Geoparsing Digitised Historical Collections

Claire Grover, Richard Tobin (School of Informatics, University of Edinburgh),
James Reid (EDINA, University of Edinburgh),
Stuart Dunn (AHeSSC, King's College London),
Matthew Woollard (UK Data Archive, University of Essex),
Julian Ball (BOPCRIS, University of Southampton)

1. Introduction

In this paper we report on two JISC-funded projects which aim to enrich the metadata of digitised historical collections with georeferences and other information automatically computed using geoparsing and related information extraction technologies. Understanding location is a critical part of any historical research, and the nature of the collections make them an interesting case study for testing automated methodologies for extracting content. The two projects (*GeoDigRef* and Embedding *GeoCrossWalk*) have looked at how automatic georeferencing of resources might be useful in developing improved geographical search capacities across collections. Using three distinct resource collections and by building two distinct search and browse interfaces, the projects have shown how a new dimension to search can be better exploited and may be added to existing collections with relative ease. These projects provide an exemplar for the broader aim of geo-enablement across the JISC Information Environment, the ultimate aim being to ensure that discovery via geography (the 'where' component of resources which is a cross cutting constant across the majority of collections e.g. country, place name, postcode, parish name etc) becomes more widely embedded into search paradigms within the JISC IE, as it has in the broader search discovery landscape via e.g. Google Maps/Earth, GeoFlickr etc.

2. Background

The projects were concerned with the following three digitised collections:

- The Stormont Papers (http://stormontpapers.ahds.ac.uk), 84 volumes of parliamentary debates from the start of the Northern Irish Parliament in 1921 to the end of Home Rule in 1972;
- Histpop (http://www.histpop.org), the Online Historical Population Reports for Britain and Ireland from 1801 to 1937;
- BOPCRIS 18th Century Parliamentary Publications (www.parl18c.soton.ac.uk), Journals of the House of Lords from 1688 to 1854.

Each of these collections has been separately digitised and OCRed with the output of OCR being in each case a set of XML documents, each conforming to a different schema. These documents are input to the geoparsing technology developed in the School of Informatics at the University of Edinburgh which has been under development for a number of years. The system combines general-purpose XML-based information extraction technology from the LT-XML2 and LT-TTT2 software tools (http://www.ltg.ed.ac.uk/software) with geoparsing-specific sub-components which were developed in collaboration with EDINA as part of the GeoCrossWalk project (http://www.geoxwalk.ac.uk).

As shown in Figure 1, the geoparser has two main components, the geotagger which is responsible for place name recognition and the georesolver which is responsible for georeferencing. The former processes an input text and identifies the strings within it which denote place names. The latter takes the pool of recognised place names as input, looks them up in a gazetteer, either GeoNames (http://www.geonames.org) or GeoCrossWalk (http://www.geoxwalk.ac.uk), and determines for each place name which of the possible referents is the correct one.

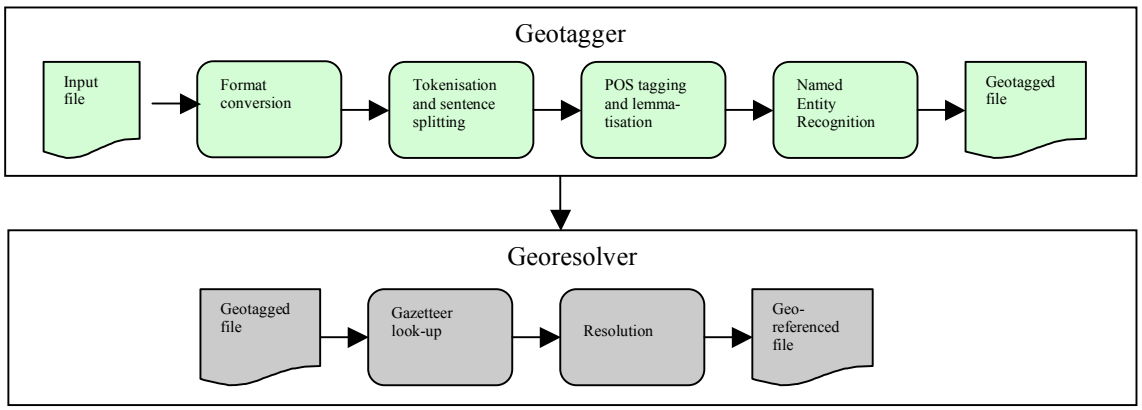

Figure 1. Overview of Geoparser architecture

The original version of the geoparser was developed mainly as a demonstrator and was configured to process web page or newspaper text. The main part of the work was therefore adaptation and extension of the system to allow it to work optimally for the three collections. Although the focus of the project was georeferencing, and thus it was a priority to accurately identify place names within the collections, the system is also capable of recognising person names. The geotagging part of the system is based on general-purpose named entity recognition software which includes the ability to recognise other kinds of entity such as person and organisation names, dates etc. One reason for including person name recognition is because it is easier to achieve accurate place name recognition by also applying the rules for person names to disambiguate cases where a person name contains the name of a place (for example, "Mrs Chichester", "Earl of Essex"). In both projects, therefore, priority was given to recognition of person and place names as well as the resolution of place names to provide georeferencing. The search interface for the *GeoDigRef* project provides both map-based search and 'people' search while the interface for the *Embedding GeoCrossWalk* project also incorporates a timeline which takes advantage of information about the dates on which the debates took place.

The geoparser provides georeferencing with reference to two gazetteers, the Ordnance Survey-derived GeoCrossWalk gazetteer and the open access GeoNames gazetteer. The Histpop and BOPCRS collections were georeferenced twice using each of the gazetteers. Since the GeoCrossWalk gazetteer covers only mainland Great Britain, only GeoNames could be used for the Stormont Papers. The fact that the GeoCrossWalk gazetteer is limited to Great Britain is a problem even for documents whose focus is Britain, since there are likely to be occasional references to other places, and the system will either return nothing or some quite irrelevant place with the same name. To mitigate this we augment GeoCrossWalk with an additional list (derived from GeoNames) of places outside Britain with a population of more than 200,000.

The BOPCRIS data comprises thirteen volumes of the Journals of the House of Lords: Volumes 14-25 (1688-1741) and Volume 50 (1814-1817). For the GeoDigRef project, each volume was split into one page per file giving a total of 9,417 pages/files containing approx 7.5 million words. The OCR output contains *Word* elements around words with attributes capturing the coordinates of each word in the image of the page thus allowing the results of processing to be mapped back onto the image if desired. The Stormont data comprises 84 volumes of parliamentary proceedings. For the Embedding GeoCrossWalk project, each volume was split into one day of proceedings per file, giving a total of 3,315 files containing approximately 67 million words. The Histpop data comprises 25,298 XML files totalling approx 10.5 million words. Each file corresponds to an individual page of the collection.

3. Configuring the Geoparser

Adaptation of the existing geotagging component involved a large number of changes and additions. Some of the more substantial adaptations include collection-specific format conversion, the addition of language recognition software (for BOPCRIS which contains frequent passages in Latin and occasional ones in French), specialisations of tokenisation and sentence splitting for BOPCRIS and Histpop, use of font information for BOPCRIS (since names tend to be italicised) and the addition of specialised lexical information (e.g. person titles such as "Epus", "Dux", "Ds." "Rt. Hon.", "Bt." and terms associated with place such as "Metropolitan Borough", "Soke", "Ward", "Diocese", "M.B.", "R.D." etc.). Furthermore, the place name recognition rules were extended to compute information that may be useful to the georesolver. For example, it may be unclear what the boundaries of some place names ought to be – e.g., "County Down" vs. "Down". Different gazetteers may have different entries in cases like this so we mark up the longer name but keep the shorter name as the value of an attribute on the entity. This allows gazetteer look-up to use the shorter name if there is no match for the longer name. A number of rules were developed to use the linguistic context to allow place names to constrain each other's recognition. For example, a frequent linguistic pattern is "PLACE, PLACE" where the first place name is interpreted as being contained in the second ("London, England"). This pattern can be use to take a decision for cases which would be unclear if they didn't appear in that context (e.g. "Drum, Argyll and Bute" where the clear place "Argyll and Bute" makes it possible to decide to mark "Drum" as a place name. Other similar patterns include coordination ("the rivers Stour, Waveney and Deben"), overt indicators of proximity ("Nuneaton to the north of Coventry") and the use of parentheses ("Coventry (Warwickshire)").

Place names recognised by the geotagger are passed to the georesolver. The first stage of georesolution involves gazetteer-dependent actions such as generating queries in an appropriate format, sending them to the relevant server, and converting the results to a common format, in terms of both structure and vocabulary. Queries are for both the name as it appears in the text and for any alternative names. Since GeoNames queries are just URL fetches ("REST style"), a query is generated and performed for each place name. This is simple but slow because it involves many transactions per document. For GeoCrossWalk, on the other hand, a single query is generated listing all the place names, which is more efficient but requires matching up the parts of the output with the place names from the input. In the output from gazetteer look-up, each place name has been associated with zero or more possible referents and the georesolver ranks these by applying a range of heuristics including, for example, preferring populated places to facilities, preferring larger places etc. We also try to capture the fact that places in a document are often close together. Intuitively we expect many of the places in a document to be

in clusters, so we try to measure that. For each candidate for a place name, we compute its distance from the nearest candidate for each other place name. We then find the average distance to the nearest five other places, and prefer candidates for which this is smaller. Each of the heuristics is scaled to be in the range 0-1, and the scaled values are combined to produce a single score for each candidate.

Collection	Documents	Sentences	Tokens	Location Entities	Person Entities
Histpop	500	9,329	261,676	6,102	298
BOPCRIS	137	7,172	161,583	2,056	5,724
Stormont	11	7,090	173,357	1,263	1,529

Table 1. Overview of the geotagger test sets

To evaluate the system it was necessary to create test data specifically for the collections being processed. Since the geoparser is composed of two main stages, the geotagger and the georesolver, we created test data for both stages. Table 1 shows information about the geotagger test sets. In terms of numbers of tokens they are broadly comparable but they differ in terms of the distribution of entities. Histpop entities are predominantly place names while the majority of BOPCRIS entities are person names (in the BOCRIS data all the lords present are listed at the start of each day's proceedings). The Stormont data shows a more balanced distribution. In the full paper we will discuss differences like these and report evaluation results for both stages of processing. We will also discuss the search interfaces developed for the collections and our vision of the future where georeferencing will provide both the ability to incorporate map based discovery tools and cross searching across collections by the use of geography.