



## Embedding GeoCrossWalk Final Report

Project Information			
<b>Project Title</b>	Embedding GeoCrossWalk		
<b>Start Date</b>	1st October 2008	<b>End Date</b>	30 <sup>th</sup> June 2009
<b>Lead Institution</b>	Centre for e-Research, King's College London		
<b>Project Director</b>	Sheila Anderson		
<b>Project Manager &amp; contact details</b>	Stuart Dunn Centre for e-Research 26-29 Drury Lane London WC2B 5RL		
<b>Partner Institutions</b>	University of Edinburgh, Queen's University Belfast		
<b>Project Web URL</b>	<a href="http://www.kcl.ac.uk/iss/cerch/projects/portfolio/embedding.html">http://www.kcl.ac.uk/iss/cerch/projects/portfolio/embedding.html</a>		
<b>Programme Name (and number)</b>	JISC IE		
<b>Programme Manager</b>	James Farnhill		

Document Name			
<b>Document Title</b>	Final Report		
<b>Reporting Period</b>			
<b>Author(s) &amp; project role</b>	Stuart Dunn (project manager) and Claire Grover (researcher)		
<b>Date</b>	22/07/09	<b>Filename</b>	Embedding_report_final.doc
<b>URL</b>			
<b>Access</b>	<input type="checkbox"/> Project and JISC internal		<input type="checkbox"/> General dissemination

Document History		
Version	Date	Comments
1.0	10 <sup>th</sup> July 2009	First edition
1.1		
1.2		

## Table of Contents

Acknowledgements.....	3
Executive Summary.....	3
Background.....	3
Aims and Objectives.....	5
Methodology.....	5
Implementation.....	7
Outputs and results.....	7
Conclusions.....	8
Recommendations.....	8

## Acknowledgements

This project was funded under the JISC Information Environment programme. It was a collaboration between the Language Technology Group at the University of Edinburgh School of Informatics, and the Centre for e-Research at King's College London, with supporting consultancy for Queen's University Belfast.

## Executive Summary

The Embedding GeoCrossWalk project sought to provide a deeper understanding of how references to place in structured text can be researched and automatically extracted. The text collection used was the Hansards (proceedings) of the Lower House of the devolved Stormont Assembly between 1921 and 1972, usually known (and referred to hereafter) as 'The Stormont Papers'. The project's aims were threefold. Firstly it sought to deploy the Geoparser tool, developed previously by the Language Technology Group of Edinburgh University's School of Informatics, to georeference the Stormont Papers, using Natural Language Processing (NLP). The project used the Geoparser in conjunction with GeoNames, an open-source global gazetteer ([www.geonames.org](http://www.geonames.org)), to identify, tag and (where appropriate) disambiguate all references to location. Secondly, the project refined and developed a better understanding of the Geoparser tool's application to content of this kind, and highlighted . Finally, it laid the foundations for an expanded geospatial browsing capability for the Stormont collections, which will be implemented alongside the existing interface.

The primary research value of the project lay not so much in what it achieved as in the possibilities it has highlighted. For example, location is only one entity that can be identified by the algorithm. Others include person names, roles, etc. Once entities such as these have been identified, the logical next step is to map them in some way. We consider that the Embedding GeoCrossWalk project has highlighted significant future possibilities in these areas.

## Background

The JISC-funded digital version of the Stormont Papers is a complete record, comprising of all eighty-four volumes, and has been made available since 2006 by the Arts and Humanities Data Service, and subsequently by the Centre for e-Research at KCL (see <http://stormontpapers.ahds.ac.uk>). The digitisation was a collaborative exercise between KCL and Queens University Belfast. This is an important resource for the history of the province's political discourse, mainly because the non-digital versions of the Stormont Papers are relatively inaccessible for the research community, still more so for the interested public. The digital medium has been used to make the Stormont Papers available online, and various search facilities provided based on the original indices, as well as Optical Character Recognition (OCR)-derived full text and JPEG images of each page. The Embedding GeoCrossWalk project sought to use this collection to deploy and document a use case for the Geoparser tool, to improve the tool's functionality and provide new possibilities for the Stormont research collection. Understanding location is a critical part of any historical research, and the highly structured nature of Hansards material means that it is a particularly suitable corpus for testing automated methodologies for extracting spatial content.

A further motivation is the reliance of the Geoparser tool on GeoCrossWalk, which is used to disambiguate places with the same names. GeoCrossWalk is, in turn, reliant on Ordnance Survey toponym data to perform the georeferencing and disambiguation, which under current licensing arrangements is not available in Northern Ireland; and in any case is relevant only to the UK. Using GeoNames therefore, which is an open-source global gazetteer, provides an opportunity for trialling Geoparser's use independently of GeoCrossWalk, and therefore with non-UK datasets.

The project has important applications in the immediate future. In 2008, JISC funded the 'Historical Hansards: Completing the Jigsaw' project, whose primary aim is to digitize the Hansards of Stormont's Upper House (Senate) for the same period as the existing Stormont Papers. The outcomes of the Embedding GeoCrossWalk project will be employed to provide a geospatial search facility across both collections, when the latter corpora is digitized.

The Geoparser has been under development for a number of years. The starting point for this project was the version which has been available as a demonstrator at (<http://scargill.inf.ed.ac.uk/geoparser.html>) since February 2008. This combines general-purpose XML-based NLP IE technology from LT-TT2 (<http://www.ltg.ed.ac.uk/software/lt-tt2>) with geoparsing-specific sub-components which the LTG has developed in collaboration with EDINA, as part of the GeoCrossWalk project.

Finally, the project builds on the LTG's Geoparser work with the BOPCRIS (British Official Publications Collaborative Reader Information Service) British Parliamentary Papers Online and GeoDifRef projects. These collections differ from the Stormont Papers in several important ways. For example the BOPCRIS material is considerably older than the Stormont collections. The OCR is therefore problematic, and any NLP approach must take account of this. Using the Geoparser on collections which differ in such ways will provide useful guidance as to its further development and application. Although these issues will be reported on more substantively elsewhere, this report highlights the main issues with applying the Geoparser to the Stormont Papers.

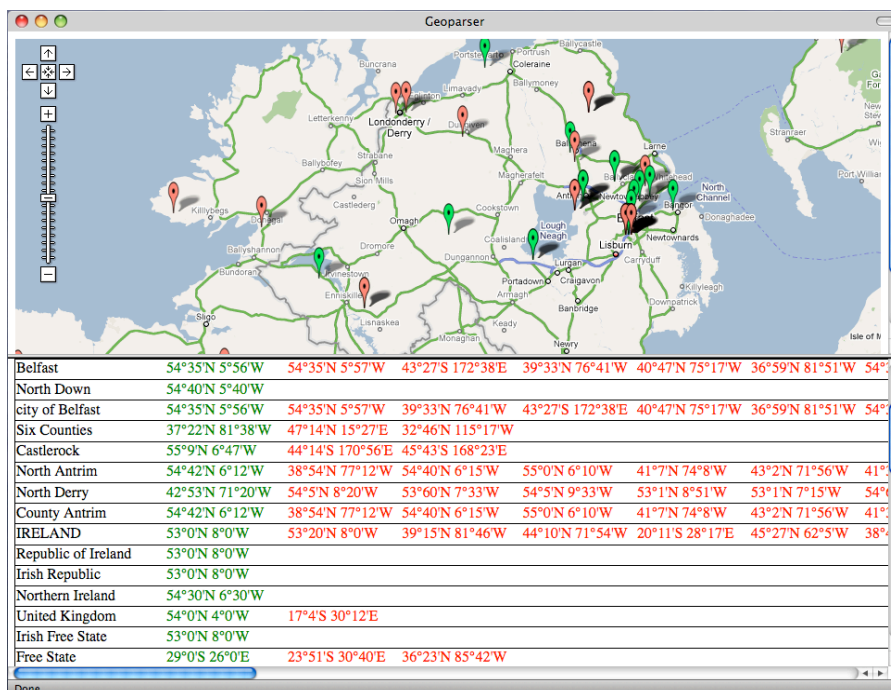


Fig. 1. Spatial locations in the Stormont Papers (from vol 45, 1959) identified with the GeoParser and resolved with GeoNames.

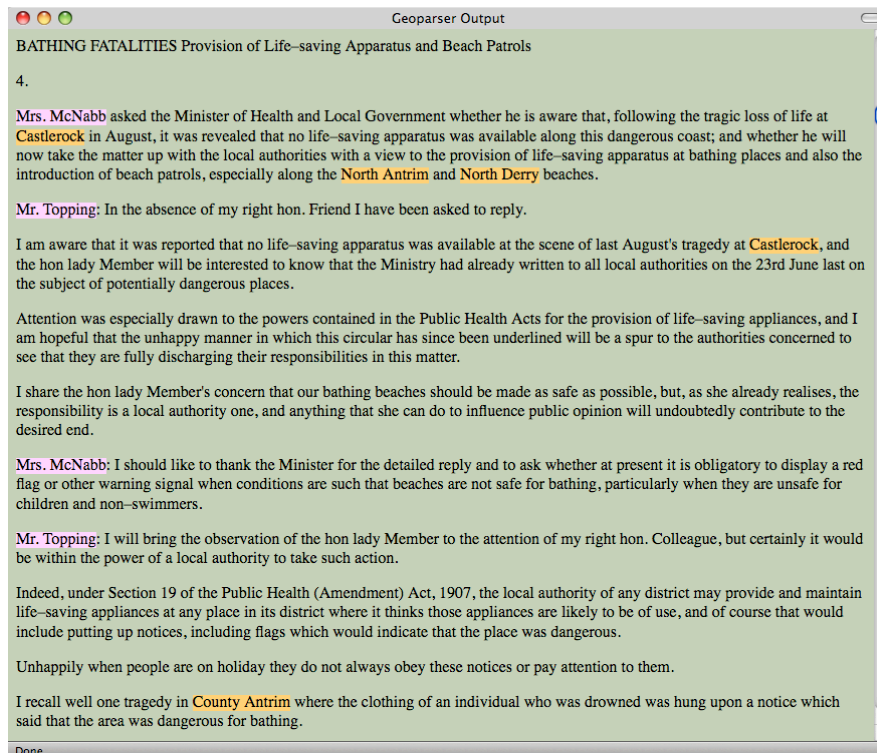


Fig. 2. Entities from the same section identified and highlighted.

## Aims and Objectives

The project's key aims and objectives were:

1. Enriched metadata:
  - To identify toponyms within enriched XML datasets and explicitly georeference them, i.e. adding geographical co-ordinates to the metadata.
  - To establish the viability of GeoNames as a disambiguation reference for non-mainland UK geodata.
2. To produce an end-of-project evaluation and summary report making suggestions for further work if appropriate.
3. To establish an interface for geospatially querying the existing Stormont Papers website, that we will seek to implement in the new Historical Hansards content when available, allowing unified and systematic geospatial cross-searching across the two sets. Due to time constraints, this will be an exemplar interface, which will be rolled out across both collections later.
4. A defined and documented workflow for deploying the GeoParser software with digitized textual content.

The principle outcome was to be an evaluation of the methodology employed for metadata enrichment and an assessment of the broader utility of georeferencing extant digitised resources.

## Methodology

The diagram in Figure 1 provides an overview of the components of the Geoparser. There are 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) or GeoCrossWalk (<http://www.geoxwalk.ac.uk>) and determines for each place name which of the possible referents is the correct one. The system also contains a component which creates a Google Map display of the place names in a document. Note that the geotagger component is based on the LT-TTT2 distribution and that much of the detailed documentation for LT-TTT2 is valid for this application.

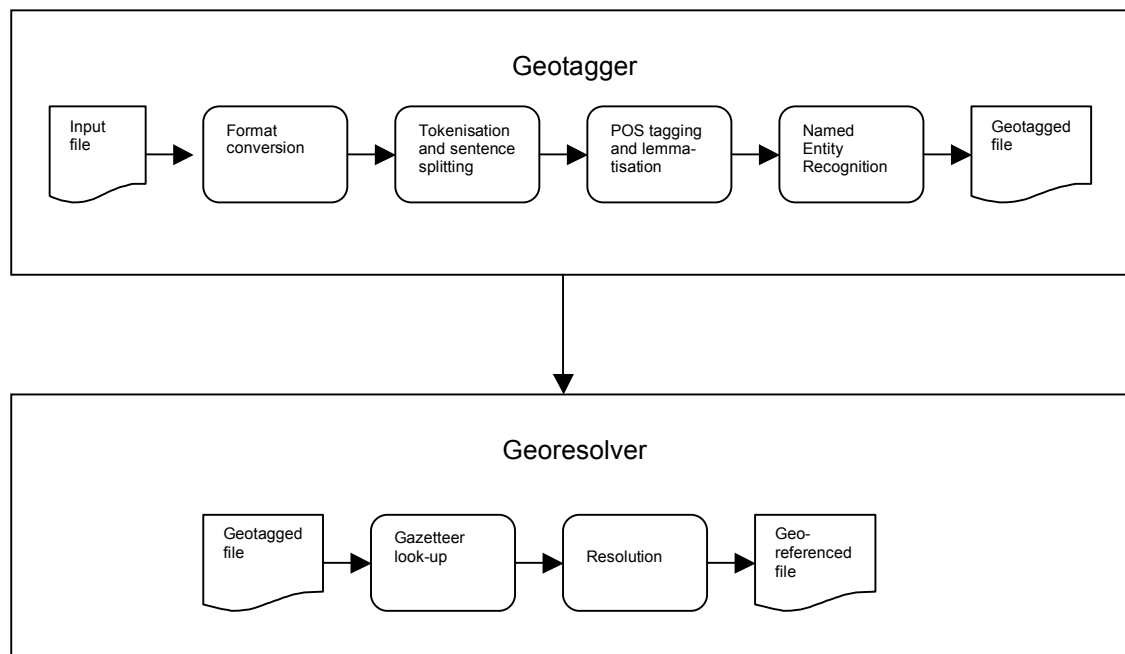


Figure 1. Overview of Geoparser architecture

The system is implemented as a set of Unix shell scripts. The top level script is *geoparse* which is invoked with an argument which is the filename of the document to be processed as well as two parameters. The first parameter specifies the format of the input file (plain, xml or html) and the second specifies which gazetteer should be used (xwalk or geonames). The file and the file format parameter are passed to the geotagger which is implemented as the script *geotag*. This creates an output file (with the extension *.geotagged.xml*) which is passed as the input to the georesolver along with the gazetteer parameter. The first step of the georesolution process is implemented as the script *geogaz* which extracts the place names from the geotagged file, creates gazetteer queries on the basis of the place names, and submits them to the appropriate gazetteer server. The output of *geogaz* is a file (with extension *.gazunres.xml*) which contains all the information returned by the gazetteer server. The script *georesolve* takes this as input and ranks candidate referents for each place name. The output of *georesolve* (a file with extension *.gaz.xml*) is optionally passed to a script called *gazmap* which creates HTML files to allow the results to be displayed in a browser using Google Maps.

The starting point was the version of the Geoparser which has been under development for a number of years as part of the GeoCrossWalk project. That version was developed mainly as a demonstrator and was configured to process general purpose 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 Stormont Papers collection. Although the focus of the project was georeferencing, and thus it was a priority to accurately identify place names within the Stormont Papers, the system is also capable of recognising person names. This is because it is easier to achieve accurate place name recognition by also applying the rules for person names so that disambiguation takes place in the frequent case where a person name contains the name of a place (for example, "Mrs Chichester"). The longer term aim is to incorporate person as well as place search into the interface, and to use the adapted Geoparser to provide a georeferenced back-end for both the Lower House and Senate papers, when

the latter are available. The existing Geoparser provides georeferencing with reference to two gazetteers, the OS-derived GeoCrossWalk gazetteer and the open access GeoNames gazetteer. Since the GeoCrossWalk gazetteer covers only mainland Great Britain, the choice to use GeoNames was easily made (see above, Background). The Geoparser works optimally on documents which are neither too short nor too long as each place name is resolved in the context of all the other place names in the document. For the Stormont Papers, an entire Volume provided too large a context while a single page appeared to be too small. As a solution, we split each Volume into documents each of which covered an entire day of proceedings using the date mark-up that was in the input. A consequence of this decision is that it is now possible to relate mentions of place names to the dates on which they were discussed and hence to provide an interface that can show timelines.

It is important to monitor progress and assess accuracy for applications such as the Geoparser. For this reason a sample of the data was hand-annotated for place and person names so that the Geoparser could be evaluated by comparing system output with the hand annotations. In addition, the georeferencing part of the system needs to be evaluated. For this we hand-annotated the correct gazetteer entry for each place name in the previously annotated set and compared the system's top-ranked choice against the hand-annotated choice.

An interim version of the Geoparser was delivered to CeRch and installed on one of their machines to test that the software could be run in-situ if necessary. In the event, since the georeferencing is a one-off batch process, the data are processed by the LTG and delivered to CeRch.

## Implementation

The technical implementation of the Geoparser mainly involved refining the rules and heuristics that are used for entity recognition and place name resolution. Details of this process can be found in the longer report on the three current applications of the Geoparser (in preparation). In addition to the usual resources, the digitised back indexes of the Stormont Papers were made available so that, for example, lists of person names could be used to improve recognition accuracy. The hand-annotation for the entity recognition and the place name resolution were achieved with customised annotation tools. For the former, the MMAX2 tool was customised to meet our requirements while for the latter we created a purpose-built web-based tool. The guidelines for the annotators are contained in the appendices of the forthcoming report.

The rules which were used to 'train' the Geoparser for the Stormont Papers were originally developed by the History Data Service, and have been used as a starting point for other collections that the Geoparser has worked with. This contributed to the geospatial elements of the development of an XML schema for the current Historical Hansards project, which will be used to mark up the digitized collections. As such, this will greatly facilitate the broader implementation of the Geoparser tool for future Hansard collections. It should be noted in this context that all parliaments in British Commonwealth countries (Australia, Canada, South Africa etc) employ the Hansard reporting system: it will therefore be easy to replicate this methodology with digitized parliamentary records of these countries, where these exist.

## Outputs and Results

Geoparser development outputs are:

- Named entity recognition rules customised for the Stormont Papers data
- Customisation of the georesolution code
- Annotated person and place name evaluation data
- Annotated georesolution data
- Customisation of MMAX2 for entity recognition for the Stormont Papers
- Purpose-built annotation tool for annotation of georesolution
- Pipeline to convert back-index person name lists to lexicon for use by Geoparser

- An exemplar interface which will be further developed once the Historical Hansards digitization is complete.

## Conclusions

At a technical level, our conclusions about the application of the Geoparser tool to the Stormont Papers concur with its application to other collections. Considerable value is added to the collections by providing a specific toponym-based search facility, and the Geoparser facility allows a method for doing this which does not place additional burdens on the user. As with other projects, such as GeoDigRef, the possibilities of georeferencing can extend far beyond simply assigning latitude/longitude coordinates. There is a wealth of possibilities both for extending this methodology to other UK Hansard collections, as will be demonstrated upon completion of the Stormont Senate digitization project; however this approach will also be applicable to other types of parliamentary publication such as Bills, Acts, committee reports, etc.

We also conclude from this project that it is possible to use the GeoNames gazetteer in conjunction with the Geoparser. Other projects have indicated that, when combined with the GeoCrossWalk, the Geoparser's performance is at least favourably comparable to that of free services provided by elements of the 'Informal Geospatial Data Infrastructure, such as the mapping elements of Yahoo (which has recently released a new mapping service), Google and Microsoft. GeoCrossWalk is, however, reliant on the Ordnance Survey database, and is thus constrained to the mainland UK. The OS reference set is not available yet for such purposes for Northern Ireland. Although JISC licensing activities undertaken since the start of this project might change this, The resolution of ambiguous toponyms from GeoNames does not appear to be quite as good as that of GeoCrossWalk. However, given the highly controlled and quality-assured nature of OS data versus the open-source GeoNames, this is not unexpected.

## Implications

Our conclusions imply that the Historical Hansards and Stormont Papers website interfaces can be easily adapted to cope with geospatial searches in more sophisticated way that has been possible previously. We will be seeking to realize these implications in as the HH project proceeds. In significant ways, the implications echo those of the GeoDigRef project: that there is 'intrinsic power' in georeferencing text collections which goes beyond simply providing a map interface. The results of parsing the Stormont Papers through the GeoParser has provoked new questions about how locations are categorized and described, but also how digitization projects should approach metadata with regard to persons, names and roles.

A major implication is that different kinds of text collections will raise different issues when applied to different text collections. There are numerous variables, including the quality of the OCR, the structure of the document, the provenance of the references to location, and the geographic scope of those references. A full understanding of these variables will be needed if wide applicability is to be assured.

## Recommendations (optional)

Again, we echo the recommendations of the GeoDigRef project: JISC should promote some kind of lightweight georeferencing in its digitization projects. The recent efforts add geographic elements to the IE should seek to underpin this.