he should have been buried in York – including a procession on Sunday through the countryside associated with the last days of the last Plantagenet, led by two knights in armour on horseback.

There will be prayers in the cathedral from two archbishops, Justin Welby and Vincent Nichols; open-air religious services; live television broadcasts to an audience of millions; lectures, art exhibitions and book launches; flower festivals; concerts of medieval music; fireworks and the first ringing of a new peal of bells composed for Richard.

In the cathedral the £2.5m remodelling outside and in is almost complete. Beautiful red sandstone slabs, inset with four small white stone roses, have been laid in the new space created behind the altar, where purple drapes now cover a brick-lined pit ready to receive the coffin. The work disturbed seven previously unrecorded burial vaults, and their dozens of occupants have been reverently gathered up and reburied. The tomb of the Vauxhns, a 19th-century vicar and his family, which began to peel off the wall from the vibrations from the work, has been boxed in for safety and will be restored later.

“We’re almost there,” the Rev Pete Hobson said. “The new altar is due to arrive with just two or three days to spare – when it comes I’ll breathe more easily.”

People are coming from all over not just for the cathedral, where every seat could have been filled a hundred times, but to myriad events in the small villages and towns usually bypassed by tourists: they include a Last Plantagenet Dinner at the Red Lion in Market Bosworth, where school children have made half a mile of bunting and the medieval costumes have arrived for all the staff of Chocolat’s sweet shop selling shields ornamented with white chocolate roses.

In the village of Sutton Cheney representatives of the local council, the highways authority, the cathedral, villagers, the vicar and churchwardens, gathered to measure with a stop watch every second of the ten minutes the cortège will stop on Sunday afternoon on its slow way to the cathedral. It won’t allow enough time to come into the church of St James the Greater, where Richard is believed to have taken his last communion before the battle, so Julia Hargreaves will hold a prayer service at the gate. Clerk Shelley Howard arrived a little late, having taken a phone call from somebody in Ohio asking if they could come to the previous night’s candlelit service. “They wanted to know how much the tickets were,” she said, shaking her head in amazement.

A country lane leads from the village street to Ambion Hill, where Richard’s troops are believed to have camped on the eve of battle, now the site of the Bosworth Battlefield Heritage Centre where a major open air ceremony will be led by the Tim Stevens, the bishop of Leicester. When tickets went online 10 days ago, the website crashed within minutes: every member of staff came...
Dadlington usually divides its sympathies between the two sides, but for this they are coming out in favour of the deposed king in the Battle of Bosworth Field.

Leicester is filling up with yellow road signs, warning of the return of the once and future king: “King Richard III Route Sunday 22 – expect delays”.

Many have been wondering how to get hold of them to keep as souvenirs of an extraordinary episode in the history of the city and county, as the final preparations get underway for the 21st century reburial of a medieval king.

It is more than two years since the bones rediscovered by archaeologists under a council car park in the heart of Leicester, were formally identified as those of the last Plantagenet king.

In 1485 the burial by Franciscan friars was unceremonious. In the heat of August, several days after his death at Bosworth, he was buried hastily in a roughly-dug grave slightly too small for his naked body. This time a week of events will surround the reinterment on Thursday 26 in the cathedral – delayed for a year by a legal challenge claiming that he should have been buried in York – including a procession on Sunday through the countryside associated with the last days of the last Plantagenet, led by two knights in armour on horseback.

There will be prayers in the cathedral from two archbishops, Justin Welby and Vincent Nichols; open-air religious services; live television broadcasts to an audience of millions; lectures, art exhibitions and book
Unlock the potential in your data with our simple web services

Unlock Places

Use Unlock Places to search for places. Search across different sources of geographic data - gazetteers - and get results in different useful forms for web and mobile apps.

The Unlock Places place search API is an open service, with extensive coverage in the UK through Ordnance Survey Open Data, and worldwide coverage added by Geonames and Natural Earth Data.

- Get started
- API documentation

Unlock Text

Use the Unlock Text place-name text-mining service to extract place-names from documents and find their locations. Our geoparser extracts place-names from text, then Unlock Text uses the Places search API to find likely locations, and rank them.

For use geo-referencing 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 adipiscing elit, vel urna mattis edinburgh
• Knowledge about possible interpretations comes from a gazetteer
  – A gazetteer pairs place names with lat/longs
  – Paris, France: 48.85339, 2.34864
  – Paris, Texas: 33.66094, -95.55551
  – Edinburgh Geoparser is configured to use gazetteers provided through EDINA’s Unlock Places service but it can be adapted to use other gazetteers.
• If a place name is not in the gazetteer, it cannot be grounded
• If the correct interpretation of a place name is not in the gazetteer, it cannot be grounded correctly
• Modern gazetteers may not be ideal for historical documents so historical gazetteers can/should be developed:
  – The Historical Gazetteer of England’s Place-Names developed in the Digitisation of English Place Names (DEEP) project.
  – Pleiades++ developed in the Google Ancient Places (GAP) project.
The Historical Gazetteer of England's Place-Names is a key piece of digital infrastructure for use in the humanities, arts, and social sciences. Its aim is to associate disparate content through place - everything happens somewhere, after all - and therefore to facilitate accurate searches across resources. Gazetteers of contemporary place-names have been available for some time, but the Historical Gazetteer's historical place-name forms add chronological depth to the mix. These forms have been collected over the last ninety years as part of the English Place-Name Society's Survey of English Place-Names. The Historical Gazetteer brings the four million + historical place-name forms of the Survey, including those for hamlets, fields, and streets, into the digital realm, heralding a new era of chronological depth and spatial granularity in gazetteer provision.

Details of how to use this site can be found through the links below.
which is of Alured's fee in Meleburn by the service of half a knight, and in Widiate by the service of one knight for all service. For this, the tenant granted and quit

2 John (1200-01).

claimed for himself and his heirs all his right in the land of Tarente to Alured and his heirs for ever, which William de Villers held. And be it known that if Walter should die before William his son, William shall by this fine be quit respecting his relief, against Alured and his heirs.

(8, new number 7) At Westminster, in the octave of St. Blasius, between Cristiana daughter of William Clunt, plaintiff, and Ermengard, Prioress of Clerkenwell, tenant, of half a virgate of land in Blaneford. The plaintiff remitted and quit claimed all her right in the same for herself and her heirs to the Prioress and her successors for ever. For this the Prioress gave to the plaintiff four marks of silver.
Google Ancient Places (GAP)

- The GAP project, funded under the Google Digital Humanities programme, aimed to identify toponym references in works such as Herodotus' *Histories*, Livy's *History of Rome* and Tacitus' *Annals*, and create a map-based visualisation tool to be used by students and researchers of the ancient world.

- Gazetteer of the ancient rather than the modern world, Pleiades.
  - Expanded to create Pleiades+ by matching, where possible, the ancient places to their modern equivalents in GeoNames.
  - Pleiades++ step to allow GeoNames alternative names lists to mediate choice of candidate from Pleiades+ (Egypt, Aegyptus).
How GapVis Works

classical texts from Google Books

Onelogging

Edinburgh Geoparser

database of toponym URIs – tied back to text snippets

Pleades gazetteer of ancient places

putting it all together – GapVis

finding relationships between places
Trading Consequences

• What does archival text say about the economic and environmental consequences of global commodity trading during the nineteenth century?

• Goals:
  – Text mining, data extraction and information visualisation to explore big historical datasets.
  – Focus on how commodities were traded across the globe in the 19th century.
  – Help historians to discover novel patterns and explore new research questions.

• Example questions:
  – What were the routes and volumes of international trade in resource commodities 1850-1914?
  – What were the local environmental consequences of this demand for these resources?
<table>
<thead>
<tr>
<th>Collection</th>
<th># of Documents</th>
<th># of Images</th>
</tr>
</thead>
<tbody>
<tr>
<td>House of Commons Parliamentary Papers (ProQuest)</td>
<td>118,526</td>
<td>6,448,739</td>
</tr>
<tr>
<td>Early Canadiana Online</td>
<td>83,016</td>
<td>3,938,758</td>
</tr>
<tr>
<td>Directors’ Letters of Correspondence (Kew)</td>
<td>14,340</td>
<td>n/a</td>
</tr>
<tr>
<td>Confidential Prints (Adam Matthews)</td>
<td>1,315</td>
<td>140,010</td>
</tr>
<tr>
<td>Foreign and Commonwealth Office Collection</td>
<td>1,000</td>
<td>41,611</td>
</tr>
<tr>
<td>Asia and the West (Gale)</td>
<td>4,725</td>
<td>948,773 (OCRed: 450,841)</td>
</tr>
</tbody>
</table>
Trading Consequences

From Padang was exported, in 1871, 6,127 piculs of cassia bark, of which a large portion was shipped to America (Fliickiger and Hanbury). ...

- Normalised and grounded entities:
  - commodity: cassia bark [concept: Cinnamomum cassia]
  - date: 1871 (year=1871)
  - location: Padang (lat=-0.94924;long=100.35427;country=ID)
  - location: America (lat=39.76;long=-98.50;country=n/a)
  - quantity + unit: 6,127 piculs
Trading Consequences

From Padang was exported, in 1871, 6,127 piculs of cassia bark, of which a large portion was shipped to America (Fliickiger and Hanbury). ...

• Extracted entity attributes and relations:
  • origin location: Padang
  • destination location: America
  • commodity–date relation: cassia bark – 1871
  • commodity–location relation: cassia bark – Padang
  • commodity–location relation: cassia bark – America
Oxford 22/03/2105

Reassembling the Republic of Letters
OCR Issues

9, Montreal
2, Montréal
2, Montrent
2, Montrea
1, MO.’N’YREUL
1, Mont-treal
1, MONTRLAL
1, Monreal
1, MONTRBl’tL
1, MONTiEAL
1, MIontret]
1, Mbonreal
1, Maontreal
1, 3MON2RRRA
1, 10TRBAL
1, 1ONTREAL
Palimpsest: an Edinburgh Literary Cityscape
Palimpsest

Literature, University of Edinburgh
James Loxley, Professor of Early Modern Literature
Miranda Anderson, Research Fellow
Tara Thomson, Research Fellow

Informatics, University of Edinburgh
Jon Oberlander, Professor of Epistemics
Beatrice Alex, Research Fellow in Text Mining
Claire Grover, Senior Research Fellow

SACHI: St Andrews Human Computer Interaction Research
Aaron Quigley, Director of SACHI & Chair of Human Computer Interaction
David Harris-Birtill, Research Fellow
Uta Hinrichs, Research Fellow

EDINA
James Reid, Workgroup Leader, Geoservices
Nicola Osborne, Social Media Officer
Approach

• Retrieve literary works which are at least partly set in Edinburgh from all literature accessible to us.
• Devise a method for identifying “loco-specificity” in literature automatically based on input from literary scholars.
• Create a fine-grained location gazetteer for Edinburgh.
• Identify and geo-reference locations (including street names and buildings) using the Edinburgh Geoparser.
• Visualise geolocated snippets on web or mobile device.
• Rank snippets for ‘interestingness’.
Datasets

• HathiTrust collection (all worldwide public domain material)
• British Library Nineteenth Century Books collection
• English Project Gutenberg books
• Oxford Text Archive data
• National Library of Scotland data
• Limited set of copyrighted material, with author/publisher agreement (Irvine Welsh, Muriel Spark, Alexander McCall Smith, Doug Johnstone, Ron Butlin, ...)

Oxford 22/03/2105
Reassembling the Republic of Letters
Assisted curation of Edinburgh-centric books

- **Hathi:**
  - 239,481 books in initial set.
  - Automatic ranking for relevance reduced set to 6,025.
  - Manual curation selected 337 items.
- **British Library Nineteenth Century Books**
  - 546 after automatic ranking
  - Manual curation selected 111 items
- **Project Gutenberg, Oxford Text Archive and National Library of Scotland** provided a further 55 items.
- Plus 42 modern books
- Publication dates from 1742 onwards but the vast majority from 19th century.
- In total 47,376 mentions of 1,360 distinct Edinburgh place names
Edinburgh Gazetteer

• Created a local Edinburgh gazetteer by aggregating records from Open Street Maps, OS Locator and the Royal Commission on the Ancient and Historical Monuments of Scotland (RCAHMS).

• 13,085 entries, e.g.:
  <place name="Grassmarket" lat="55.946734" long="-3.19771" source="rcahms"/>
  <place name="Grassmarket" lat="55.9474" long="-3.1960" source="osl"/>

• Fairly good coverage but some items missing and a lot of noise from Open Street Maps and RCAHMS.
Thank you!

- Trading Consequences: http://tradingconsequences.blogs.edina.ac.uk/
- Palimpsest: http://palimpsest.blogs.edina.ac.uk/
- The Historical Gazetteer of England's Place-Names: http://www.placenames.org.uk/
- Google Ancient Places https://googleancientplaces.wordpress.com/