Adapting the Edinburgh Geoparser for Historical Georeferencing

ADAPTING THE EDINBURGH GEOPARSER FOR HISTORICAL GEOREFERENCING

BEATRICE ALEX, KATE BYRNE, CLAIRE GROVER AND RICHARD TOBIN

Abstract *Place name mentions in text may have more than one potential referent (e.g. Peru, the country vs. Peru, the city in Indiana). The Edinburgh Language Technology Group (LTG) has developed the Edinburgh Geoparser, a system that can automatically recognise place name mentions in text and disambiguate them with respect to a gazetteer. The recognition step is required to identify location mentions in a given piece of text. The subsequent disambiguation step, generally referred to as georesolution, grounds location mentions to their corresponding gazetteer entries with latitude and longitude values, for example, to visualise them on a map. Geoparsing is not only useful for mapping purposes but also for making document collections more accessible as it can provide additional metadata about the geographical content of documents. Combined with other information mined from text such as person names and date expressions, complex relations between such pieces of information can be identified. The Edinburgh Geoparser can be used with several gazetteers including Unlock and GeoNames to process a variety of input texts. The original version of the Geoparser was a demonstrator configured for modern text. Since then, it has been adapted to georeference historic and ancient text collections as well as modern-day newspaper text.*[1–4] *Currently, the LTG is involved in three research projects applying the Geoparser to historical text collections of very different types and for a variety of end-user applications. This paper discusses the ways in which we have customised the Geoparser for specific datasets and applications relevant to each project.*

Keywords: Georeferencing, georesolution, text mining, domain adaptation

International Journal of Humanities and Arts Computing ... , Edinburgh University Press
DOI: ...

http://www.euppublishing.com/ijhac

INTRODUCTION: THE EDINBURGH GEOPARSER

The Edinburgh Geoparser is a Natural Language Processing (NLP) system designed to analyse text in order to identify occurrences of locations and 'pin' them to a map by determining their correct latitude and longitude. This involves disambiguation wherever a location has more than one possible interpretation. An detailed introduction to geoparsing and the steps involved can be found in the paper by Kalev H. Leetaru (2012).[5] The Geoparser's functionality is comparable to other software such as CLAVIN[6], OpenCalais[7] or Yahoo PlaceSpotter.[8] The Geoparser described here is an update of the system which we reported on previously.[2,9] In those papers, we evaluated its performance against the SpatialML corpus[10] as well as against historical English documents. The software can be downloaded from the LTG website[11] and a version of the Geoparser is also the backend of EDINA's Unlock Text,[12] a RESTful API for geoparsing texts on the Web.

There are two main components in the Geoparser, a named entity recognition (NER) or geotagging component and a georesolution component. The former uses NLP techniques to identify named entities in text, specifically location, person and date entities. The latter looks up the location names in a gazetteer and resolves ambiguities to suggest the most likely interpretation (i.e. latitude/longitude, country and type) for each location given its context in the text being processed. The recognition component is a pipeline of sub-components built using XML tools in combination with Unix shell scripting. The XML tools are LT-XML2[13] and LT-TTT2[14] which we have been specifically designed for NLP applications. The pipeline converts input text to XML, performs low-level analysis such as tokenisation and sentence-splitting and then applies part-of-speech (POS) tagging and lemmatisation, syntactic chunking and NER. For POS tagging and lemmatisation we use third party software.[15,16] The chunking and NER steps are rule-based as opposed to components using machine learning to make predictions. The output of the recognition component is a linguistically annotated version of the input text with the location, person and date entities marked up in XML format. The georesolution component takes this as input and looks up the location names in a gazetteer. The standard version of the Geoparser allows for the use of a two gazetteers, namely GeoNames[17] and Ordnance Survey data (both available in Unlock[18]). In the projects we report on here, we have extended the Geoparser to allow the use of historical

gazetteers or adjusted its feature set used for the georesolution. Queries to a gazetteer return all the records which match the input location name and the job of georesolution is to rank these records in order of likelihood in the given context. The georesolver uses heuristics combined with weighting of information to arrive at its rankings. For example, populated places are preferred over places described in a gazetteer as facilities, and larger places (by population) are weighted more highly than smaller ones. The possible interpretations of other locations in the document are used so that all locations mutually constrain one another to be as close together as possible. Thus a document mentioning *Portsmouth*, *Southampton* and *Bournemouth* will be analysed so that the place-names resolve to towns on the south coast of England, while a document containing *Portsmouth*, *Hampton* and *Chesapeake* will resolve the names to places in Virginia, USA.

The Edinburgh Geoparser has been in development for a number of years and has been used in several projects with good effect. The rule sets for recognition and resolution have been tuned and tested in many contexts and are reasonably stable. However, potential users of the Geoparser are very wide-ranging and the texts that they wish to analyse are of all kinds, in many formats and put to many different purposes. This makes it extremely difficult to create a robust Geoparser which will please all of the users at all times. The easiest access to the Geoparser is via the Unlock Text API which accepts a range of parameters but cannot be fully customised for a particular purpose, and we expect that many users will want to download the Geoparser source and adjust it for their own needs, much as we have done in the projects described in this paper.

The named entity recogniser in the Geoparser has advantages and disadvantages: its behaviour is more transparent than supervised machine learning NER systems and rule sets can be altered relatively straight-forwardly; however, because it relies on lexicons and hand-written rules, it will stumble in cases not foreseen by the authors. The pipeline architecture of the system allows for the user to completely replace the NER component with a component of their own choosing, or to input documents where the named entities have been annotated by hand. Similarly, users may have documents with very specific formats such as tabular data where they may wish to restrict which parts of the document get processed.

The georesolution component has been tuned primarily to modern newswire text. Texts from

newswire are typically quite short and deal with a fairly specific topic with a fairly specific geographic focus. When processing longer documents, the user needs to be aware of the way in which the location names influence each other's interpretations and they should consider segmenting the document into smaller, geographically coherent pieces. Furthermore, the weights that we have chosen for the features that contribute to the resolution have been optimised for newswire and in different settings users may want to adjust these weights. Often, a user will know what the geographical focus of their document is, whether this focus is a continent, a country or a smaller area. We provide command line options to allow this area to be specified either as a bounding circle or a bounding box so that interpretations inside the area can be more highly ranked according to a user specified weight. Places outside the bounding circle or box may still be selected, so users wishing for an absolute constraint would need to filter the results to exclude the outliers.

ADAPTING THE GEOPARSER

In this section, we report on adjustments made to the Edinburgh Geoparser for Trading Consequences, GAP and DEEP, three research projects all processing historical text of different kinds.

(1) The Trading Consequences Project: Georeferencing Nineteenth Century Text

In Trading Consequences,[19] a Digging into Data II project (CIINN01), the aim was to assist historians in understanding economic and environmental consequences of commodity trading in the nineteenth century British Empire. We applied text mining to large quantities of digitised historical text, which when combined with visualisations presented in a web interface enables historians to analyse trends in commodity trading for a broad range of commodities (see Figure 1).[20]

We analysed textual data from major British and Canadian datasets, including the House of Commons Parliamentary Papers available through ProQuest,[22] the Early Canadiana Online data archive[23] and a sub-part of the Foreign and Commonwealth Office Collection from JSTOR.[24] We are also analysing Adam Matthew's Confidential Print collections,[25] the Directors' Correspondence Collection from the Archives at Kew Gardens available at JSTOR Global Plants[26] and several hundred manually selected titles relevant to this domain. With the exception of the Kew data, all datasets were digitised via Optical Character Recognition (OCR) and text quality varies

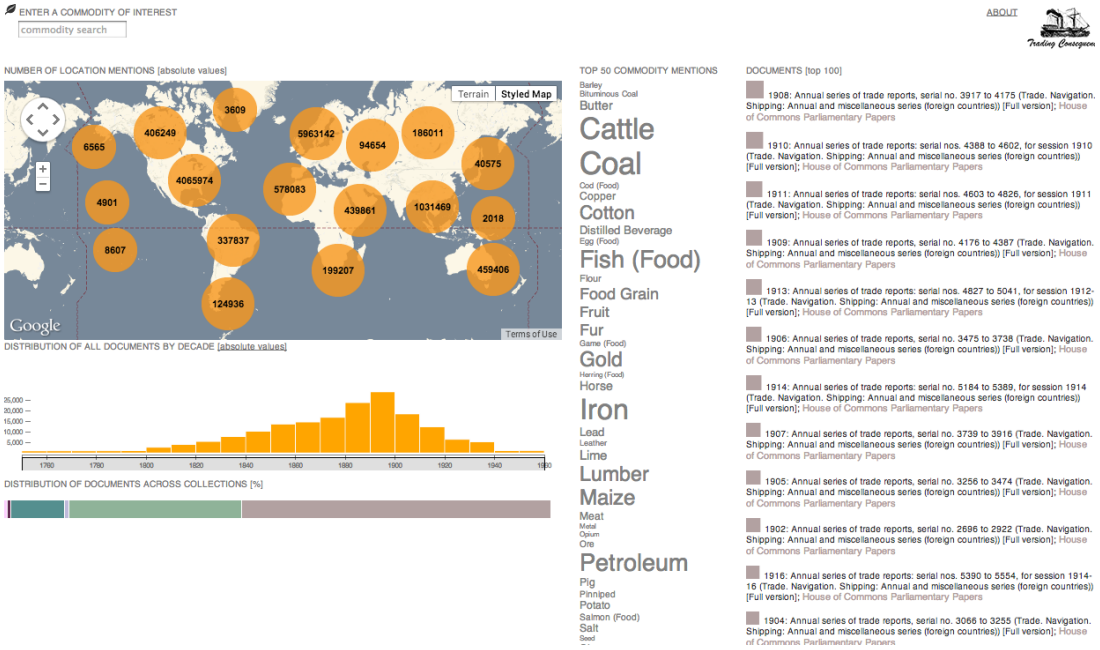


Figure 1. Web interface to the interlinked visualisation of Trading Consequences.[21]

considerably for and within each collection. Together these sources amount to over 10 million pages of text and over 7 billion word tokens. We used the Edinburgh Geoparser combined with the GeoNames gazetteer as part of the text mining component to identify and ground locations in these collections. From previous experience, we knew early on that we had to make some adjustments to the Geoparser to process historical collections relevant to Trading Consequences. At the recognition step, for example, we found that identifying person names and location names in parallel, even though there is no need to extract person name information for the intended application, helped to improve the overall quality of the text mined output. There are many location names which are made up of person names or which are similar to them. For example, there is the location *Markham* in Ontario and the person name *Clements Markham*, the British official who was responsible for collecting cinchona plants from their native Peruvian forests and transplanting them to India. In the initial Trading Consequences prototype the person name *Markham* was wrongly identified as a location mention and grounded accordingly. Consequently, this error appeared in the map visualisation for the commodity *cinchona*, the plant whose bark was processed into quinine. We therefore switched on the named entity recognition step for person names to

avoid such entity type confusion. This approach is not specific to this project but works equally well for other datasets and applications. In the experiments presented next we evaluate the georesolution step of the Edinburgh Geoparser for adjustments we made to its feature set specifically for Trading Consequences.

A. Gold Standard Data

In order to evaluate the effect of the changes we made to the Edinburgh Geoparser, we created a gold standard dataset containing manually annotated location mentions georeferenced to GeoNames. The gold standard is made up of document extracts from 25 randomly selected documents for each of the five collections processed in Trading Consequences and for the manually selected documents. Extracts were created to reduce the load of the annotator by splitting the document into equal sized chunks of 5 KB and randomly selecting one extract per document. The gold standard therefore contains a total of 150 document extracts. The annotation was performed in two steps. We firstly asked an annotator to mark up the entire gold standard with location mentions even if they contained errors introduced through the digitisation process. In total, the gold standard contains 4,373 manually identified location mentions. We then ran the Edinburgh Geoparser over the manually annotated data without applying a cut-off to the number of locations returned by GeoNames and without ranking the results. The annotator then carried out the georesolution annotation using the Edinburgh Geo-Annotator[27] by selecting one of the suggested candidates. He was able to do that for 3,109 locations. He selected none of the suggested candidates for 283 locations; and for 981 locations GeoNames did not return any candidate, so no candidate resolution could be made. One of the reason for the high number of locations without any GeoNames candidates is that 14.8% of location mentions in the gold standard contain OCR errors. For example, all mentions referring to the location name *Montreal* containing at least one error are listed in Figure 2 along with the number of times they occur in the gold standard. OCR errors affect named entities worse than common vocabulary, as this percentage decreases, for example, to 9.1% for commodity mentions in text.[28] This is most likely because OCR engines used to digitise documents rely on a dictionary or language model which does not contain many proper nouns. More detailed information on the effect that OCR errors have on named entity recognition for the historical texts

processed in Trading Consequences can be found in Alex and Burns (2014).[29]

```
9,Montreai
2,Montroal
2,Montrent
2,Montrea
1,MO.'N'YREUL
1,Mont- treal
1,MONTRLAL
1,Montreali
1,MONTREAL
1,Mont real
1,MONTRBf'tL
1,MONTIiEAL
1,MIontret]
1,Mbontreal
1,Maontreal
1,3MON2RRA
1,10TRBAL
1,10NTREAL
```

Figure 2. Forms of *Montreal* containing OCR errors and their counts in the gold standard.

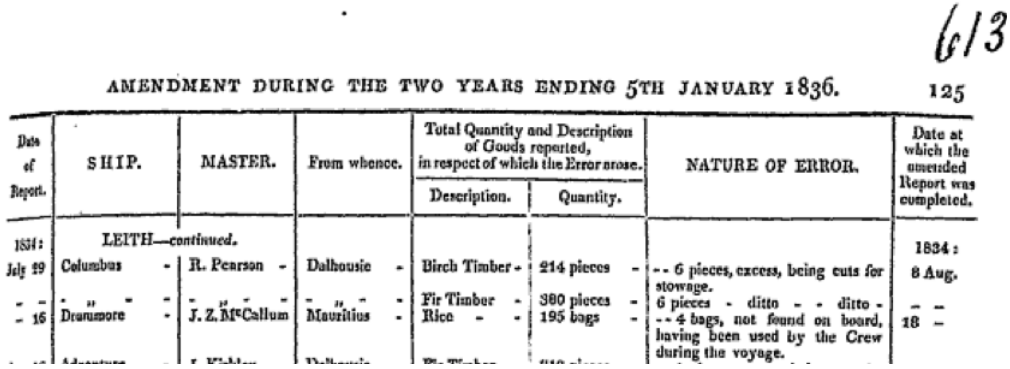
B. Georesolution Experiments

The georesolution step of the Geoparser uses a combination of heuristics such as location feature type, population size, contextual information of location mentions combined with location clustering to disambiguate between multiple locations with the same name in the gazetteer.[1] In the prototype Geoparser integrated at the start of Trading Consequences, features and parameters had been applied based on empirical analysis of georeferenced newspaper text but without methodical parameter tuning for performance optimisation. For example, a cut-off parameter was applied to consider the top 20 locations returned for a given GeoNames search in the case where more results were returned. We first processed the gold standard using the Geoparser with its default settings and compared the output to the manual annotations (see Table 1). Of the 3,109 locations which were resolved by the annotator, 2,586 (83.2%) were correctly resolved (exact match of the GeoNames identifier) and 2,626 (84.5%) fall within a 5km radius of the gold resolution.

A large majority of trading during the nineteenth century was carried out by ship, making locations with ports extremely important in this context. We therefore gave the Geoparser access to a

	Exact Match		Within a 5km Radius	
Feature settings	**# Correct**	**Score**	**# Correct**	**Score**
Default settings	2,586	83.2%	2,626	84.5%
1. port feature	2,528	81.3%	2,565	82.5%
2. increase country feature	2,585	83.1%	2,625	84.4%
3. decrease spot feature	2,585	83.1%	2,625	84.4%
Combination of 1. to 3.	2,601	83.7%	2,638	84.9%
Combination of 1. to 3. and optimised cut-off	2,608	83.9%	2,645	85.1%

Table 1. Georesolution performance of the Edinburgh Geoparser for its default settings, new features and a combination of them on the Trading Consequences gold standard. We report number of correct locations (# Correct) and accuracy scores for two types of evaluation (exact match of GeoNames identifier and occurrence within a 5km radius).



6/3

AMENDMENT DURING THE TWO YEARS ENDING 5TH JANUARY 1836. 125

Date of Report.	SHIP.	MASTER.	From whence.	Total Quantity and Description of Goods reported, in respect of which the Error arose.		NATURE OF ERROR.	Date at which the amended Report was completed.
				Description.	Quantity.		
1834:	LEITH—continued.						1834:
July 29	Columbus -	R. Pearson -	Dalhousie -	Birch Timber -	214 pieces -	-- 6 pieces, excess, being cuts for stowage.	8 Aug.
- -	- „ - -	- „ - -	- „ - -	Fir Timber -	380 pieces -	6 pieces - ditto - - ditto -	- -
- 16	Drummore -	J. Z. M'Callum	Mauritius -	Rice - -	195 bags -	-- 4 bags, not found on board, having been used by the Crew during the voyage.	18 -

Figure 3. Top of page 125 from the Ships' Reports of the House of Commons Parliamentary papers from 1836.[30]

gazetteer of ports (with latitude and longitude values).[31] It contains a list of 1,646 ports collected from early-mid 20th century Royal Navy logs, provided to us by Philip Brohan at the Met Office Hadley Centre in Exeter, which we manually supplemented with 136 additional ports listed in the gazetteer of Colonial and Foreign ports.[32] We adjusted the Geoparser by assigning a higher weight to location candidates within 0.1 degree to a port. For example, the location mention *Dalhousie*, is clearly referring to a port when mentioned in the *From whence* column of a table in the Ships' Reports of the House of Commons Parliamentary papers from 1836 shown in Figure 3. Incidentally, tables, as shown in this example, also have a negative effect of the performance of our text mining tools which are optimised for running text but we will explore this problem in future work.

The previous version of the Geoparser grounded the mention of *Dalhousie* wrongly to Dalhousie in India (GeoNames ID: 1273648; lat: 32.5333300, long: 75.9833300) as a result of the population size heuristic and other factors, such as locations in context and location clustering. The ports-based adjustment means that the correct Dalhousie in Canada (GeoNames ID: 6943599; lat: 48.0550200, long: -66.3847200) is ranked as the top candidate by the georesolution component. However, Table 1 shows that the ports-based adjustment (port feature) deteriorated the resolution on the gold standard. Error analysis showed that the port feature gave too much weight to smaller locations stored in Geonames, which is why we added two new features to overcome this problem. We increased the weight for GeoNames locations of type *PCLI* (independent political entity) which usually refer to countries (see *2. increase country feature*). We also reduced the weight of GeoNames locations of class *S* (spot), including buildings, facilities and farms (see *3. decrease spot feature*). Both features do not damage the performance of the default Geoparser when applied in isolation, but in combination with the port feature they result in a small improvement of 0.5% exact match accuracy.

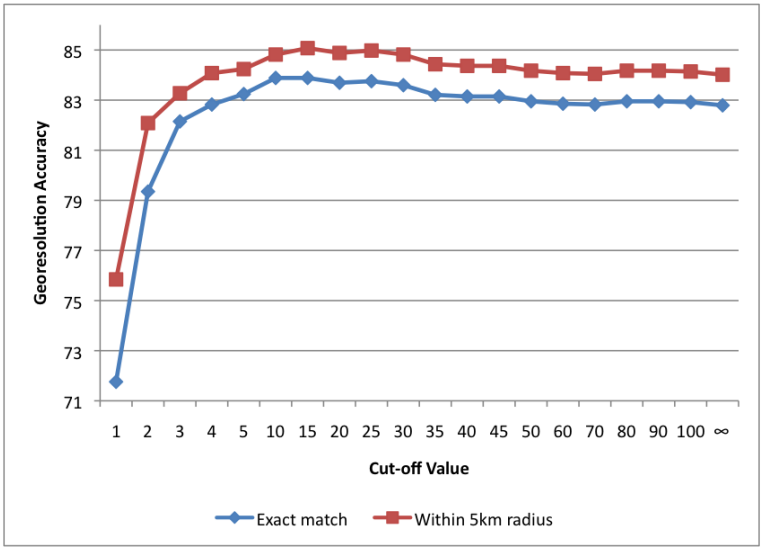


Figure 4. Georesolution accuracy with cut-off values varying between 0 and ∞.

We also optimised the cut-off parameter applied in the Geoparser when retrieving multiple entries from the GeoNames database for one location mention. Figure 4 shows the results we obtained for

varying the cut-off between 1 and ∞. In the default settings, the cut-off was set to 20. The graph illustrates that selecting the first entry extracted from the GeoNames database is not an adequate method to perform georesolution. Not applying a cut-off and considering all possible locations for a given mention also does not result in an optimal performance and it means the resolver needs to work a lot harder when ranking the candidates returned for highly ambiguous location names. The best performance for both types of evaluation is achieved when limiting the number of entries returned from the GeoNames database to 15 before ranking. This results in an overall accuracy of 83.9% for exact match evaluation and 85.1% for evaluation within a 5 km radius. Given the quality of the OCRed text and the historical nature of the Trading Consequences data, these scores are surprisingly high. To put them into perspective, Catherine D'Ignazio et al. (2014) report georesolution scores of 96.3% using the Yahoo Placespotter, 90.3% using OpenCalais and 89.9% using CLAVIN when processing modern news article data from the New York Times, Huffington Post and the BBC.[33]

(2) The GAP Project: Georeferencing Classical Texts

In 2010, the Language Technology Group was approached by the Google Ancient Places (GAP) team who were looking for a tool capable of georeferencing English translations of Greek and Roman classical texts, available as Google Books. The GAP project,[34] funded under the Google Digital Humanities programme, aimed to identify place name 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. This project was the beginning of a collaboration with the members of the GAP team that has spanned several related projects and is still continuing.[3,35,36] The team is international and interdisciplinary, comprising specialists from classics, archaeology, language engineering and visualisation.

The first adaptation needed was to enable the Geoparser to use a gazetteer of the ancient rather than the modern world, namely Pleiades,[37] a freely available scholarly resource run by Sean Gillies and Tom Elliott, of the Institute for the Study of the Ancient World at New York University. The Pleiades team allowed us to take a copy of their entire dataset, which we turned into a relational

database with a schema approximately mirroring that of GeoNames, as this minimised the customisation required in the Geoparser code.

Drawing on the expert knowledge of the classicists on the GAP team, the dataset was expanded to create "Pleiades+" by matching, where possible, the ancient places to their modern equivalents in GeoNames. This provided much more precise latitude/longitude positioning and also added alternative spellings or representations of the place-names in many cases. At run time we introduced a further enhancement using GeoNames (the "Pleiades++" step), for cases where a place name candidate found by the Geoparser was not present in Pleiades+. In these cases we checked the candidate against GeoNames, to collect alternative names that could then be sourced in Pleiades+. In all cases Pleiades+ was the sole source for successful candidate place names, as we only want places existing in the ancient world. An example may make the Pleiades++ step clearer. Translators will often replace the names of well known places with their modern equivalents, so a Google Book text in translation might mention *Egypt*. However, Pleiades only contains *Aegyptus*, the equivalent ancient name. Looking up *Egypt* in GeoNames produces *Aegyptus* as one of the alternative names, and hence we are led to the correct entry in Pleiades+.

This project raised other issues that are relevant to how feasible it is to adapt the Geoparser for widely varying texts. Just as in the Trading Consequences work described above, it proved necessary to disambiguate personal and location names. In the geotagging phase of the Geoparser pipeline, lexicon lists of personal names and location names are used to help determine whether a candidate entity should be categorised as a place or a person. For the GAP project both of these lexicons had to be tailored for ancient texts. For example, *Paris*, *Priam* and *Medea* are obviously people in this context, whereas in a modern text they are probably places. This means not only that suitable lexicons of common ancient personal names had to be used but that the standard lexicons in the Geoparser had to be switched off as they reduced classification performance when included. The input texts for GAP were mainly Google Books, though some Open Library[38] texts were included to test the adaptability of the pipeline. These texts are typically quite untidy, being scanned and OCRed on a large scale. Some pre-processing was done to remove extraneous characters and the books were divided into smaller chunks (typically chapters). These processes were made as

generic as possible, but it is difficult to split an arbitrary text into smaller pieces in a coherent manner without some hand-tailoring. The successor projects to GAP wish to process complete books with minimal user intervention, which raises yet further questions. As explained in the introduction, the clustering algorithms of the georesolution step may not be appropriate if the context is unreasonably large: an entire book rather than a single chapter, say.

Because the GAP project worked with raw unannotated text it was not possible to produce normative evaluations of the geoparser's performance over the texts processed, nor was such evaluation one of the objectives of this humanities project. However, some form of benchmark was required, in order to test improvements during the configuration phase. For this we used the output of an earlier project, Hestia[39,40] to gauge accuracy over comparable ancient text. The Hestia project used a hand-annotated version of Herodotus' *Histories* from the Perseus Digital Library.[41] The precision and recall scores for place name recognition over this text were 77.74% and 95.58% respectively, giving F-score of 85.74%.[42] It was only possible to evaluate the geoparser's first step of geotagging by this method, as we had no gold standard for the georesolution step.

One of the products of GAP was the GapVis online interface[43] illustrated in Figure 5. This presents a selection of classical texts and is intended to assist scholarly interpretation of the ancient world. The user can choose from the "Book Summary" or "Reading" views, or examine a chosen place in detail. The summary view shows the distribution of place names throughout the text, giving an overview of the key locations relevant to the text. The Reading view is that shown in Figure 5, where the text is presented beside a map showing the locations of places mentioned. A scrolling bar beneath the map allows the user to move forwards and backwards through the pages of the text, seeing the places come in and out of focus on the map as they are mentioned in the narrative. The "Place Detail" option gives a network diagram showing possible relationships between the chosen location and others, based on co-occurrence frequency of the place names in a moving text window of a fixed size. The interface builds on earlier visualisation work in the Hestia project.

The GapVis interface has recently been evaluated qualitatively by using it in an undergraduate course on the Ancient World at the University of Texas. The georeferenced text makes it possible to set student exercises with detailed questions about where events happened – questions it

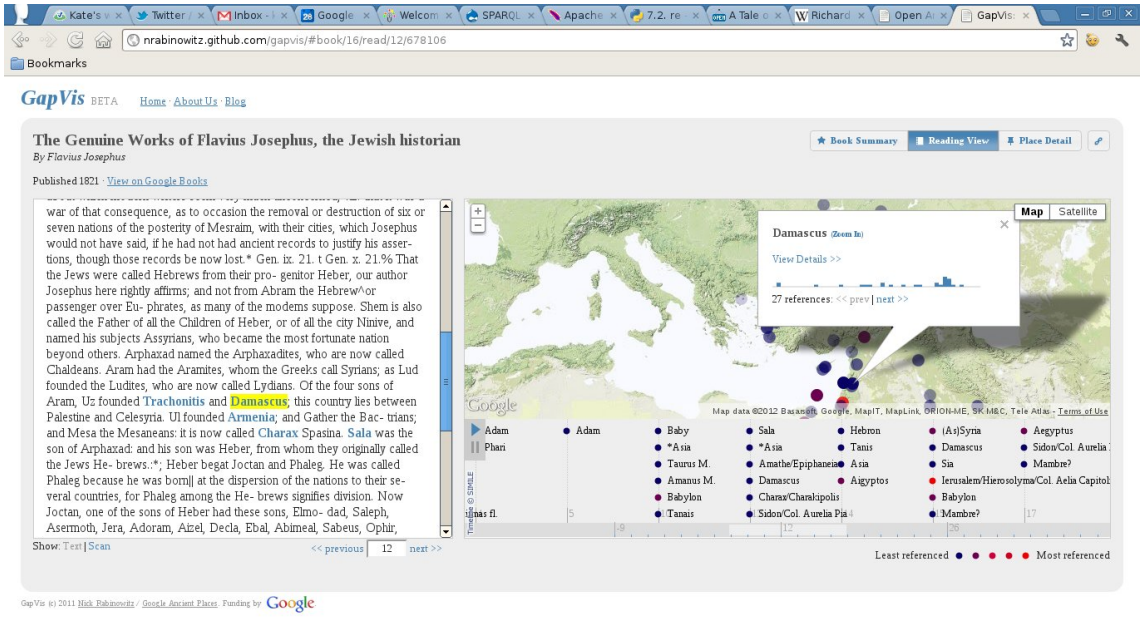


Figure 5. The GapVis web interface.

would be unreasonable to expect to be answered following a traditional book-based reading of the text. Detailed analyses of the results of this case study have been published on the Hestia project blog.[44–46] These posts provide valuable insights into the advantages and shortcomings of automated spatial annotation such as geoparsing, from a humanities perspective. As with all software projects involving a user interface, it is proving difficult to test the underlying functionality as distinct from the user experience – many of the students' queries relate to issues that are not part of the research project, such as problems with operating the interface on a touch-screen.

This set of projects has been an interesting application of the Geoparser. The priorities of a humanities led project have been different, with less interest in formal performance against gold standards, and more in practical use in real-world situations. The fact that high-performing automatic text-processing tools typically achieve precision and recall scores somewhere in the 80s means that up to 20% of the target is mis-identified, and this inaccuracy is sometimes hard for inexperienced users to deal with. Even sophisticated users tend to expect the results shown on screen to be totally correct, and may lose confidence in the entire methodology if they spot an obvious error.

(3) The DEEP Project: Georeferencing Historical English Place-names

The Digital Exposure of English Place-names project (2011-2013) was a JISC-funded collaboration between ourselves, the Institute for Name-Studies in Nottingham, the Centre for Data Digitisation and Analysis in Belfast, the Centre for e-Research at King's College London and EDINA. The project has digitised all 86 volumes of the Survey of English Place-Names (SEPN), the ultimate authority on historic place-names in England. These volumes were compiled over a period of nine decades by the English Place-Name Society and work is still ongoing on outstanding counties. One outcome of the DEEP project is an immensely detailed historical gazetteer for most of the counties in England which can be accessed as a gazetteer service via EDINA's Unlock. It can be also browsed and searched independently.[47] As described earlier EDINA also hosts Unlock Text, a means to access the Edinburgh Geoparser, and we have modified the Geoparser to allow georeferencing of historical documents against the DEEP gazetteer.[48] The Edinburgh Geoparser has thus been used in two ways in the project, firstly to assist in calculating coordinates for all the parishes and other place-names in the DEEP gazetteer, and secondly to allow access to the resulting gazetteer for historical text georeferencing. In the following sections we describe the modifications needed for each of these in turn.

A. Adding Georeferences to the DEEP Data

The LTG's main role in the DEEP project was to transform the output of the OCR process into structured data which can be used for a variety of purposes. Our focus in this section is on the use we have made of the Edinburgh Geoparser to assign georeferences to DEEP place-names and to provide links between historical gazetteer records and their counterparts in the Unlock and GeoNames gazetteers.

The first county survey to be published by SEPN was Buckinghamshire in 1925 with the remaining eighty plus volumes appearing regularly up until the present day. The surveys follow broadly the same format but their appearance over such a long time-span means that there is considerable variation in the type, amount and formatting of information in the volumes. Given the nature and layout of the text, the geotagging component of the Edinburgh Geoparser would have been inappropriate for identifying the place-names so we have instead developed specialised rule sets

for identifying all the relevant pieces of information in the SEPN volumes. We make extensive use of an adapted version of the georesolution component.

FLEET (SY 634805)
Flete 1086 DB, 1212 Fees, 1227 FF, 1235–6 Fees, 1244 *Salis et freq* to 1428 FA, *Fleta* 1086 Exon, 1227 (1372) *ChrP*, 1280 *Ass*, QW, *Flethe* m13 *Salis*
Flote 1086 DB, Exon, *Floeta* c.1160 QuarrCh, *Fleota* 1213 Osm, *Salis*, 1215 (1372) *ChrP*, *Fleote* e13 (1372), 1270 (1372), 1367 (1372) all *ChrP*
Flicta 1157 Sarum
Fluetu 1160–82 QuarrCh, *Fluta* l12 *Salis*, *Flute* 1227 FF
Fleet(*e*) 1646 SC, 1653 *ParlSurv et passim*

From OE **flēot** 'estuary, inlet, creek', with reference to the long channel (still called East and West Fleet) between Chesil Beach and the mainland, see foll.

Figure 6. Excerpt from Survey for Fleet in Dorset.

An typical entry from one of the most recent volumes (Dorset Part 4 published in 2010) is shown in Figure 6. This is the entry for the township of Fleet in the parish of the same name. An OS grid reference (SY 634805) is provided. The entry starts with a list of historical variants of the name where each variant is associated with at least one attestation indicating a historical source in which the name occurred and the date of that source. Thus the first attestation for *Flete* shows it occurring in the Domesday Book (DB) in 1086. It occurs in several other sources up to the last one in 1428 in a source abbreviated 'FA' (*Feudal Aids* in the Public Record Office).

The entry goes on to discuss etymology and then lists smaller places in the vicinity, including East & West Fleet, the inlet alluded to in the extract in Figure 6, Bagwell Barn, Bagwell Barn Cottages, Crook Hill and Fleet Common. After that there is a list of modern field-names followed by a list of historical field-names. Dated, attested historical variants of modern names are provided at all levels from county name through hundreds/wards/wapentakes etc., to parishes, townships, minor names, street names and field-names. These historical names are converted to records in the DEEP gazetteer along with their date, source and latitude/longitude. The modern names in SEPN are also included in the DEEP gazetteer.

The example given above is one where the volume itself provides authoritative georeferencing but, in fact, only a minority of SEPN volumes contain grid references. There is however, a second authoritative source of this information, the Key to English Place-Names (KEPN) database developed and maintained by INS.[49] In creating the DEEP gazetteer we have used the Edinburgh Geoparser to aggregate information from the volumes, the KEPN database, Unlock and GeoNames in order to provide highly accurate, multi-faceted georeferencing focused on the parishes and the major places within them. By preserving the containment relationships between the larger and smaller places, we can allow smaller places without authoritative georeferences to share the georeference of their containing place.

```
<mads ID="epns-deep-86-b-subparish-000004">
 <authority ID="epns-deep-86-b-name-subparish-000004">
   <geographic valueURI="http://placenames.org.uk/id/placename/86/000831">Fleet
   </geographic>
 </authority>
 <related xlink:href="#epns-deep-86-a-parish-000004" type="broader">
  <geographic>Fleet</geographic>
 </related>
 <variant ID="epns-deep-86-b-name-w42045">
  <geographic valueURI="http://placenames.org.uk/id/placename/86/000832">Flete
  </geographic>
 </variant>
 <variant ID="epns-deep-86-b-name-w42126">
   <geographic valueURI="http://placenames.org.uk/id/placename/86/000833">Fleta
   </geographic>
 </variant>
 ...
 <recordInfo>
  <recordCreationDate>2013-10-23</recordCreationDate>
  <recordContentSource valueURI="http://epns.nottingham.ac.uk/England/Dorset/Uggescombe
  %20Hundred/Fleet/Fleet"/>
 </recordInfo>
 <extension>
  <geo source="epns" lat="50.6224698" long="-2.517456837"/>
  <geo source="kepn" kepnref="11636" lat="50.6231" long="-2.51733"/>
  <geo source="unlock" gazref="unlock:11135945" lat="50.62247651321654"
  long="-2.5160432262957 89"/>
  <geo source="geonames" gazref="geonames:7295017" lat="50.6199200"
  long="-2.5273800"/>
  <attestation variantID="epns-deep-86-b-name-w42045">
   <date end="1086" begin="1086" subtype="simple">1086</date>
   <source style="" id="do165">DB</source>
  </attestation>
  <attestation variantID="epns-deep-86-b-name-w42045">
   <date end="1212" begin="1212" subtype="simple">1212</date>
   <source style="" id="do267">Fees</source>
  </attestation>
  ...
 </extension>
</mads>
```

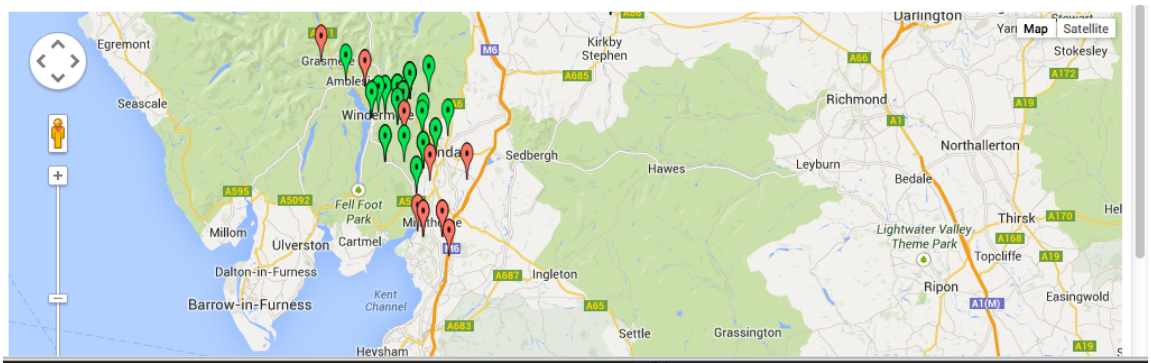
Figure 7. Extract of MADS record for Fleet.

The SEPN-text-to-structured-data process results in output files in MADS.[50] A cut-down version of the MADS for the example in Figure 6 is shown in Figure 7. The <geo> elements in the <extension> element contain the georeferencing for the subparish Fleet, and for its historical variants. This place has the maximum number of <geo> elements: one derived from the grid reference in Figure 6 (source="epns"), one derived from the KEPN database (source="kepn"), and two more created by using the Geoparser to select the most likely records from Unlock and GeoNames (source="unlock" and source="geonames"). The coordinates are all slightly different but they each approximate the position of the historical names associated with Fleet. When the DEEP data is ingested into the Unlock service, one of the sets of coordinates has to be treated as primary, and the preference order for selecting the source of the primary coordinates is epns, kepn, unlock, geonames. In cases where there are no <geo> elements, an entry is given the coordinates of the closest containing element in the hierarchy. Note that the presence of coordinates from multiple sources provides a sort of linking between the sources and it would be relatively straightforward to convert the MADS format of the DEEP data into proper linked data.

In order to achieve multiple georeferencing, we needed to make a number of extensions to the Geoparser for the DEEP system, including implementing a mapping from modern OS grid references as well as older OS sheet-number grid references to latitude/longitude coordinates. We have implemented "known-lat", "known-long" and "known-gridref" parameters and heuristics to allow the georesolution component to be provided with known coordinates and to weight the ranking of Unlock or GeoNames records to strongly prefer those close to the known coordinates. In addition, we have extended the gazetteer look-up output to include information about distance to the known coordinates so that we can discard any Unlock or Geonames records that are not within a reasonable distance of the KEPN record that is our authoritative source of information. In this way we compute highly accurate links between the historical gazetteer and entries in modern gazetteers. Moreover, where KEPN lacks information, the links to Unlock obtained via georesolution can provide the missing information and for smaller places can sometimes provide more accurate coordinates.

B. Using the DEEP Gazetteer in the Geoparser

In order to use the DEEP gazetteer as the source of information for georeferencing historical documents, it was necessary to make alterations to both the place-name recognition and the geo-resolution components of the Geoparser. To give a flavour of some of the issues involved we illustrate the discussion of this work with reference to Figure 8. This figure shows a visualisation of the results of the Geoparser on an input text which is a sample taken from Farrer and Curwen (1923),[51] a collection of summaries and transcripts of documents for townships of the parish of Kendal, accessed via British History Online.[52]



Staveley, which William de Lancastre and Helewise his wife gave to him and his heirs, with liberty to make improvements between the highway and Kent to the bounds of Bolteston and common of pasture with the grantor's men who dwell around, and that he, his heirs and their men shall be quit of pannage, saving the grantor's full multure and hunting; to hold by rendering yearly 3s. of farm and 16d. of cornage, half at Easter and half at Michaelmas, for all service except foreign service. Witnesses: Gilbert de Lancastre, Lambert de Bussei, Gilbert de Lancastre, younger, Adam Gernet, William de Windesore, Gamel the forester, William son of Geoffrey, Walter . . . ., William son of Ketel; Reg. of Deeds at Levens, f. 192.

1256 Perambulation of the boundary between the land of Peter de Brus in Hogayl, Foulbarg and Fayrhayt (Fairbank in the vill of Stirkeland Ketel) and the land of Walter de Lyndesey in Applethwayt and Micheleselet (now Misslett). The bounds commenced at Bredestrete as far as the burial place for dead men, and from that burial place as far as Whystoner next Wodewardehowe. And from thence as far as to Banerhowe in Thwaytlenkyld, and from thence as far as the south door of the barn of Adam de Porta. And from thence in a straight line to Hoggehalebek. And from thence in a straight line between Hethementer and the house of William de Porta, and so as far as the fountain of Gilbert de Lancaster, and so next the fountain which is called

STAVELEY	54°24'N 2°49'W	54°23'N 2°50'W	54°23'N 2°
Staveley	54°24'N 2°49'W	54°23'N 2°50'W	54°23'N 2°
Bolteston	54°22'N 2°47'W		
Levens	54°16'N 2°48'W	54°16'N 2°48'W	
Hogayl	54°23'N 2°50'W		
Foulbarg			
Fairbank	54°22'N 2°50'W		
Stirkeland Ketel	54°21'N 2°47'W		
Applethwayt	54°23'N 2°54'W		
Micheleselet	54°23'N 2°52'W		
Misslett	54°23'N 2°52'W		
Wodewardehowe			
Banerhowe	54°23'N 2°53'W		
Thwaytlenkyld			
Hoggehalebek	54°23'N 2°54'W		
Hethementer			

Figure 8. Visualisation of Geoparser output for subsection of *Records relating to the Barony of Kendale*.

The place-name recognition component needs to be able to recognise DEEP historical names in English historical texts, for example *Banerhowe* and *Hoggehalebek* in Figure 8. These names do not occur in the lists of modern place-names that we use in the location recognition part of the Geoparser and if run in 'modern' mode, many of the places are not recognised. In addition,

many subparts of the person names are mistaken for place-names. To address these issues, we first converted all of the DEEP modern and historical names into a lexicon to be used by the recogniser. This step is similar to the way lexicons had to be tailored for ancient texts in the GAP project described above. We used the DEEP lexicon instead of the other location lexicons but left the remainder of the NER component in place. As with the other projects described here, we found it essential to recognise person names in tandem with locations and we also tailored the person name rules to deal more effectively with names such as *Walter de Lyndesey* and *Peter de Brus*. As can be seen from the lower left frame in Figure 8, many of the place-names have been recognised, but some have not. The names *Foulbarg*, *Wodewardehowe*, *Thwaytlenkyld* and *Hethementer* are all field-names in the relevant SEPN volume (vol. XLII, part 1 of Westmorland). *Fayrhayt* and *Whystoner* have been missed by the recogniser. The SEPN volume, and therefore the DEEP data, has the field *Fayrhayk*, instead of *Fayrhayt*. *Whystoner* is not in SEPN but there is a field *Whystan'* mentioned in the same section as *Wodewardehowe*.

The georesolution component looks up the recognised names in the Unlock ingest of the DEEP gazetteer, accessed through the same API as is used for Unlock but with "gazetteer=deep" as part of the query. The run shown in Figure 8 used a prototype version of DEEP in Unlock which does not include the field-names, so they have not been georeferenced. In the visualisation these are the location mentions without links, as links are created from the relevant placenames.org.uk URL returned as part of the response from Unlock. We have implemented a new feature to be used with the DEEP gazetteer that allows the user to specify which SEPN county (or counties) the document is about. In our example we specified "Westmorland" and this caused the gazetteer look-up to reject any records outside of this area. If the user does not wish to use such an absolute constraint, the alternative is to use the standard Geoparser mechanism of weighting more highly those entries which are inside a bounding circle or bounding box.

The place-names in our example are so distinct that there is very little ambiguity for the geo-resolution component to resolve. *Staveley* matches more than one Westmorland record, *Staveley Chapelry* and the settlements *Over Staveley* and *Nether Staveley* as well as the minor places *Staveley stone*, *Staveley Head Fell*, *Staveley Park* and *Staveley-gate*. *Levens* appears twice with the

same coordinates as it occurs in the SEPN volume as both a modern name and a recorded historical variant of that modern name (in 1352 and 1376, Inquisitions post mortem).

A version of the Geoparser adapted to use DEEP is accessible in Unlock Text. We have attempted to fine-tune it on the basis of a small number of test documents chosen because they are among the sources cited by the SEPN editors and are therefore known to contain historical names. It has not been possible to perform a formal evaluation of this version of the Geoparser though we suspect that the range of possible historical input documents is so wide that a one-size-fits-all version in Unlock Text is unlikely to lead to high performance for many users. It may be necessary for users to adapt the Geoparser source for their own needs and they may also benefit from using it in an assisted-curation scenario where the output is manually post-edited.

SUMMARY AND CONCLUSION

The Edinburgh Geoparser has been in development for a number of years and has now become a practical and useful tool for georeferencing many kinds of texts. As the back-end to Unlock Text it is now available to a wide range of users. The API for Unlock Text is evolving in response to requests from projects such as GAP and more of the underlying functionality is gradually being exposed in the API. However, the Geoparser itself is evolving as we, its developers, put it to use in various projects, as illustrated above. It is becoming clear that customisation of the Geoparser is frequently needed to achieve optimal performance in a particular context and this means that there is an issue as to how we can provide a tool that meets everybody's needs. As we take development forward we will need to address this issue. However, the Edinburgh Geoparser has shown its flexibility over very disparate texts and we are optimistic that future versions will continue to support scholars working with a range of texts. The need to disambiguate places from people in different types of text has been found to be an important step throughout our research. Not every geoparsing system may be set up to deal with this task prior to georesolution.

We can conclude that producing a general purpose geoparsing tool that works "off the shelf" with any type of text is difficult given the current state-of-the-art. Developing a geoparser which can be easily adapted to new domains and types of text by users who do not always want to delve deep into the code is therefore crucial for such technology to be used widely in new and

emerging digital humanities research. As there are many benefits of geoparsing texts, it is starting to be recognised as an important method to analyse text in humanities and social science research. Locations are key for connecting separate datasets and can add a new dimension to longitudinal studies. Plotting place mentions on a map givers users a visual connection between quite separate source documents. Geoparsing can also be a very efficient shortcut to linking big datasets, which is notoriously challenging to achieve through close reading of documents, even for domain experts.

ACKNOWLEDGEMENTS

We are greatly indebted to our project partners on Trading Consequences, GAP and DEEP and the project funders (Jisc, AHRC, SSHRC and Google) for making this research possible. Further information can be found on each respective project website.[19,34,47]

END NOTES

[1] C. Grover, S. Givon, R. Tobin, and J. Ball. Named entity recognition for digitised historical texts. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pages 1343–1346, Marrakech, Morocco, 2008.

[2] C. Grover, R. Tobin, K. Byrne, M. Woollard, J. Reid, S. Dunn, and J. Ball. Use of the Edinburgh Geoparser for georeferencing digitised historical collections. *Philosophical Transactions of the Royal Society A*, 368(1925):3875–3889, 2010.

[3] L. Isaksen, E. Barker, E.C. Kansa, and K. Byrne. GAP: A NeoGeo Approach to Classical Resources. *Leonardo Transactions*, 45(1), 2011.

[4] B. Alex and C. Grover. Labelling and spatio-temporal grounding of news events. In *Proceedings of the workshop on Computational Linguistics in a World of Social Media at NAACL 2010*, Los Angeles, CA, 2010.

[5] K. H. Leetaru. Fulltext geocoding versus spatial metadata for large text archives: Towards a geographically enriched wikipedia. *D-Lib Magazine*, 18(9/10), 2012.

[6] Cartographic Location And Vicinity INdexer (CLAVIN). `http://clavin.bericotechnologies.com`.

[7] OpenCalais. `http://www.opencalais.com`.

[8] Yahoo PlaceSpotter,Yahoo BOSS Geo Services. `http://developer.yahoo.com/boss/geo`.

[9] R. Tobin, C. Grover, K. Byrne, J. Reid, and J. Walsh. Evaluation of georeferencing. In *Proceedings of*

the 6th Workshop on Geographic Information Retrieval, GIR '10, pages 7:1–7:8, New York, US, 2010. ACM.

[10] I. Mani, J. Hitzeman, J. Richer, D. Harris, R. Quimby, and B. Wellner. SpatialML: Annotation scheme, corpora, and tools. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, 2008.

[11] Edinburgh Language Technology Group. `http://www.ltg.ed.ac.uk`.

[12] Unlock Text. `http://edina.ac.uk/unlock/texts`.

[13] LT-XML2. `http://www.ltg.ed.ac.uk/software/ltxml2`.

[14] LT-TTT2. `http://www.ltg.ed.ac.uk/software/lt-ttt2`.

[15] J.R. Curran and S. Clark. Investigating GIS and smoothing for maximum entropy taggers. In *Proceedings of the 11th Meeting of the European Chapter of the Association for Computational Linguistics (EACL-03)*, pages 91–98. Budapest, Hungary, 2003.

[16] G. Minnen, J. Carroll, and D. Pearce. Robust, applied morphological generation. In *Proceedings of the 1st International Natural Language Generation Conference*, Mitzpe Ramon, Israel, 2000.

[17] GeoNames. `http://www.geonames.org`.

[18] Unlock. `http://edina.ac.uk/unlock`.

[19] Trading Consequences. `http://tradingconsequences.blogs.edina.ac.uk`.

[20] U. Hinrichs, B. Alex, J. Clifford, and A. Quigley. Trading Consequences: A Case Study of Combining Text Mining and Visualisation to Facilitate Document Exploration. In *Proceedings of DH2014*. Lausanne, Switzerland, 2014.

[21] Trading Consequences' Interlinked Visualisation. `http://http://tcqdev.edina.ac.uk/vis/tradConVis`.

[22] House of Commons Parliamentary Papers, ProQuest. `http://parlipapers.chadwyck.co.uk/home.do`.

[23] Early Canadiana Online. `http://eco.canadiana.ca`.

[24] Foreign and Commonwealth Office Collection, JSTOR. `http://www.jstor.org`.

[25] Confidential Print collections, Adam Matthrew. `http://www.amdigital.co.uk`.

[26] Directors' Correspondence Collection from the Archives, Kew Gardens, available at JSTOR Global Plants. `http://plants.jstor.org`.

[27] B. Alex, Byrne K, C. Grover, and R. Tobin. A web-based geo-resolution annotation and evaluation tool. In *Proceedings of the 8th Linguistic Annotation Workshop (LAW VIII)*. Dublin, Ireland, 2014.

[28] Commodity. A commodity is defined as a natural resource or a lightly processed product.

[29] B. Alex and J. Burns. Estimating and rating the quality of optically character recognised text. In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage (DATeCH 2014)*, pages 97–102, Madrid, Spain, 2014. ACM.

[30] Ships' reports. Return to an order of the Honourable the House of Commons, dated 31 May 1836;–for, a return of the number of ships' reports that required amendment during the two years ending 5th January 1836; the date of each ship's arrival; and the date at which the amended report was completed; stating the nature of the error in each case. In *House of Commons Parliamentary Papers*, 1836. Document id: 1836-016588.

[31] Ports gazetteer used in Trading Consequences available on GitHub. `https://github.com/digtrade/digtrade/blob/master/lexical-resources/ports.csv`.

[32] F. Miltoun. *Ships and Shipping*. Alexander Moring Ltd., De La More Press, 1903.

[33] C. D'Ignazio, R. Bhargava, and E. Zuckerman. CLIFF-CLAVIN: Determining Geographic Focus for News Articles. In *Proceedings of NewsKDD 2014*. New York, US, 2014.

[34] Google Ancient Places. `http://googleancientplaces.wordpress.com`.

[35] L. Isaksen, E. Barker, E.C. Kansa, and K. Byrne. Googling Ancient Places. In *Proceedings of Digital Humanities 2011 (DH2011)*, Stanford, CA, 2011.

[36] E. Barker, K. Byrne, L. Isaksen, E. Kansa, and N. Rabinowitz. The Geographic Annotation Platform – a Framework for Unlocking the Places in Free-text Corpora. In *NeDiMAH workshop at Digital Humanities 2012 Conference (DH2012)*, Hamburg, Germany, 2012.

[37] Pleiades. `http://pleiades.stoa.org/home`.

[38] Open Library. `http://openlibrary.org`.

[39] The Herodotus Encoded Space-Text-Image Archive. `http://hestia.open.ac.uk`.

[40] E. Barker, S. Bouzarovski, C. Pelling, and L. Isaksen. Mapping an Ancient Historian in a Digital Age: the Herodotus Encoded Space-Text-Image Archive (HESTIA). *Leeds International Classical Journal*, 9:1–24, 2010.

[41] Perseus Digital Library. `http://www.perseus.tufts.edu/hopper`.

[42] K. Byrne. Matching lexicons to gazetteers. GAP project blog post. April 18, 2011. `http://googleancientplaces.wordpress.com/2011/04/18/matching-lexicons-to-gazetteers`.

[43] GapVis online interface. `http://nrabinowitz.github.io/gapvis`.

[44] Reading Herodotus spatially in the undergraduate classroom, Part I. Hestia project blog post. June 5, 2014. `http://hestia.open.ac.uk/`

reading-herodotus-spatially-in-the-undergraduate-classroom-part-i.

[45] Reading Herodotus spatially in the undergraduate classroom, Part II. Hestia project blog post. June 8, 2014. http://hestia.open.ac.uk/reading-herodotus-spatially-in-the-undergraduate-classroom-part-ii.

[46] Reading Herodotus spatially in the undergraduate classroom, Part III. Hestia project blog post. June 22, 2014. http://hestia.open.ac.uk/reading-herodotus-spatially-in-the-undergraduate-classroom-part-iii.

[47] The Historical Gazetteer of England's Place-Names. http://placenames.org.uk.

[48] C. Grover and R. Tobin. A gazetteer and georeferencing for historical English documents. In *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2014)*, pages 119–127, Gothenburg, Sweden, 2014.

[49] Key to English Place-Names (KEPN). http://kepn.nottingham.ac.uk.

[50] Metadata Authority Description Standard (MADS). http://www.loc.gov/standards/mads/mads-doc.html.

[51] W. Farrer and J.F. Curwen, editors. *Records relating to the Barony of Kendale: volume 1*. Cumberland and Westmorland Antiquarian and Archaeological Society, 1923.

[52] Records relating to the Barony of Kendale: volume 1. Available at British History Online. http://www.british-history.ac.uk/report.aspx?compid=49295.