Text Mining and Geo-referencing Historical Text

Beatrice Alex
Edinburgh Language Technology Group
School of Informatics
balex@inf.ed.ac.uk
@bea_alex

DCHRN Workshop Cultural Heritage Sparks, Edinburgh, Jan 29th 2016

- Language Technology Group: www.ltg.ed.ac.uk
- Research and development of natural language processing techniques and technology.
- Collaboration in projects with partners in a range of different disciplines (biodiversity, biomedicine, education, cultural heritage, history and literature).
- Aggregation, text mining, geo-parsing, natural language generation, linking of data.

- Recent projects:
  - **Palimpsest** (Mining Literary Edinburgh, AHRC)
  - **UK Connect** (Analysis of social media, British Council)
  - **BotaniTours** (Information aggregation and presentation of botanical points of interest in the Scottish Borders, dot.rural).
  - **Trading Consequences** (Text mining trends in commodity trading of large 19th century text collections, Digging into Data).
  - **New:**
    - **HistText:** geo-parsing the Historical Texts data (Jisc)
    - Text mining brain scan reports for clinical neurologists (MRC).

- Describes a set of linguistic, statistical and/or machine learning techniques that model and structure the information content of textual resources.
- Turns unstructured text into structured data (e.g. relational database or linked data).
- Is very useful for analysing large text collections automatically (overcoming data paralysis).
- Goal: Analyse large (or small) textual collections to enable scholars to discover novel patterns and explore hypotheses.

- Jul 2015-Feb 2016 (Jisc)
- Jisc created the Historical Texts portal to EEBO, EECO, and the British Library Nineteenth Century Books collection
- University of Edinburgh is currently not licensing access to this portal. :-(

Your institution does not have access to Historical Texts

Historical Texts is available via subscription to UK Higher Education and Further Education institutions who are full members of Jisc Collections. Your institution may not subscribe to the service or, if it does, it may have changed its UK Federation (Shibboleth/OpenAthens) identity provider details.

Please contact our helpdesk at historicaltexts@mimas.ac.uk to confirm whether your institution subscribes and for further assistance.

Please include the following error message in any email:

Identity provider lookup failed at (https://historicaltexts.jisc.ac.uk/Shibboleth.sso/Login)

EntityID: https://idp.ed.ac.uk/shibboleth

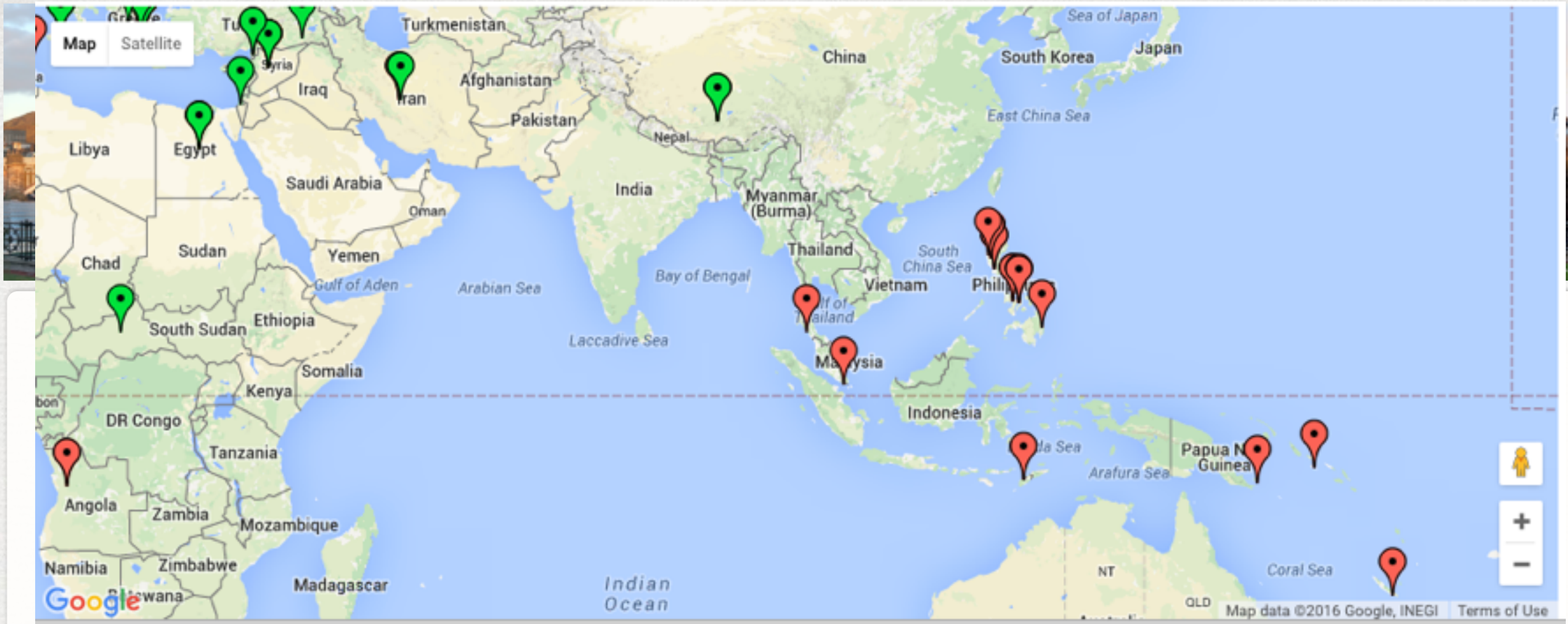
opensaml::saml2md::MetadataException: Unable to locate metadata for identity provider (https://idp.ed.ac.uk/shibboleth)

HISTORICAL TEXTS

- Describes a set of linguistic, statistical and/or machine learning techniques that model and structure the information content of textual resources.
- EEBO-TCP (1473-1700)
  - 29,548 books
  - 113,869 MARC records
- ECCO-TCP (1701-1800)
  - 2,398 books
  - 182,157 MARC records
- BL Nineteenth Century (1789-1914)
  - Over 65,000 books
  - ? MARC records

HISTORICAL TEXTS

- Our job is to geo-parse all of this data to create more location meta-data and thereby improve search and discovery.
- Challenges:
  - Historical place names: Some place names were reused by explorers and discoverers of the USA, Australia and New Zealand. We employ a bounding box to excludes locations which have not been discovered at a certain point in time.
  - Lack of availability of historical gazetteers: had to select sub-set of locations with GeoNames for example, we also applied the Pleiaded-Plus gazetteer of ancient places.
  - Language variation and case (mostly EEBO): Grasse (grass) versus Grasse (France), Hamme (ham) vs. Hamme (Belgium)… we use a list of common words to help distinguish between them.



A CONFERENCE ABOVT THE NEXT SVCCESSION TO THE CROWNE OF INGLAND, DIVIDED IN TO TVVO PARTES.

VVHERE-OF THE FIRST CONTEYNETH THE discourse of a ciuill Lavvyer, hovv and in vvhat manner propinquity of blood is to be preferred. And the second the speech of a Temporall Lavvyer, about the particuler titles of all such as do or may pretende vvithin Ingland or vvithout, to the next succession.

VVhere vnto is also added a new & perfect arbor or genealogie of the discents of all the kinges and princes of Ingland, from the conquest vnto this day, whereby each mans pretence is made more plaine.

DIRECTED TO THE RIGHT HOnorable the earle of ESSEX of her Maiesties priuy councell, & of the noble order of the Garter.

Published by R. DOLEMAN.

Imprinted at N. with Licence.

M. D. XCIIII.

Click on a lat/long to centre the map there.

Notingham	52°57'N 1°9'W			
Holland	52°15'N 5°45'E	50°55'N 3°57'E	52°20'N 4°50'E	1
Israel	31°30'N 34°45'E	9°38'S 124°14'E	10°20'S 15°1'E	8
Yorke	4°45'N 7°34'W			
Braganza	41°50'N 6°46'W	31°7'N 82°15'W		
Africa Africa				
Burgundy	47°15'N 4°10'E	47°0'N 4°30'E		
Castile	41°0'N 3°30'W	42°38'N 78°3'W	45°30'N 77°19'W	3
Aragon	41°30'N 0°40'W	43°18'N 2°19'E	41°0'N 1°0'W	4
England	52°10'N 0°42'W	53°0'N 2°0'W	54°29'N 8°53'E	6
Austria	47°20'N 13°20'E	15°47'N 120°18'E	17°37'N 92°2'W	1
Normandy	51°15'N 0°40'W	51°16'N 0°40'W	49°0'N 0°0'E	4
Gascony	43°30'N 0°0'E			
Guyenne	44°35'N 1°0'E	48°47'N 78°28'W	19°2'N 72°37'W	4
Babilon	49°56'N 11°53'E	51°9'N 21°38'E	13°21'N 16°37'W	

HISTORICAL TEXTS

- Our job is to geo-parse all of this data to create more location meta-data and thereby improve search and discovery.
- Challenges:
  - Historical place names: Some place names were reused by explorers and discoverers of the USA, Australia and New Zealand. We employ a bounding box to excludes locations which have not been discovered at a certain point in time.
  - Lack of availability of historical gazetteers: had to select sub-set of locations with GeoNames for example, we also applied the Pleiaded-Plus gazetteer of ancient places.
  - Language variation and case (mostly EEBO): Grasse (grass) versus Grasse (France), Hamme (ham) vs. Hamme (Belgium)… we use a list of common words to help distinguish between them.

- Tools (https://www.ltg.ed.ac.uk/software/):
  - **The Edinburgh Geoparser:** an open-source tool for geo-referencing text. See also our online demo at: http://jekyll.inf.ed.ac.uk/geoparser.html
  - **LT-XML2 and LT-TTT2:** XML-based software for shallow linguistic processing of text.

NATURAL LANGUAGE GENERATION

THE UNIVERSITY of EDINBURGH

The University of Edinburgh Homepage Link

MUSICAL INSTRUMENTS MUSEUMS EDINBURGH

- Amy Isard, PhD candidate: Natural Language Generation for cultural heritage data
- Structured data -> natural language
- Contact: amyi@inf.ed.ac.uk

- Questions?
- Contact: balex@inf.ed.ac.uk
- Website: http://homepages.inf.ed.ac.uk/balex/
- Twitter: @bea_alex