PHILOSOPHICAL
TRANSACTIONS
OF
THE ROYAL
SOCIETY

Phil. Trans. R. Soc. A (2010) **368**, 3875–3889
doi:10.1098/rsta.2010.0149

Use of the Edinburgh geoparser for georeferencing digitized historical collections

By Claire Grover[1,*], Richard Tobin[1], Kate Byrne[1], Matthew Woollard[2], James Reid[3], Stuart Dunn[4] and Julian Ball[5]

[1] *School of Informatics, University of Edinburgh, Edinburgh EH8 9AB, UK*
[2] *UK Data Archive, University of Essex, Colchester CO4 3SQ, UK*
[3] *EDINA, 160 Causewayside, Edinburgh EH9 1PR, UK*
[4] *Centre for e-Research, King's College London, Strand, London WC2R 2LS, UK*
[5] *Hartley Library, University of Southampton, Southampton SO17 1BJ, UK*

We report on two JISC-funded projects that aimed to enrich the metadata of digitized 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 makes 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. In this paper, we describe the work that was undertaken to configure the geoparser for the collections as well as the evaluations that were performed.

Keywords: geoparsing; georeferencing; natural language processing; information extraction; digitized historical text

1. Overview

Understanding location is a critical part of any historical research, and highly accurate, automatic, geographical referencing promises to allow historians to discover information relating to regions beyond the simple name on which they search. If all digitized collections were georeferenced or 'geo-enabled', this would provide a common unifying aspect that allows for pooling of disparate resources, thereby contributing to the aims of the Linked Data community. The GeoDigRef and Embedding GeoCrosswalk projects were both concerned with georeferencing digitized historical collections. In the GeoDigRef project we worked with two collections, Histpop, the Online Historical Population Reports for Britain and Ireland from 1801 to 1937 (http://www.histpop.org), and BOPCRIS, the Journals of the House of Lords (1688 to 1854) from the

[*] Author for correspondence (grover@inf.ed.ac.uk).

One contribution of 16 to a Theme Issue 'e-Science: past, present and future I'.

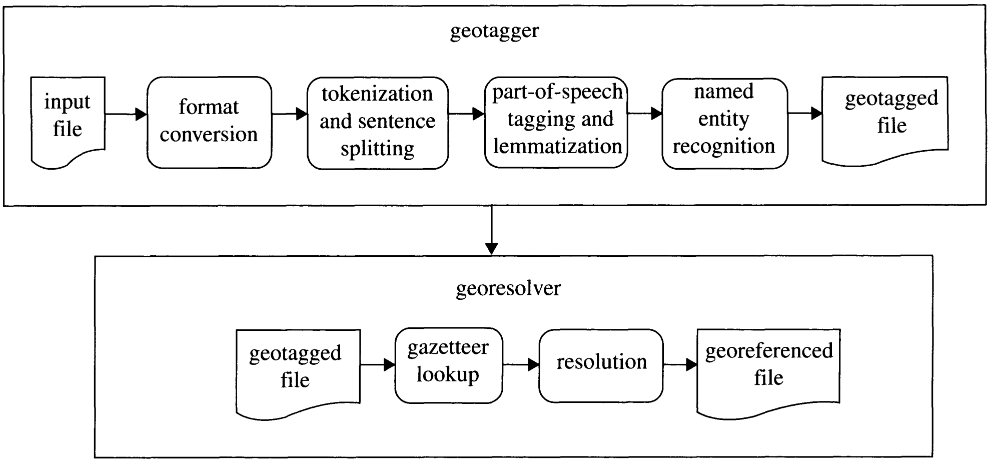

Figure 1. Overview of georeferencing system.

BOPCRIS 18th Century Parliamentary Publications (www.parl18c.soton.ac.uk). In the Embedding GeoCrossWalk project we worked with the Stormont Papers: 84 volumes of parliamentary debates from the start of the Northern Irish Parliament in 1921 to the end of Home Rule in 1972 (http://stormontpapers.ahds.ac.uk).

Each collection has been separately digitized, processed by optical character recognition (OCR) and edited, and exists as a set of XML (eXtensible Markup Language) documents, where each set conforms to a different schema. These documents are input to the geoparsing technology developed in the School of Informatics at the University of Edinburgh. 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; Grover & Tobin 2006) with geoparsing-specific subcomponents that were developed in collaboration with EDINA as part of the GeoCrossWalk project (http://edina.ac.uk/projects/GeoCrossWalk_summary.html). 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 that denote place names. The latter takes the pool of recognized place names as input, looks them up in a gazetteer and determines for each place name which of the possible referents is the correct one. This two-stage architecture is similar to other georeferencing systems, for example, Clough (2005), where the two components are named geo-parser and geo-coder, respectively. At the time of the projects, gazetteer lookup options were either the open-access GeoNames gazetteer (http://www.geonames.org) or the Ordnance Survey-derived GeoCrossWalk gazetteer. Since then, Unlock (http://unlock.edina.ac.uk/) has been added as a replacement for GeoCrossWalk.

The original version of the geoparser was a demonstrator configured for modern text. The current work has therefore involved adaptation and extension of the system to allow it to work optimally for the three collections. (We have, however, been careful to ensure that performance on modern text does not deteriorate: see

Tobin *et al.* (2010) for evaluation on the SpatialML corpus (Mani *et al.* 2008).) Although the focus of the project was georeferencing, and thus it was a priority to accurately identify place names within the collections, the system also recognizes person names. The geotagger is based on a system that also recognizes person and organization names and we decided to retain person name recognition. The reason for this is that it is easier to achieve accurate place name recognition by also applying the rules for person names for cases where a person name contains the name of a place (e.g. 'Mrs Chichester', 'Earl of Essex'). After recognition of person and place names, the georesolver provides georeferencing of place names. 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 that takes advantage of information about the dates on which the Stormont debates took place. The Histpop and BOPCRIS collections were georeferenced twice, using each of the gazetteers. Since the GeoCrossWalk gazetteer does not cover Northern Ireland, it was unsuitable for the Stormont Papers and GeoNames alone was used. For evaluation purposes, we hand-annotated samples of the data (see §§5 and 6).

2. Configuration for the collections

The data from the three collections are the output of OCR on the original documents. The Histpop data comprise 25 298 XML files totalling approximately 10.5 million words. Each file corresponds to an individual page of the collection. The following is an extract from one of the files.

```
<pages>
  <fk_mno>275</fk_mno>
  <page_seq>8</page_seq>
  <ocr_text>
    viii Shrewsbury M.B. and Hereford M.B. are the most populous areas with
    populations of 32,372 and 24,163 respectively. There are two other urban
    areas with populations over 10,000, seven with populations between 10,000
    and 5,000 and ...
  </ocr_text>
  <fulltitle>
    Census of England and Wales, 1931, Counties of Herefordshire and
    Shropshire (Part I)
  </fulltitle>
</pages>
```

The BOPCRIS data comprise 13 volumes of the Journals of the House of Lords: volumes 14–25 (1688–1741) and volume 50 (1814–1817). Each volume was split into one page per file, giving a total of 9417 pages and files containing approximately 7.5 million words. The following is an extract from one of the files, where it can be seen that the OCR output contains *Word* elements around words with attributes x, y, h and w capturing the coordinates of each word in the images of the page. (This allows the results of processing to be mapped back onto the image if desired.)

```
<page number="107">
  <Page>107</Page>
  <Word x="412" y="341" w="142" h="46">Cities</Word>
  <Word x="587" y="339" w="53" h="46">of</Word>
  <Word x="671" y="336" w="174" h="47" font="it">London</Word>
  <Word x="871" y="338" w="83" h="44" font="it">and</Word>
  <Word x="962" y="338" w="170" h="44" font="it">Westm</Word>
  <Word x="1140" y="346" w="89" h="35">to</Word>
  <Word x="1263" y="334" w="215" h="60">expedite</Word>
  ...
</page>
```

The Stormont Papers collection 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 3315 files containing approximately 67 million words. The following is an extract from one of the files.

```
<day value="1932-06-07" id="d31">
  <p><pb n="1549" id="v14p1549"/>HOUSE OF COMMONS.</p>
  <p><date value="1932-06-07">Tuesday, 7th June, 1932.</date> </p>
  <p>The House—which had stood adjourned from Wednesday 1st June—met
    at Twelve noon, Mr. SPEAKER in the Chair.</p>......
  <p>The MINISTER OF FINANCE (Mr. Pollock) (at the Bar), reported
    That His Grace the Governor of Northern Ireland, in the name of and on
    behalf of His Majesty the King, has been pleased to give his Assent to the
    following Bills agreed upon ...</p>...
</day>
```

In order to process texts from the collections it was necessary to make a range of adjustments and additions to some of the geotagger components, as follows.

Format conversion. The data from the collections are provided as XML but each conforms to a very different schema. It was necessary to add collection-specific format conversion for the projects.

Tokenization. This process identifies word tokens and results in *w* elements being wrapped around the tokens (i.e. a word or a punctuation mark). For BOPCRIS the specialization of the tokenizer is more complex than for the other collections because the input file already contains XML markup around words. The original token splitting is not what the pipeline expects: punctuation characters and following white space are included inside *Word* elements (see example above), so retokenization is required to provide tokens that are of the right form.

Sentence splitting. The tokenizer also recognizes sentences and wraps *s* elements around them. For Histpop and Stormont certain abbreviations needed to be added to the list of known abbreviations (e.g. 'Rt. Hon.') to prevent their full stops being interpreted as sentence boundaries. For BOPCRIS, a specialized sentence splitter was implemented partly because of the tendency for semi-colons to be used where full stops are used nowadays and partly because it was convenient to wrap each item in the frequent long lists of person names as a separate sentence.

Language identification. The BOPCRIS data contain frequent passages in Latin and occasional ones in French. It was necessary to identify the language of any given part of the text to prevent the named entity recognition from applying to Latin and French passages. We used Van Noord's language guesser, TextCat (http://www.let.rug.nl/vannoord/TextCat/), applied on a per sentence basis, using the English, French and Latin language models. In BOPCRIS documents, place and person names are only in English text.

Part-of-speech tagging and lemmatization. The POS tagger determines the most likely part of speech (POS). Here, we used the C&C tagger (Curran & Clark 2003) trained on the Penn Treebank (Marcus *et al.* 2000). Lemmatization is the process of finding the stem form of inflected words and for this we used *morpha* (Minnen *et al.* 2000). Neither the POS tagger nor the lemmatizer was changed for the two projects, though a maximum sentence length parameter for the POS tagger was significantly increased (from 250 to 1600 words) to deal with long sentences from Histpop and BOPCRIS.

3. Named entity recognition

The named entity recognition (NER) component is the main component of the geotagger. It is based on the rule-based named entity recognizer in LT-TTT2, though for the two projects it has been configured to recognize only person and place names, disregarding the LT-TTT2 rules for dates, numerical expressions and organization names. Person names in the Histpop collection are very infrequent; however, the BOPCRIS data contain many more person names than place names and a person search facility would be useful. Therefore, the system is configured to recognize both.

The NER component is made up of a number of subcomponents. The first stage is lexical lookup, where words or sequences of words are looked up in a variety of lexicons, including one for common English words, one for forenames, and two geographic lexicons. One of the geographic lexicons is derived from the name list of the Alexandria Digital Library Project Gazetteer (http://www.alexandria.ucsb.edu/), a very extensive world-level gazetteer, while the other is derived from the list of place names in the GeoCrossWalk gazetteer, which provides fine-grained information about Great Britain. Because many place names are ambiguous, such lists must be used with caution. In general, multi-word lexical entries are much more likely to be true place names when encountered in text than single word entries. For example, 'Shepherd's Bush' is a place name composed of English common nouns but, with capitalization and occurring together, these words are unlikely to denote anything other than a place (except as part of a larger name, e.g. 'Shepherd's Bush Empire'). By contrast, single words that can be found in a place name lexicon will frequently not denote a place. For example 'Kendal' is a place name but it could also be a person name or a brand name, so it would be inadvisable to tag every occurrence as a place. It is even more inadvisable to tag very common words that are also place names as places (e.g. 'Best', 'Drum', 'Start', etc.). The following shows an example of XML output from the NER component and illustrates some aspects of the lexical lookup process:

```
<s>
 <enamex type="person">
  <w pername="true" p="NNP" l="john">John</w>
  <w pername="true" p="NNP" l="kendal" locname="single" alsource="true"
     edsource="true">Kendal</w>
 </enamex>
 <w p="VBD" l="live">lived</w>  <w p="IN">in</w>
 <enamex type="location" edsource="true">
  <w p="NNP" l="shepherd" common="true">Shepherd</w>
  <w p="POS">'s</w>
  <w p="NNP" l="bush" common="true">Bush</w>
 </enamex>
 <w p=".">.</w></s>
```

The first lexical lookup stage identifies any multi-word sequences that have been successfully looked up in one of the location lexicons. Thus 'Shepherd's Bush' is wrapped by an *<enamex type="location">* element because it is a multi-word sequence and was found in one of the place name lexicons. The next stage adds attributes to all other words that match entries from one of several lexicons: in the example above, 'Shepherd' and 'Bush' are marked as *common="true"* because they were found in the common noun lexicon, 'John' and 'Kendal' are marked as *pername="true"* because they were found in the forenames lexicon, and 'Kendal' is marked as *locname="single"* because it matched a single word entry in a place name lexicon. Where the match was found in the Alexandria-derived place name lexicon, the words receive *alsource="true"* markup; where it was found in the GeoCrossWalk-derived lexicon, the words receive *edsource="true"* markup. From the example it can be seen that 'Kendal' matched in both lexicons but that 'Shepherd's Bush' was found only in the GeoCrossWalk lexicon.

A special lookup stage was implemented for the Stormont Papers for person names using lexicons derived from the published indices. This proved particularly useful for segmenting complex lists of person names. The lexical lookup stage provides all the information needed by the main NER rules in the next stage of processing, which marks up further *<enamex type="location">* elements as well as *<enamex type="person">* elements. For example, 'John Kendal' in the current example is recognized as a person.

The third stage of the NER component builds a small 'on-the-fly' lexicon of variations on the entities that have already been found, while the final stage uses this lexicon for a second round of lookup followed by application of the final NER rules. The use of the on-the-fly lexicon has the effect of spreading information about entities from clear cases found in the first pass to less clear cases, while the final rules make some decisions about potential single-word place names. If the previous example continued 'Kendal now lives in Morpeth', the final stage would identify 'Kendal' as a person entity because it would be in the on-the-fly lexicon. The on-the-fly lexicon process was used for all the collections except Histpop.

Adaptation of the existing NER component to the four collections involved a large number of changes and additions, many of which involve small details rather than major changes. The following are some of the more substantial changes that needed to be made.

Alternative names. In the evaluation data for Histpop, words such as ‘County’ were included in the markup of place names (e.g. ‘County of Norfolk’, ‘Tyrone Co.’, etc.) and the NER rules were configured to do the same. However, entries in gazetteers may be shorter (‘Norfolk’, ‘Tyrone’) and gazetteer lookup would fail on the longer version. To compensate, the longer name is marked up while the shorter name appears as the value of an *altname* attribute on the entity, allowing the georesolver to use the shorter name if there is no match for the longer name.

Context features. Rules were developed to use the linguistic context to allow place names to constrain each other’s recognition and to supply information to the georesolver. For example, in one pattern a first place name is interpreted as being contained in a second and this can be used to decide cases that would otherwise be unclear (e.g. ‘Drum, Argyll and Bute’, where the clear place ‘Argyll and Bute’ makes it possible 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)’). Features based on the context can be used by the georesolver.

Lexicon and rule additions. The BOPCRIS data required the addition of a number of new titles for persons (e.g. ‘Epus’, ‘Dux’, ‘Ds.’). Similarly the Histpop data prompted the addition of many new terms associated with place names (e.g. ‘Borough’, ‘District’, ‘Soke’, ‘Ward’, ‘Diocese’, ‘M.B.’, ‘R.D.’). For Stormont, rules were included to respond to regular patterns (e.g. ‘The MINISTER OF COMMERCE (Mr Barbour):’).

Font information. In the BOPCRIS texts, italic font is used consistently around names (e.g. ‘An Act to enable *John Keeble* Gentleman to sell certain Lands in *Stow Markett*’). The output of BOPCRIS OCR retains font information, which is extremely useful given that it is difficult to tell proper nouns from common nouns since the latter are usually capitalized.

4. The georesolver

The output of the geotagger is input to the resolution process, which consists of two main stages, lookup of the place names in a gazetteer, and resolution, which ranks the resulting matches. For visualization of the results there is an optional third stage, which uses the Google Maps application programming interface (API) to display the ranked locations.

In the first step of gazetteer lookup the place names are extracted from the geotagged file and duplicate place names are reduced to a single representative. The result is an XML file containing a top-level <*placenames*> element and a <*placename*> child for each unique place name. The place names are then passed to the gazetteer lookup script for the gazetteer being used. Each gazetteer’s script does the lookup in an appropriate way but produces a gazetteer-independent output.

The gazetteer-dependent actions are: 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 (feature type, for example). Queries are for both the name as it appears in the text and for any alternative names (encoded in the *altname* attribute). In the output from gazetteer lookup, each <*placename*> element contains a number of <*place*> elements, which are the candidate places from the gazetteer. These elements have the following attributes: *lat* (latitude); *long* (longitude); *gazref* (an id formed from the gazetteer name and an id returned by the gazetteer); *in-cc* (where available, the ISO country code of the containing country); and *type* (feature type). Each gazetteer has a large set of feature types and we reduce this to a small set for use by the disambiguation code as follows: *water* (river, lake, etc.); *civil* (administrative division); *civila* (top-level administrative division); *country* (country); *fac* (building, farm, etc.); *mtn* (mountain or valley); *ppl* (populated place); *ppla* (capital of top-level administrative division); *pplc* (capital of a country); *rgn* (region); *road* (road, railway, etc.); and *other* (other).

Each gazetteer script is responsible for doing this mapping. After gazetteer lookup, duplicate elimination is done on the candidates for each place name, as the alternative names may have resulted in duplicate results from the gazetteer. The output from gazetteer lookup using GeoNames for the place ‘Morpeth’ is this:

```
<placenames>
  <placename name="Morpeth" id="1">
    <place name="Morpeth" gazref="geonames:2205622" type="ppl" lat="-32.7333333"
      long="151.6333333" in-cc="AU"/>
    <place name="Morpeth" gazref="geonames:2642182" type="ppl" lat="55.1666667"
      long="-1.6833333" in-cc="GB"/>
    <place name="Morpeth" gazref="geonames:6078707" type="rgn" lat="42.383391788"
      long="-81.84977586" in-cc="CA"/>
    <place name="Morpeth" gazref="geonames:885826" type="fac" lat="-19.7"
      long="31.0166667" in-cc="ZW"/>
    <place name="Morpeth" gazref="geonames:973798" type="ppl" lat="-26.9333333"
      long="23.4666667" in-cc="ZA"/>
  </placename>
</placenames>
```

From this it can be seen that GeoNames contains three populated place entries for ‘Morpeth’, one in Australia, one in Great Britain and one in South Africa. It also contains one region entry in Canada and one facility entry in Zimbabwe.

The GeoCrossWalk gazetteer is confined to Great Britain, as is its recent replacement, Unlock. Even for documents about Britain this leads to problems, 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 use an additional list (derived from GeoNames) of places outside Britain with a population of more than 200 000 when using GeoCrossWalk or Unlock.

The second stage of georesolution takes the output from gazetteer lookup and applies heuristics in order to rank the candidate entries. Before applying any of the heuristics described below, we first try to augment the information about each candidate place with information about population and containing country. This is done by consulting lists of large places derived from GeoNames and Wikipedia. If there is a place in the lists with the same name and similar latitude and longitude (within one degree), we assume a match. The information added is containing country (the attribute *in-cc*), if not already present, and population (the attribute *pop*), if available. Most georesolution systems use heuristics to disambiguate place names: Leidner (2007) provides a useful summary of the heuristics used in systems such as Amitay *et al.* (2004), Clough (2005), Rauch *et al.* (2003), Schilder *et al.* (2004) and Smith & Crane (2001). The heuristics that we use are as follows.

Feature type. For example, we prefer populated places to 'facilities'.

Population. We prefer bigger places (for newspaper text, we found that more than 90% could be correctly identified using this alone).

Contextual information. Containment and proximity information may be present in the text and marked up by the geotagger, e.g. the containment relation in 'Leith, Edinburgh'. We strongly favour candidates consistent with such contextual information, i.e. candidates for Leith and Edinburgh that are near each other.

A locality parameter from the user. The georesolver can be called with a parameter that specifies the geographic focus of a document as a latitude, longitude and radius.

Clustering. Intuitively, we expect many of the places in a document to be in clusters and we try to measure this. 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 these heuristics is scaled to be in the range 0–1, using logarithmic scaling for the population and clustering. This scaling has not been thoroughly explored since we do not have sufficient data to experiment properly. The scaled values are combined to produce a single score for each candidate. Again the combination has not been thoroughly explored and in future work we will consider varying the formula in accordance with what we know about the text. For example, we might weight population higher and clustering lower for news articles.

The output of the georesolver is the same list as was input except that the entries for each place are ranked, with the one ranked number 1 as the preferred reading. Features computed for use by the heuristics are also present in the output.

5. Geotagger evaluation

It is important to be able to report the quality of a system's performance in concrete, quantifiable terms. We follow standard practice in comparing system output to hand-annotated 'gold standard' evaluation data. In the case of Histpop and Stormont, completely new data were created. For BOPCRIS there was a pre-existing named entity recognition evaluation set (Grover *et al.* 2008) but this consisted of annotations on the output of an earlier OCR stage and was not directly re-usable; however, it was possible to transfer the previous annotations semi-automatically to the new OCR output. The geotagger test sets were manually annotated for person and location entities. The georesolver test sets were the same data where the human annotated location entities had been

Table 1. Overview of Histpop, BOPCRIS and Stormont test sets.

collection	documents	sentences	tokens	entities	
Histpop	500	9329	261 676	*location*	5890
				person	300
				total	6427
BOPCRIS	92	5486	102 851	*location*	1181
				person	4809
				total	5990
Stormont	12	7601	185 503	*location*	1216
				person	1634
				total	3055

manually resolved twice, first using the GeoCrossWalk gazetteer and then using the GeoNames gazetteer. For both stages, annotation guidelines were drawn up. Ideally some of the data would have been doubly annotated in order to monitor annotation quality through the calculation of inter-annotator agreement. Unfortunately, project resources did not allow this.

Table 1 shows information about the three geotagger test sets. Each Histpop document is a randomly OCRed page, while the BOPCRIS documents are pages randomly selected from volumes 14 and 50 of the Journals of the House of Lords. The Stormont set contains a small number of randomly chosen documents, but since each represents a day of proceedings it can contain several pages—the 12 documents contain a total of 471 pages. On average, a BOPCRIS page is twice the length of a Histpop page and a Stormont document is more than ten times the length of a BOPCRIS page. The BOPCRIS set contains nearly as many entities as Histpop even though it contains less than half the number of tokens, and the Stormont data are comparatively sparsely populated with entities. Table 1 also provides a breakdown of the entities. Histpop entities are predominantly place names; the majority of BOPCRIS entities are person names, and the two kinds of entities are more evenly balanced in the Stormont data. One cause of the large number of person names in BOPCRIS is the listing of all the Lords present at the start of each day's proceedings.

To assess the performance of the geotagger, the system is run over the documents in the test sets and its output is compared with the gold standard human annotations. Results are reported in terms of precision and recall, where precision is the percentage of system-predicted entities that were correct, and recall is the percentage of gold standard entities that the system correctly identified. It is usually important for an application that a balance be struck between precision and recall. F-score is the harmonic mean of precision and recall ($F = 2 \times (\text{precision} \times \text{recall})/(\text{precision} + \text{recall})$) and gives an idea of overall system performance. Table 2 shows the geotagger evaluation results.

Table 2 shows that the results for Histpop and Stormont are considerably better than the results for BOPCRIS. Furthermore, in each collection there are differences in accuracy between the *location* and *person* entities, with the worst *location* score being in the BOPCRIS data and the worst *person* score being in Histpop. The main factors that contribute to the pattern of results relate to

Table 2. Geotagger evaluation results.

		precision (%)	recall (%)	*F*-score
Histpop	*location*	82.09	80.78	81.43
	person	53.07	59.51	56.11
	total	80.77	79.93	80.34
BOPCRIS	*location*	55.92	61.56	58.61
	person	81.83	82.57	82.20
	total	76.35	78.43	77.38
Stormont	*location*	71.72	74.67	73.17
	person	85.71	88.55	87.10
	total	79.64	82.55	81.07

Table 3. Entity frequencies in the corpora.

Stormont		Histpop		BOPCRIS	
Northern Ireland	298	Scotland	476	Ireland	107
Belfast	120	England	286	Earl of Shaftesbury	51
Ulster	101	Wales	171	Great Britain	42
Mr Grant	69	London	142	Comes Mulgrave	32
Mr Diamond	43	Ireland	128	Ds. Maynard	31
Rev. Dr Paisley	41	Edinburgh	80	Epus. London	29
Mr F. V. Simpson	40	Glasgow	74	Ds. Colepeper	29
Mr Donald	39	United Kingdom	61	Scotland	29
Mr McGuffin	37	Perth	56	Epus. Winton	28
Mr Henderson	36	Dundee	50	Ds. Cornwallis	27
Mr Andrews	36	Aberdeen	49	Comes Rochester	27
United Kingdom	32	England	48	England	27
Great Britain	30	Scotland	46	Epus. Sarum	26
England	29	Leith	46	Comes Carnarvon	26
Londonderry	28	Greenock	46	Ds. Lucas	25
total	979 (32%)	total	1759 (27%)	total	536 (9%)

the differing relative frequencies of the entities in the original texts, mismatches between the extents of gold and system entities, and difficulties in distinguishing person from place.

We have calculated type : token ratios for the entities in the collections to discover whether there are differences in lexical variation. The least varied set is Stormont (1 : 4, 25% approx.), Histpop is in the middle (1 : 3, 33% approx.) and BOPCRIS is the most varied (1 : 2.4, 42% approx.). Table 3 shows the top 15 most frequent entities in each corpus with their counts. In the Stormont corpus these account for about 32 per cent of the tokens, while for Histpop the proportion is 27 per cent and, at the other extreme, for BOPCRIS it is 9 per cent. The relative distributions of *location*/*person* entities reflect the distributions shown in table 1: BOPCRIS has more *person* entities, Histpop has more *location* entities and Stormont is roughly balanced.

Another factor concerns overlap in the extent of entities. If a long gold standard entity is compared with a system entity that covers some of the same string but is shorter, then the system entity is penalized even if it is plausible. An example is the gold entity 'Ann Dowager Baroness Southampton' where the system outputs two entities, 'The Right Honourable Ann' and 'Baroness Southampton'. This counts as two false positives and one false negative when in fact it is not completely wrong. OCR quality also has repercussions: the annotators marked up entities that had been mangled by OCR as if they were not mangled, and this led to cases where it would be extremely hard for the system to perform well. For example, 'tomes Rochester' is an OCR misrepresentation of 'Comes Rochester' (Comes = Count). The result was that the system did not recognize 'tomes' as a person title, so it failed to recognize a person entity and instead marked 'Rochester' as a place.

6. Georesolver evaluation

In order to assess the accuracy of the georesolution part of the system, it was necessary to hand annotate data. There are some freely available georesolution test sets (e.g. Leidner 2007; Mani *et al.* 2008) but these are for newspaper text and would not properly reflect the performance of the system on the GeoDigRef and Embedding GeoCrossWalk collections. To create the test data, we took the gold standard annotated test data for the geotagger and hand annotated the correct interpretation for each place name. For Histpop and BOPCRIS we did this twice, once using the GeoCrossWalk gazetteer and a second time using GeoNames. For Stormont, we produced gold georesolution annotation only using GeoNames as GeoCrossWalk does not cover Northern Ireland.

The georesolver was evaluated against the test data under the following conditions: (i) the input was the gold standard entity markup to ensure evaluation of only the georesolver and not the full pipeline; (ii) the gazetteer entries for the resolver to rank were exactly the entries that were available to the human annotators; (iii) the gold standard entity markup was augmented with linguistic context features (the *altname*, *contains*, etc. attributes) in order to provide all the information that the georesolver would normally work with; and (iv) georesolution was tested both with the user supplied locality parameter switched off and switched on. For Histpop and BOPCRIS the setting was '55.45 −5.2 655' (to cover the British Isles) and for Stormont the setting was '54.6 −6.8 92' (to cover Northern Ireland). Two kinds of comparison were considered: strict matching, where gold and system choices should be identical (i.e. have the same id), and within 5 km matching, where gold and system choices would be counted the same if their grid references were within 5 km of each other. The latter is useful for cases where the gazetteer has more than one entry for essentially the same place, e.g. a populated place entry as well as administrative district entry.

During gold annotation, a number of cases arose. For some place names no gazetteer entry was found during gazetteer lookup, while for others there were entries but the human annotator considered that none of them was correct (they selected 'none'). Occasionally entries were found but the human annotator neither chose one nor selected 'none'—we consider these to be annotation errors. In the vast majority of cases entries were found and the human annotator selected

Table 4. Georesolver evaluation.

		GeoNames	GeoCrossWalk
Histpop	place names	5882	5890
	no candidate	424	1203
	'none' selected	349	252
	no selection	18	0
	non-'none' selected	5091	4435
	baseline	1113 (21.9%)	1983 (44.7%)
	strictly correct without locality	3554 (69.8%)	2833 (63.9%)
	strictly correct with locality	3835 (75.3%)	2835 (63.9%)
	correct within 5 km without locality	3875 (76.1%)	4110 (92.7%)
	correct within 5 km with locality	4177 (82.0%)	4112 (92.7%)
BOPCRIS	place names	1181	1181
	no candidate	339	462
	'none' selected	80	43
	no selection	27	26
	non-'none' selected	735	650
	baseline	156 (21.2%)	233 (35.8%)
	strictly correct without locality	494 (67.2%)	515 (79.2%)
	strictly correct with locality	565 (76.9%)	515 (79.2%)
	correct within 5 km without locality	523 (71.2%)	592 (91.1%)
	correct within 5 km with locality	598 (81.4%)	593 (91.2%)
Stormont	place names	1216	
	no candidate	150	
	'none' selected	74	
	no selection	7	
	non-'none' selected	985	
	baseline	480 (48.7%)	
	strictly correct without locality	836 (84.9%)	
	strictly correct with locality	888 (90.2%)	
	correct within 5 km without locality	855 (86.8%)	
	correct within 5 km with locality	907 (92.1%)	

one of them as correct. In the evaluation we exclude all cases except the case where the annotator chose one entry as correct, though in table 4, which shows the evaluation results, we indicate the numbers of each of the cases. We exclude the 'none' case, as the system is not designed to make a 'none of the above' judgement and will always choose one. The vast majority of instances are part of the evaluation and here we look at the first ranked entry from the system. If the first ranked entry is the same entry as the gold entry, then it is correct in the strict sense; if it falls within 5 km of the gold entry, then it is correct in a looser sense. Table 4 also shows the effects of the use of the locality parameter. We have included a baseline, which is the score that would be obtained by randomly selecting entries.

The lower baselines for GeoNames indicate that georesolution is harder with this gazetteer, which contains many more entries from all over the world. In most cases, georesolution performance is worse with GeoNames than with

GeoCrossWalk, a result that reflects the difference suggested by the baselines. In both cases, however, the system achieves good results that are significantly better than the baselines. As would be expected, the strict measure of success gives rise to lower scores than the 'within 5 km' measure. When using GeoNames, the use of the locality parameter results in higher scores. Its use with GeoCrossWalk makes little difference since it provides the British Isles as the locality and GeoCrossWalk is confined to Great Britain anyway. There are relatively large numbers of 'no candidate' place names for BOPCRIS, and an examination of a sample of these suggests that many are OCR errors (e.g. 'County of flits') or possibly older spellings (e.g. 'Materdale').

7. Conclusions

We have described the development and evaluation of a geoparsing system for georeferencing digitized historical collections. The system has been integrated into two distinct interfaces: the search interface for the GeoDigRef project (http://unlock.edina.ac.uk/geodigref/search) provides both map-based search and people search, while the interface for the Embedding GeoCrossWalk project (http://kcl.ac.uk/iss/cerch/projects/portfolio/embedding.html) also includes a timeline that takes advantage of information about the dates of the Stormont debates. The results shown above demonstrate the robustness of the system and suggest that the automatic georeferencing of numerous other collections is a practical possibility. While the geoparser can be used in its current form with a minimum of effort, we have shown that the types of historical source and the quality of OCR can significantly affect the results of the process—for this reason, a degree of tuning may be needed to achieve the best results on a new collection. In future work we hope to continue to test and develop the system to improve performance on a wider range of data, as well as to demonstrate the practicality of using georeferencing to link across document collections. We also hope to use the geoparser to mine a variety of historical resources to create a new gazetteer that would contain more information about place name changes over time.

References

Amitay, E., Har'El, N., Sivan, R. & Soffer, A. 2004 Web-a-where: geotagging web content. In *Proc. 27th Annu. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, Sheffield, UK*, pp. 273–280. New York, NY: ACM Press. (doi:10.1145/1008992.1009040)

Clough, P. 2005 Extracting metadata for spatially-aware information retrieval on the internet. In *Proc. 2005 Workshop on Geographic Information Retrieval, Bremen, Germany*, pp. 25–30. New York, NY: ACM Press. (doi:10.1145/1096985.1096992)

Curran, J. R. & Clark, S. 2003 Investigating GIS and smoothing for maximum entropy taggers. In *Proc. 10th Meeting of the European Chapter of the Association for Computational Linguistics* (*EACL-03*), Budapest, Hungary, pp. 91–98.

Grover, C. & Tobin, R. 2006 Rule-based chunking and reusability. In *Proc. 5th Int. Conf. on Language Resources and Evaluation* (*LREC 2006*)*, Genoa, Italy*, pp. 873–878.

Grover, C., Givon, S., Tobin, R. & Ball, J. 2008 Named entity recognition for digitised historical texts. In *Proc. 6th Int. Conf. on Language Resources and Evaluation* (*LREC '08*)*, Marrakech, Morocco, 28–30 May* (eds N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis & D. Tapias), pp. 1343–1346. Paris, France: European Language Resources Association (ELRA). See http://www.lrec-conf.org/proceedings/lrec2008/.

Leidner, J. L. 2007 Toponym resolution in text: annotation, evaluation and applications of spatial grounding of place names. Technical report, PhD thesis, School of Informatics, University of Edinburgh.

Mani, I., Hitzeman, J., Richer, J., Harris, D., Quimby, R. & Wellner, B. 2008 SpatialML: annotation scheme, corpora, and tools. In *Proc. 6th Int. Conf. on Language Resources and Evaluation (LREC'08), Marrakech, Morocco, 28–30 May* (eds N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis & D. Tapias), pp. 410–415. Paris, France: European Language Resources Association (ELRA). See http://www.lrec-conf.org/proceedings/lrec2008/.

Marcus, M. P., Santorini, B. & Marcinkiewicz, M. A. 1993 Building a large annotated corpus of English: the Penn Treebank. *Comput. Linguistics* **19**, 313–330.

Minnen, G., Carroll, J. & Pearce, D. 2000 Robust, applied morphological generation. In *Proc. 1st Int. Natural Language Generation Conf., Mitzpe Ramon, Israel, 12–16 June*, pp. 201–208. Stroudsburg, PA: Association for Computational Linguistics. See http://www.cs.bgu.ac.il/~nlg2000/.

Rauch, E., Bukatin, M. & Baker, K. 2003 A confidence-based framework for disambiguating geographic terms. In *Proc. HLT-NAACL Workshop on Analysis of Geographic References* (eds A. Kornai & B. Sundheim), pp. 50–54. Stroudsburg, PA: Association for Computational Linguistics. See http://www.aclweb.org/anthology/W03-0108.pdf.

Schilder, F., Versley, Y. & Habel, C. 2004 Extracting spatial information: grounding, classifying and linking spatial expressions. In *Proc. Workshop on Geographic Information Retrieval (GIR'04)* at *Proc. 27th Annu. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, Sheffield, UK.* See http://www.versley.de/schilder-geoir04.pdf

Smith, D. A. & Crane, G. 2001 Disambiguating geographic names in a historical digital library. In *Research and Advanced Technology for Digital Libraries, Proc. 5th European Conf., ECDL 2001, Darmstadt, Germany, 4–9 September* (eds P. Constantopoulos & I. Sølvberg). Lecture Notes in Computer Science, vol. 2163, pp. 127–136. Berlin, Germany: Springer. (doi:10.1007/3-540-44796-2_12)

Tobin, R., Grover, C., Byrne, K., Reid, J. & Walsh, J. 2010 Evaluation of georeferencing. In *Proc. 6th Workshop on Geographic Information Retrieval (GIR'10), Zurich, Switzerland*, pp. 1–8. New York, NY: ACM Press. (doi:10.1145/1722080.1722089)