Use of the Edinburgh Geoparser in the GeoDigRef and Embedding GeoCrossWalk Projects

Claire Grover, Richard Tobin
Kate Byrne, Matthew Woollard

November 2009

Table of contents

1. The Edinburgh Geoparser

1.1 Background

The GeoDigRef and Embedding GeoCrosswalk projects were both concerned with georeferencing digitised collections. In the case of GeoDigRef, the collections were Histpop, BOPCRIS and Archival Sound Recordings, while for Embedding GeoCrosswalk the collection was the Stormont Papers. In this report we describe the work that was undertaken to configure the geoparser for the collections as well as the evaluations that were performed.

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-TTT2 (http://www.ltg.ed.ac.uk/software/lt-ttt2) with geoparsing-specific sub-components which the LTG has developed in collaboration with EDINA, as part of the GeoCrossWalk project.

1.2 System Overview

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 (http://www.geonames.org) or GeoCrossWalk (http://www.geoxwalk.ac.uk)[1] 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.

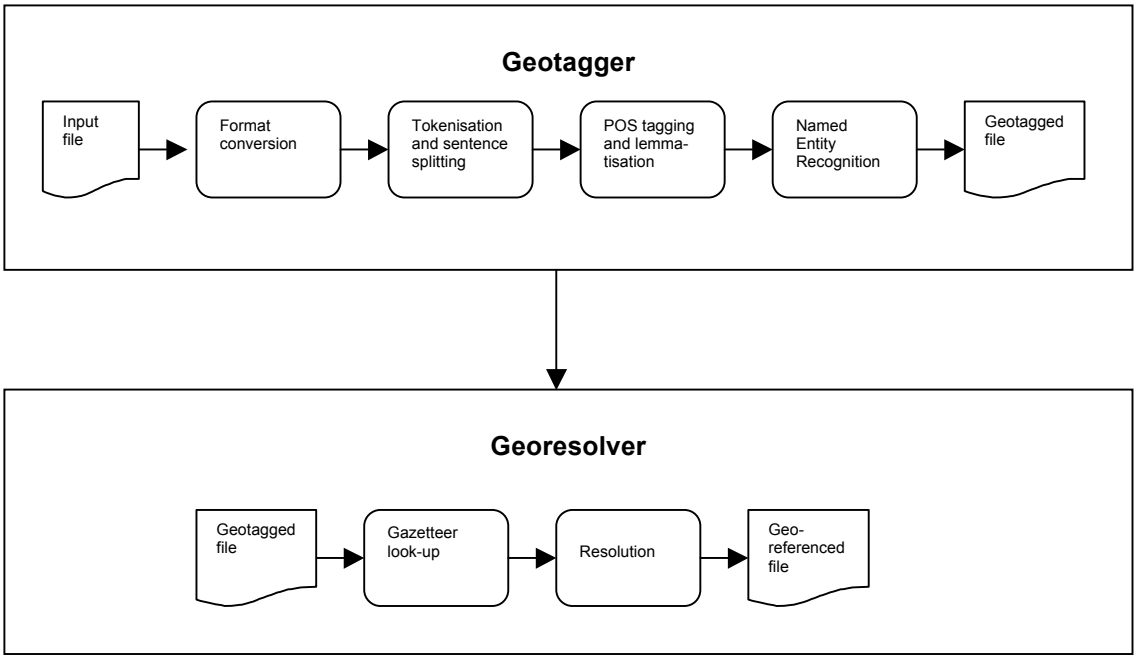

Figure 1. Overview of Geoparser architecture

[1] Recently the GeoCrossWalk project has been replaced by the Unlock service which is accessed at this URL or at the alternative http://unlock.edina.ac.uk/.

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.

1.3 Example

The workings of the system can be illustrated with a small example. The following is a small plain text input file, called example.txt:

> Some of the time savings will be remarkable: Canterbury will be an hour from London, a saving of 40 minutes; the journey from Dover will be slashed by 47 minutes and those living around Ebbsfleet near Gravesend will be just 18 minutes from St Pancras.

The output of the geotagger is an XML file, *example.geotagged.xml*, containing linguistic mark-up, including *enamex* elements which wrap recognised place names. Suppressing all other mark-up, the output looks like this:

> Some of the time savings will be remarkable:
> <enamex type='location' id='1'>Canterbury</enamex> will be an hour from
> <enamex type='location' id='2'>London</enamex>, a saving of 40 minutes; the
> journey from <enamex type='location' id='3'>Dover</enamex> will be slashed by
> 47 minutes and those living around
> <enamex type='location' id='4'>Ebbsfleet</enamex> near
> <enamex type='location' id='5'>Gravesend</enamex> will be just 18 minutes from
> <enamex type='location' id='6'>St Pancras</enamex>.

The *enamex* elements are extracted and converted into gazetteer queries, the format of which is dependent on the gazetteer being used. The gazetteer server responds with a list of entries which are converted to an appropriate XML format and saved in the file *example.gazunres.xml*. If the GeoCrossWalk gazetteer is used, the first part of this file looks like this:

```
<placenames>
 <placename id="1" name="Canterbury">
  <place long="1.083751" lat="51.275371" type="other" gazref="xwalk:256539"
name="Canterbury"/>
  <place long="1.070476" lat="51.2728005" type="other" gazref="xwalk:277683"
name="Canterbury"/>
  <place long="1.097974" lat="51.279467" type="other" gazref="xwalk:278734"
name="Canterbury"/>
  <place long="-2.728292" lat="57.62284" type="other" gazref="xwalk:43669"
name="Canterbury"/>
  <place long="1.0817925" lat="51.279124" ctv="city" type="ppl" gazref="xwalk:43670"
name="Canterbury"/>
 </placename>
 <placename id="2" name="London">
```

```
 <place type="ppl" name="London" gazref="geonames:6058560" in-cc="CA"
lat="42.9833893" long="-81.2330424"/>
 <place long="-0.108457" lat="51.5103495" ctv="city" type="ppl" gazref="xwalk:147216"
name="London"/></placename>
 <placename id="3" name="Dover">
  <place long="1.317512" lat="51.124058" type="other" gazref="xwalk:256262"
name="Dover"/>
  <place long="1.276323" lat="51.178629" type="other" gazref="xwalk:277732"
name="Dover"/>
  <place long="1.2787685" lat="51.2123225" type="other" gazref="xwalk:278736"
name="Dover"/>
  <place long="-2.580522" lat="53.508131" ctv="village" type="ppl" gazref="xwalk:74800"
name="Dover"/>
  <place long="1.2994465" lat="51.139001" ctv="towns" type="ppl" gazref="xwalk:74801"
name="Dover"/>
 </placename>
 <placename id="4" name="Ebbsfleet" near="5">
  <place long="1.351725" lat="51.318616" type="other" gazref="xwalk:81153"
name="Ebbsfleet"/> </placename>
```

The results of gazetteer look-up for any one place will contain zero, one or more than one entry. In the case of zero, if the gazetteer has no information, the place will have to remain unresolved. In the case of one entry (e.g. for *Ebbsfleet*), the georesolver takes this to be the correct resolution. In the case of multiple entries (as for *Canterbury*, *London, Dover* etc.), the georesolver ranks the entries in order of likelihood that they are correct in this context. To do this it uses the entries for all the other places in the document to provide context. The results of georesolution are saved in the file *example.gaz.xml*. the first part of which looks like this:

```
<placenames>
 <placename id="1" name="Canterbury">
  <place rank="1" score="1.775078808" scaled_contained_by="0" scaled_contains="0"
scaled_near="0" in-cc="GB" pop="46978" long="1.0817925" lat="51.279124" ctv="city"
type="ppl" gazref="xwalk:43670" name="Canterbury" clusteriness="52.31186385"
scaled_clusteriness="0.640699903" clusteriness_rank="2" scaled_pop="0.5343789047"
scaled_type="0.6"/>
  <place rank="2" score="1.175270453" scaled_contained_by="0" scaled_contains="0"
scaled_near="0" in-cc="GB" pop="46978" long="1.070476" lat="51.2728005" type="other"
gazref="xwalk:277683" name="Canterbury" clusteriness="52.26571584"
scaled_clusteriness="0.6408915485" clusteriness_rank="1" scaled_pop="0.5343789047"
scaled_type="0"/>
  <place rank="3" score="1.174713859" scaled_contained_by="0" scaled_contains="0"
scaled_near="0" in-cc="GB" pop="46978" long="1.083751" lat="51.275371" type="other"
gazref="xwalk:256539" name="Canterbury" clusteriness="52.39985572"
scaled_clusteriness="0.6403349544" clusteriness_rank="3" scaled_pop="0.5343789047"
scaled_type="0"/>
  <place rank="4" score="1.174162787" scaled_contained_by="0" scaled_contains="0"
scaled_near="0" in-cc="GB" pop="46978" long="1.097974" lat="51.279467" type="other"
gazref="xwalk:278734" name="Canterbury" clusteriness="52.53300402"
scaled_clusteriness="0.639783882" clusteriness_rank="4" scaled_pop="0.5343789047"
scaled_type="0"/>
  <place rank="5" score="0.09696540236" scaled_contained_by="0" scaled_contains="0"
scaled_near="0" long="-2.728292" lat="57.62284" type="other" gazref="xwalk:43669"
name="Canterbury" clusteriness="639.8367713" scaled_clusteriness="0.09696540236"
clusteriness_rank="5" scaled_pop="0" scaled_type="0"/>
 </placename>
```

The entry which has the highest score is ranked number 1. In the case of *Canterbury*, this is an entry which has been identified as a city and a populated place (in contrast to the other entries which are all of type other).

The optional *gazmap* component allows the results of the georesolver to be explored in a browser using Google Maps. Figure 2 shows a snapshot of the display for the current example.

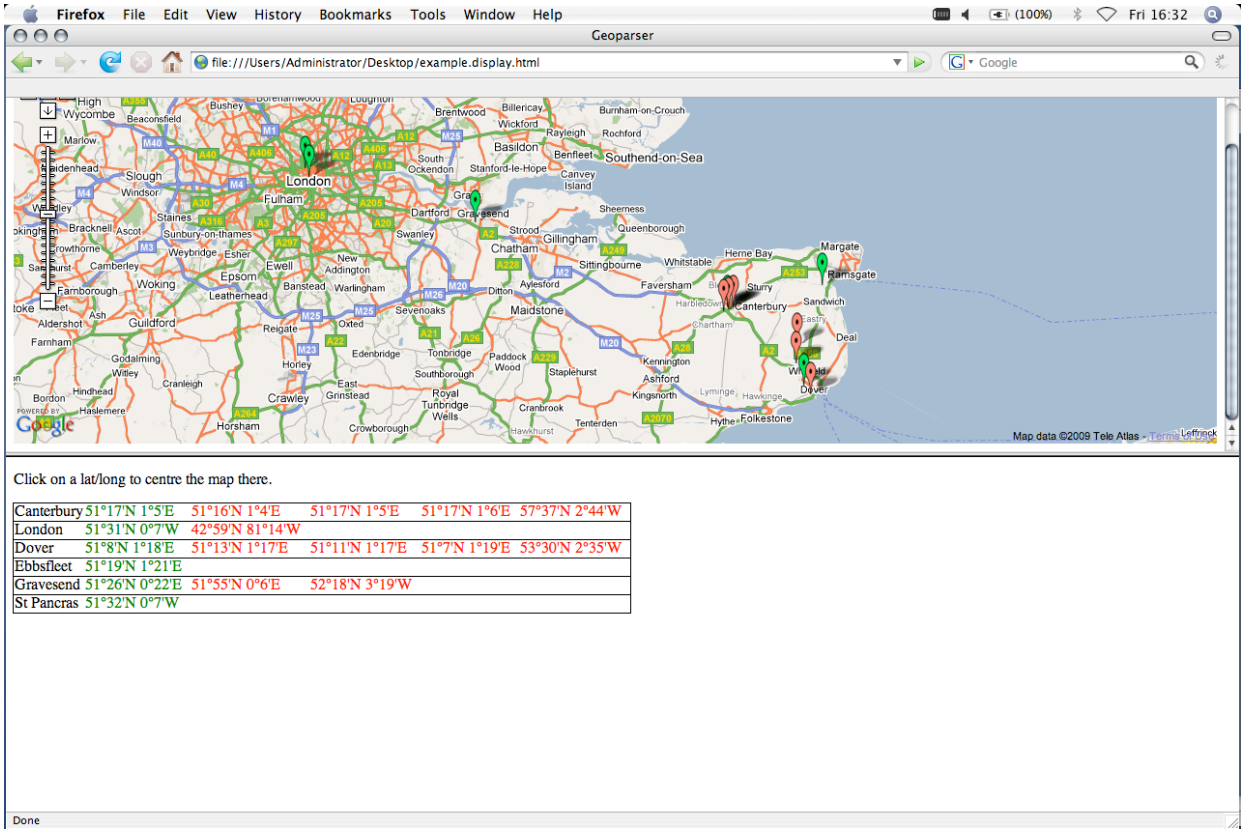

Figure 2. Display of results using GeoCrossWalk

The green markers and the corresponding green lat/long entries in the table below the map represent the highest ranked entries (for *Canterbury* the green marker is obscured by red markers corresponding to the less highly ranked entries).

The GeoCrossWalk gazetteer is derived from Ordnance Survey sources and only covers Great Britain. The Geonames gazetteer, on the other hand, covers the entire world. When the Geonames gazetteer is used, the entries returned are places with the relevant names from all over the world. Figure 3 shows a display of Geonames results for the current example. The majority of possible places are found in the U.S., however, the resolution process correctly ranks the appropriate UK entries first.

Both gazetteers find only one possible entry for each of *Ebbsfleet* and *St. Pancras.* The *St. Pancras* entry is correct but the *Ebbsfleet* one is not: the correct resolution is a very small place called Ebbsfleet close to Gravesend but the gazetteers both only have an entry for a larger Ebbsfleet further to the east near Ramsgate. This illustrates one fundamental issue for geo-referencing, namely that performance can never be better than the gazetteer permits.

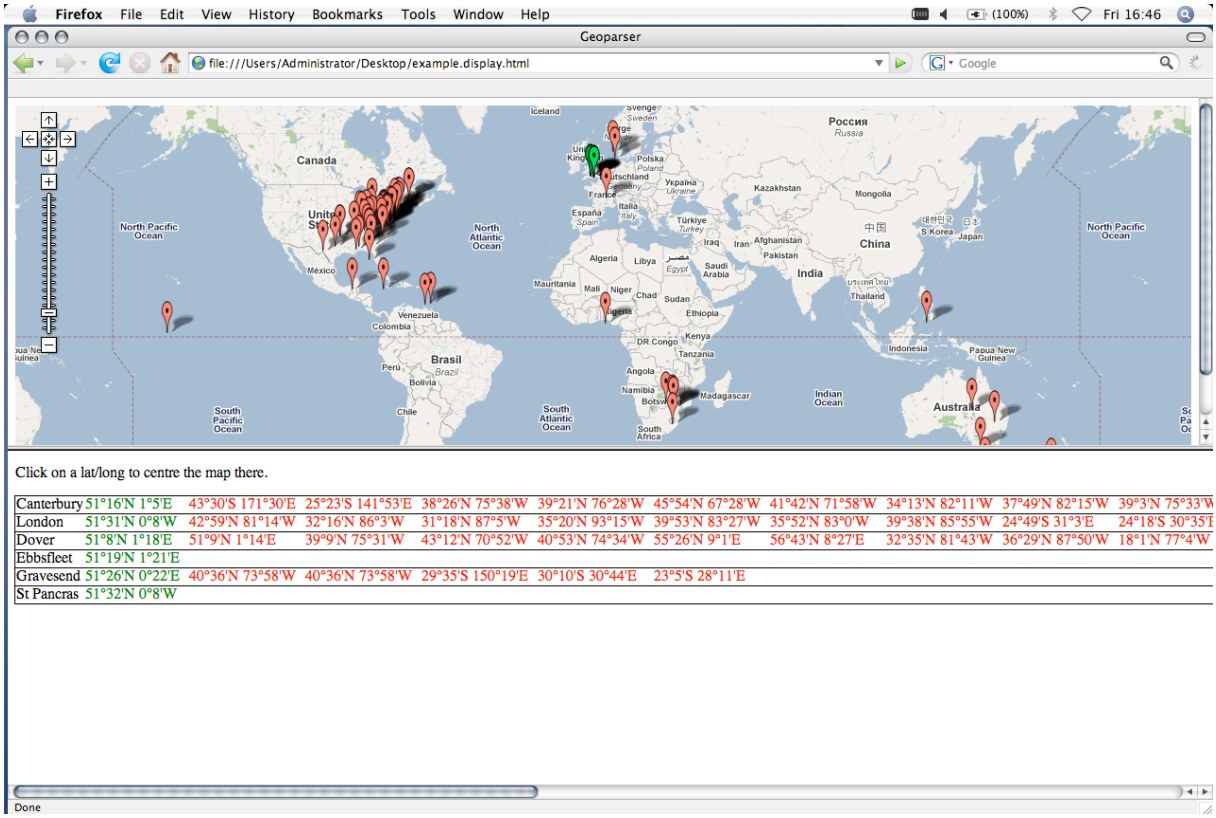

Figure 3. Display of results using GeoNames

2. Geoparsing for GeoDigRef and Embedding GeoCrossWalk

The two projects, GeoDigRef and Embedding GeoCrossWalk, used the geoparser to geo-reference four digitised collections:

- Histpop (History Data Service, HDS) - The Online Historical Population Reports (http://www.histpop.org)
- BOPCRIS 18th Century Parliamentary Publications – (www.parl18c.soton.ac.uk)
- Archival Sound Recordings (British Library, BL) - http://sounds.bl.uk/
- The Stormont Papers (Arts and Humanities Data Service, AHDS) - http://stormontpapers.ahds.ac.uk)

The geoparser as described in the previous section was capable of processing all of the collections once initial format conversion was taken care of. However, each collection has its own peculiarities and optimal performance could only be achieved through a process of adaption and configuration. In this section we describe the changes that have been made to the geoparser specifically for the four collections.

2.1 Input formats

The data from Histpop, BOPCRIS and the Stormont Papers are the output of OCR on the original documents. The data from BL are certain metadata fields from the sound archive entries.

2.1.1 Histpop

The Histpop data comprises 25,298 XML files totalling approx 10.5 million words. Each file corresponds to an individual page of the collection. The following is an extract from one of the files.

```
<?xml version="1.0" encoding="UTF-8"?>
<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 in ten cases the populations number less
than 5,000. The largest percentage increase is that recorded for Newport U.D. (Shropshire), viz.
12.5 per cent, (comprising 383 persons) and the largest numerical increase is that for
Shrewsbury M.B., viz. 1,366 persons (4.4 per cent.). ……
</ocr_text>
<fulltitle>
Census of England and Wales, 1931, Counties of Herefordshire and Shropshire (Part I)
</fulltitle>
</pages>
```

2.1.2 Bopcris

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 GeoDigRef, each volume was split into one page per file giving a total of 9,417 pages/files containing approx 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.)

<?xml version="1.0" encoding="UTF-8"?>
<page number="35">
<Page>35</Page>
<Word x="382" y="492" w="217" h="114">DIE</Word> <Word x="625" y="495" w="162" h="51">Lunae,</Word> <Word x="829" y="484" w="78" h="56">16</Word> <Word x="938" y="482" w="266" h="53">Februarii.</Word> <Word x="418" y="673" w="189" h="53">Domini</Word> <Word x="646" y="685" w="89" h="36">tam</Word> <Word x="761" y="669" w="271" h="64">Spirituales</Word> <Word x="1056" y="686" w="133" h="44">quam</Word> <Word x="1222" y="667" w="299" h="65">Temporales</Word> <Word x="1550" y="664" w="248" h="64">praesentes</Word> <Word x="1001" y="731" w="216" h="54">fuerunt.</Word> <Word x="1207" y="754" w="10" h="10">.</Word> <Word x="95" y="926" w="139" h="78">Epus.</Word> <Word x="258" y="930" w="188" h="65" font="it">Wigorn.</Word> <Word x="96" y="1012" w="138" h="58">Epus.</Word> <Word x="259" y="999" w="170" h="53" font="it">Ellen.</Word> <Word x="95" y="1077" w="140" h="58">Epus.</Word> <Word x="258" y="1060" w="199" h="53" font="it">Lich. &</Word><Word x="480" y="1058" w="103" h="51" font="it">Cov..</Word> <Word x="95" y="1144" w="121" h="57">Epus</Word> <Word x="258" y="1128" w="172" h="50" font="it">Petrib.</Word> ……
</page>

The following is a more readable excerpt from the same text where the *Word* element mark-up has been removed:

<page number="35">
<Page>35</Page>
DIE Lunae, 16Âº Februarii. Domini tam Spirituales quam Temporales praesentes fuerunt. . Epus. Wigorn. Epus. Ellen. Epus. Lich. &Cov.. Epus Petrib. Epus. Gloucestr. Epus. Oxon Epus. or. Ds; Custos Magni Sigilli Dux Somerset, P. Dux Devon, Senescallus. Dux Richmond. Dux Ormonde. Din St. Albans. : Dux Bolton. March. Normanby. Comes Lindsey, Magnus Camerarius. Comes Carlisle, Marescallus. Comes jersey, Camerarius. Comes Huntingdon. Comes Leicester. Copies Northampton. Comes Manchester. Comes Peterborow. Comes Stanford. Comes Kingston, 2. Comes Winchilsea. 1. Comest Thanet. Comes Scarsdale. Comes Shaftesbury. Comes Feversham. Comes Radnor. Comes Berkeley. Comes Nottingham. Comes Plimouth. Comes Portland. Comes Marlborough. Comes Torrington Comes Warrington. Comes Bradford. Comes Romney. Viscount Townshend. Viscount Weymouth. Viscount Longueville. Ds. Bergevenny; Ds. Lawarr. Ds. Ferrers. Ds. Wharton. Ds. North & Grey. Ds. Grey W. Ds. Howard Esc. Ds. Mohun. Ds. Leigh. Ds. Culpeper. Ds. Lucas. Ds. Rockingham. Ds. Cornwallis. Ds. Craven. : Ds. Ossulstone. Ds. Stawel. DS. Guilford. Ds. Jeffreys. Ds. Ashburnham. Ds. Lempster. . Ds. Herbert. Ds. Sommers. Ds. Halifax. PRAYERS. Hodie 2a vice lecta est Billa, intituled, "An to enable Trustees to sell certain Lands, Tithes, and Tenements, for the Payment of the Debts of Francis Purefoy Esquire, deceased." ORDERED, That the Consideration of the said Bill be committed to the Lords following; videlicet,) Dux Somerset, Praeses. Dux Bolton. Comes Leicester. Comes Northampton. Comes Denbigh. Comes Stamford. Comes Winchilsea. Comes Kingston Comes Essex. Comes Shaftesbury. I Epus. Wigorn. Epus. Sarum. Epus. Ellen. Epus. Lich. & Cov Epus. Petrib. Epus. Gloucestr. Epus. Oxon. Epus. Bangor. Ds. Bergevenny. Ds. Ferrers. Ds. North &Grey. Ds. Grey W. Ds. Howard Esca Ds. Mohun. Ds. Leigh. Ds. Culpeper. Ds. Rockingham. Ds Craven. Ds. Ossulstone Comes Radnor Comes Nottingham; Comes Plimouth. Comes Torrington; Comes Warrington; Comes Bradford. Viscount Townshend; Viscount Weymouth. Viscount Longueville: Ds. Dartmouth. Ds. Stawel. Ds. Guilford; Ds. Jeffreys. Ds. Ashburnham; Ds. Herberth Their Lordships, or any Five of them ; to meet on Tuesday the Third Day of March next; at Ten of the Clock in the Forenoon, in the Prince's Lodgings near the House of Peers ; and to adjourn as they please. This Day Charles Earl of Carlisle took the Oaths, and made and subscribed the Declaration, pursuant to the Statute ; and signed The Association. Upon reading the Petition of Robert Thurston; shewing, That he hath put in his Answer to the Appeal of jonathan Vaughan and his Wife ; and that the Appellants have not entered into a Recognizance to answer Costs, as is usual ; and that the Appeal is only for Delay; and praying a Day of Hearing may be appointed; and that the Appellant may, in the mean Time, enter into a Recognizance; or that. his Appeal be dismissed :" It. is ORDERED, by the Lords Spiritual and

rem in Parliament assembled, That this House will hear the said Cause, by Counsel, at the Bar, on Monday thc Three and Twentieth Day of this Instant February, at Eleven a Clock in the Forenoon ; and that, in mean Time, the Appellant do enter into a Recognizance for Costs, as is as. ……
</Page>

2.1.3 The Stormont Papers

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 3,315 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>Papers Presented to Parliament. Ministry of Agriculture:</p>
    <p>The Marketing of Dairy Produce Amendment No. 2 Rules 1932. (By Act.) Ministry of
        Education:</p>
    <p>Annual Report of the Ministry of Education for the year 1931–32. (By Act.)</p>
    <p>Ministry of Labour:</p>
    <p>The National Health Insurance (Approved Societies) Amendment Regulations 1932.
        (By Act.)</p>
    <p>Royal Assent.</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 by both Houses:–
    </p>
    <p>New Industries (Development), Housing (Grants), Railways (Valuation for Rating),
        Poultry Diseases, Loans Guarantee, Companies, Exported Animals (Compensation),
        Slaughter of Animals, </p>
……..
</day>
```

2.1.4 Archival Sound Recordings

The data in the British Library's Archival Sound Recordings collection are audio files. In this case the Geoparser is used to geo-reference certain of the metadata fields for the items in the collection. It is currently set up to process the contents of four fields, *dc:title*, *dc:description*, *dcterms:spatial* and *dcterms:abstract.* The metadata for an ASR record can be viewed when accessing that record on the ASR website (http://sounds.bl.uk). The following is an extract from the metadata associated with one of the records from the accents and dialects part of the collection:

```
<mets:xmlData>
<dc:title>Gwinear, Cornwall</dc:title>
<dc:description>
John explains a local expression, mentions his work in the mines, talks about his enjoyment of
cricket and reflects with pride on his continued robust health. Hayle (just to the west of
Gwinear) and Camborne (to the northeast) are nearby towns.
</dc:description>
<dc:identifier>sounds.bl.uk/021M-C0908X0027XX-0200V0</dc:identifier>
<dc:source>C908/27</dc:source>
<dcterms:created>1963/03/28</dcterms:created>
```

<dc:rights>http://sounds.bl.uk/JISC ASR IPR STATUS LIST.xls</dc:rights>
<dcterms:contributor>
<marcrel:PRO>University of Leeds</marcrel:PRO>
<marcrel:SPK>
Goldsworthy, John (b.1882; male, retired farm worker and tin miner)
</marcrel:SPK>
<marcrel:RCE>Ellis, Stanley (b.1926; male, SED fieldworker)</marcrel:RCE>
</dcterms:contributor>
<dcterms:spatial>Gwinear, Cornwall: OS Grid Reference(159500,37500)</dcterms:spatial>
<dcterms:temporal> </dcterms:temporal>
<dc:language>English</dc:language>
<dc:subject></dc:subject>
<rdf:about xlink:href="http://cadensa.bl.uk/"> </rdf:about>
<dc:type>sound</dc:type>
</mets:xmlData>

2.2 Configuring the Geotagger

In order to process texts from the collections it was necessary to make a range of adjustments and additions to the components that make up the geotagger (see Figure 1).

2.2.1 Format Conversion

The data from the collections is provided as XML but each conforms to a very different schema. The geoparser is able to handle relatively uncomplicated XML input files in a generic way but it was necessary to add collection-specific format conversion for the projects. This was implemented by allowing four more possible values for the *–t* parameter that the top-level script *geoparse* requires, these values being *histop*, *bopcris, stormont* and *bl*. When the geoparser is run with *–t histpop* the input file is passed to a tokeniser component that has been specialised to accept files in the Histpop format. Similarly, there are specialised tokeniser components for BOPCRIS, Stormont Papers and ASR files. Each of these identifies which parts of the XML files are to be processed.

2.2.2 Tokenisation

Tokenisation is the process of identifying word tokens as well as segmenting the text into sentences. This process is the same as in the original geoparser pipeline for Histpop, Stormont and ASR and results in *w* elements being wrapped around word tokens (i.e. a word or a punctuation mark) and *s* elements being wrapped around sentences:

<s><w>viii</w> <w>Shrewsbury</w> <w>M.B.</w> <w>and</w> <w>Hereford</w>
<w>M.B.</w> <w>are</w> <w>the</w> <w>most</w> <w>populous</w> <w>areas</w>
<w>with</w> <w>populations</w> <w>of</w> <w>32,372</w> <w>and</w>
<w>24,163</w> <w>respectively</w><w>.</w></s> <s><w>There</w> <w>are</w>
<w>two</w> <w>other</w> <w>urban</w> <w>areas</w> <w>with</w>
<w>populations</w> <w>over</w> <w>10,000</w><w>,</w> <w>seven</w> <w>with</w>
<w>populations</w> <w>between</w> <w>10,000</w> <w>and</w> <w>5,000</w>
<w>and</w> <w>in</w> <w>ten</w> <w>cases</w> <w>the</w> <w>populations</w>
<w>number</w> <w>less</w> <w>than</w> <w>5,000</w><w>.</w></s> <s><w>The</w>
<w>largest</w> <w>percentage</w> <w>increase</w> <w>is</w> <w>that</w>
<w>recorded</w> <w>for</w> <w>Newport</w> <w>U.D.</w>
<w>(</w><w>Shropshire</w><w>)</w><w>,</w> <w>viz</w><w>.</w></s>
<s><w>12.5</w> <w>per</w> <w>cent</w><w>,</w> <w>(</w><w>comprising</w>
<w>383</w> <w>persons</w><w>)</w> <w>and</w> <w>the</w> <w>largest</w>
<w>numerical</w> <w>increase</w> <w>is</w> <w>that</w> <w>for</w>
<w>Shrewsbury</w> <w>M.B.</w><w>,</w> <w>viz</w><w>.</w></s>
<s><w>1,366</w> <w>persons</w> <w>(</w><w>4.4</w> <w>per</w>
<w>cent</w><w>.</w><w>)</w><w>.</w></s>

<s><w>The</w> <w>MINISTER</w> <w>OF</w> <w>FINANCE</w>
<w>(</w><w>Mr.</w> <w>Pollock</w><w>)</w> <w>(</w><w>at</w> <w>the</w>
<w>Bar</w><w>)</w><w>,</w> <w>reported</w> <w>That</w> <w>His</w>
<w>Grace</w> <w>the</w> <w>Governor</w> <w>of</w> <w>Northern</w>
<w>Ireland</w><w>,</w> <w>in</w> <w>the</w> <w>name</w> <w>of</w> <w>and</w>
<w>on</w> <w>behalf</w> <w>of</w> <w>His</w> <w>Majesty</w> <w>the</w>
<w>King</w><w>,</w> <w>has</w> <w>been</w> <w>pleased</w> <w>to</w>
<w>give</w> <w>his</w> <w>Assent</w> <w>to</w> <w>the</w> <w>following</w>
<w>Bills</w> <w>agreed</w> <w>upon</w> <w>by</w> <w>both</w>
<w>Houses</w><w>:</w><w>–</w></s>

<dc:description>
<s><w>John</w> <w>explains</w> <w>a</w> <w>local</w>
<w>expression</w><w>,</w> <w>mentions</w> <w>his</w> <w>work</w> <w>in</w>
<w>the</w> <w>mines</w><w>,</w> <w>talks</w> <w>about</w> <w>his</w>
<w>enjoyment</w> <w>of</w> <w>cricket</w> <w>and</w> <w>reflects</w>
<w>with</w> <w>pride</w> <w>on</w> <w>his</w> <w>continued</w> <w>robust</w>
<w>health</w><w>.</w></s> <s><w>Hayle</w> <w>(</w><w>just</w> <w>to</w>
<w>the</w> <w>west</w> <w>of</w> <w>Gwinear</w><w>)</w> <w>and</w>
<w>Camborne</w> <w>(</w><w>to</w> <w>the</w> <w>northeast</w><w>)</w>
<w>are</w> <w>nearby</w> <w>towns</w><w>.</w></s>
</dc:description>

For BOPCRIS the specialisation of the tokeniser is more complex because the input file already contains XML mark-up around words. The transformation of the input involves retokenising because the token splitting is not what the pipeline expects: punctuation characters and following white space are included inside *Word* elements (see example above). The retokenisation results in tokens which are like those created for the other collections:

<page>
<s><w>DIE</w> <w>Lunae</w><w>,</w> <w>16</w> <w>Februarii</w><w>.</w></s>
<s><w>Domini</w> <w>tam</w> <w>Spirituales</w> <w>quam</w> <w>Temporales</w>
<w>praesentes</w> <w>fuerunt</w><w>.</w></s> <s><w>.</w></s>
<s><w>Epus</w><w>.</w> <w>Wigorn</w><w>.</w></s> <s><w>Epus</w><w>.</w>
<w>Ellen</w><w>.</w></s> …….</page>

Attributes on *w* and *s* elements have been suppressed in the previous examples for clarity. In the case of BOPCRIS, the coordinate attributes from the original *Word* elements are preserved along with a *bwid* attribute (bopcris word id) in order that the geoparser output can be mapped back to the input file and thus to the image:

<page number="35"><Page>35</Page> <s><w h="114" w="217" y="492" x="382"
bwid="bw1">DIE</w> <w h="51" w="162" y="495" x="625" bwid="bw2">Lunae</w><w
h="51" w="162" y="495" x="625" bwid="bw2">,</w> <w h="56" w="78" y="484" x="829"
bwid="bw3">16</w> <w h="53" w="266" y="482" x="938" bwid="bw4">Februarii</w><w
h="53" w="266" y="482" x="938" bwid="bw4">.</w></s> <s><w h="53" w="189" y="673"
x="418" bwid="bw5">Domini</w> <w h="36" w="89" y="685" x="646"
bwid="bw6">tam</w> <w h="64" w="271" y="669" x="761" bwid="bw7">Spirituales</w>
<w h="44" w="133" y="686" x="1056" bwid="bw8">quam</w> <w h="65" w="299" y="667"
x="1222" bwid="bw9">Temporales</w> <w h="64" w="248" y="664" x="1550"
bwid="bw10">praesentes</w> <w h="54" w="216" y="731" x="1001"
bwid="bw11">fuerunt</w><w h="54" w="216" y="731" x="1001" bwid="bw11">.</w></s>
<s><w h="10" w="10" y="754" x="1207" bwid="bw12">.</w></s> <s><w h="78" w="139"
y="926" x="95" bwid="bw13">Epus</w><w h="78" w="139" y="926" x="95"
bwid="bw13">.</w> <w font="it" h="65" w="188" y="930" x="258"
bwid="bw14">Wigorn</w><w font="it" h="65" w="188" y="930" x="258"
bwid="bw14">.</w></s> <s><w h="58" w="138" y="1012" x="96"
bwid="bw15">Epus</w><w h="58" w="138" y="1012" x="96" bwid="bw15">.</w> <w

font="it" h="53" w="170" y="999" x="259" bwid="bw16">Ellen</w><w font="it" h="53" w="170" y="999" x="259" bwid="bw16">.</w></s> ….
</page>

The tokeniser component also does sentence splitting. For Histpop, Stormont and ASR, this process is much the same as in the original geoparser except that 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 boundary full stops. For BOPCRIS, the usual sentence splitter was not used and a specialised one was implemented. The motivation for this was partly 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 separate sentences.

The BOPCRIS data contains frequent passages in Latin and occasional ones in French. These had a detrimental effect on later stages of processing and it was decided that it would be useful to identify the language of any given part of the text to prevent the named entity recognition from applying to Latin and French passages. To achieve this Van Noord’s language guesser, TextCat (www.let.rug.nl/vannoord/TextCat/) was applied on a per sentence basis, using the English, French and Latin language models. For each sentence, TextCat outputs one more or guesses where, in the case of multiple guesses, the most likely is ordered first. In the XML representation, TextCat’s output is encoded in the *lang* attribute on *s* elements so, for example, the first sentence below was categorised as Latin or French, and the third sentence as Latin or English or French:

<page number="35"><Page>35</Page> <s lang="lf"><w>DIE</w> <w>Lunae</w><w>,</w> <w>16</w> <w>Februarii</w><w>.</w></s> <s lang="l"><w>Domini</w> <w>tam</w> <w>Spirituales</w> <w>quam</w> <w>Temporales</w> <w>praesentes</w> <w>fuerunt</w><w>.</w></s> <s lang="fl"><w>.</w></s> <s lang="lef"><w>Epus</w><w>.</w> <w font="it">Wigorn</w><w font="it">.</w></s> <s lang="fle"><w>Epus</w><w>.</w> <w font="it">Ellen</w><w font="it">.</w></s>…</page>

When processing BOPCRIS documents, place names are not recognised in text which has not been classified as English (either as the first guess or as the only guess).

2.2.3 POS tagging and lemmatisation

The next stage of the geotagger pipeline determines the most likely part-of-speech (POS) for each word and lemmatises nouns and verbs. The POS tagger is the C&C tagger (Curran and Clark 2003) trained on the Penn Treebank (Marcus, Santorini, and Marcinkiewicz 1993). The lemmatiser is *morpha* (Minnen, Carroll, and Pearce 2000). Lemmatisation is the process of finding the stem form of inflected words. To illustrate, the following is the output for the two sentences, “Best is a town near Eindhoven. George Best was one of the best footballers”:

<s><w l="best" p="NNP">Best</w> <w l="be" p="VBZ">is</w> <w p="DT">a</w> <w l="town" p="NN">town</w> <w p="IN">near</w> <w l="eindhoven" p="NNP">Eindhoven</w><w p=".">.</w></s>
<s><w l="george" p="NNP">George</w> <w l="best" p="NNP">Best</w> <w l="be" p="VBD">was</w> <w p="CD">one</w> <w p="IN">of</w> <w p="DT">the</w> <w p="JJS">best</w> <w l="footballer" p="NNS">footballers</w><w p=".">.</w></s>

The *p* attribute contains the part-of-speech and the *l* attribute contains the lemma. Both instances of “Best” are categorised as NNP, which is the label for proper noun. The word “best” is categorised as JJS, which is the label for superlative adjective. The named entity recognition rules take part-of-speech into account and generally only consider proper nouns as potential entities. Lemmatisation affects only nouns and verbs. In this example, “is” and “was” have the lemma “be” while the plural “footballers” has the lemma “footballer”. Lemmatisation information is mostly used to regularise inflected forms when performing look-up in lexicons. Neither the POS tagger nor the lemmatiser was changed for the two projects,

though a maximum sentence length parameter for the POS tagger was significantly increased (from 250 to 1,600 words) to deal with long sentences from Histpop and BOPCRIS.

2.2.4 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 recogniser in LT-TTT2 though for the two projects it has been configured to recognise only person and place names, disregarding the LT-TTT2 rules for dates, numerical expressions and organisation names. The primary aim for the projects is to achieve highly accurate place name recognition, however it is easier to achieve this goal 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, Duke of Norfolk, Francis Chichester, George Best etc.). The person names recognised in the Histpop and ASR collections were not initially considered to be of value, especially since they are infrequent in Histpop. However, the BOPCRIS data contains many more person names than place names and BOPCRIS have indicated that users would be interested in a person search facility. For this reason, the geotagger outputs both and the interface developed by EDINA allows both place and person search. Similarly, person names in the Stormont Papers are relatively frequent and it was decided that it would be useful to compute them for the Embedding GeoCrossWalk project as well.

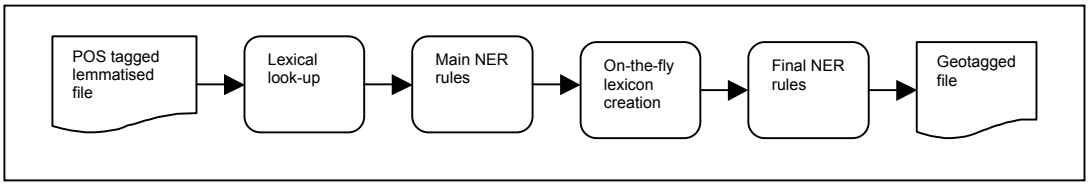

Figure 4. NER Component

The NER component is made up of a number of subcomponents, as shown in Figure 4. The first stage involves looking up words or sequences of words in a variety of lexicons, including a lexicon of common English words, a lexicon of male and female forenames, and two geographic lexicons. One geographic lexicon is derived from the name list of the Alexandria Digital Library Project Gazetteer (http://www.alexandria.ucsb.edu), a very extensive world-level gazetteer. The second 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 entries in the lexicons 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 capitalisation and occurring together, these words are unlikely to denote anything other than a place (except as part of a larger name such as “Shepherd’s Bush Empire”). By contrast, single words which can be found in a place name lexicon will frequently not denote a place in a particular context. For example “Kendal” is a place name and not a common noun 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 which are also place names as places (e.g. Best, Drum, Start etc.).

The output of the first lexical look-up stage identifies multi-word sequences which have matched successfully as location entities (<*enamex type=”location”*>) and adds attributes to all other words which have matched entries from one of the lexicons. Thus in the following, “Shepherd” and “Bush” are marked as *common=”true”* because they matched entries in the lexicon of common nouns. “John” and “Kendal” are marked as *pername=”true”* because they matched entries in the lexicon of forenames. “Shepherd’s Bush” is wrapped by an <*enamex type=”location”*> element because it was found in one of the place name lexicons. “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”* mark-up; where it was found in EDINA’s GeoCrossWalk-derived lexicon the words receive *edsource=”true”* mark-up. 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.

```
<s>
 <w pername="true" p="NNP" l="john">John</w>
 <w pername="true" p="NNP" l="kendal" locname="single" alsource="true"
    edsource="true">Kendal</w>
 <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>
```

A special look-up stage was implemented for the Stormont Papers for person names. The volumes of proceedings have indexes including a list of persons mentioned in the text. To ensure higher person name recognition accuracy, these lists were combined and converted to a lexicon which was used for Stormont-specific look-up. This proved particularly useful for segmenting complex lists of person names which occur whenever a vote is reported, for example:

> McConnell, Robert William Brian Maginess, The Rt. Hon. William Brian Minford, Nathaniel Owens Morgan, The Rt. Hon. William James Neill, The Rt. Hon. Ivan O'Neill, The Hon. Phelim Robert Hugh O'Neill, Capt. The Rt. Hon. Terence Marne Scott, Walter Warnock, The Rt. Hon. John Edmond West, The Rt. Hon Henry William

The lexical look-up stage concludes all the computation that is needed to provide information to the main NER rules in the next stage of processing. The NER rules and formalism are extensively documented in the LT-TTT2 documentation and will not be discussed here. As a result of the main NER rules, further *<enamex type="location">* elements as well as *<enamex type="person">* elements are marked up. For example, "John Kendal" in the current example is recognised as a person.

```
<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 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 look-up 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 Kensington", the final stage would find two entities:

```
<s>
 <enamex type="person" subtype=”otf”>
  <w pername="true" p="NNP" l="kendal" locname="single" alsource="true"
      edsource="true">Kendal</w>
 </enamex>
 <w p="RB">now</w>
 <w p="VBZ" l="live">lives</w>
 <w p="IN">in</w>
 <enamex type="location">
  <w p="NNP" l="kensington" locname="single" alsource="true"
      edsource="true">Kensington</w>
 </enamex>
 <w p=".".>.</w>
</s>
```

The word “Kendal” has been identified as a person entity because it was found in the on-the-fly lexicon (*otf=”true”*). “Kensington” has been marked as a location because even though it is a single word, it is not a common noun and the context doesn’t suggest that it might not denote a place. The on-the-fly lexicons 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. In the remainder of this section we describe some of the more substantial changes that needed to be made.

- In the annotated evaluation data for Histpop (see Section 3), words such as “County” were included in the mark-up of place names (e.g. “County of Norfolk”, “Tyrone Co.”, “County Down” etc.) and the NER rules were accordingly configured to do the same. However, entries in gazetteers are potentially shorter (“Norfolk”, “Tyrone”, “Down”) and gazetteer look-up in the georesolver would fail on the longer version. To compensate, the NER rules were extended so that the longer name is marked-up while the shorter name appears as the value of an *altname* 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)”).
- The BOPCRIS data required the addition of a number of new titles for persons (e.g. “Epus”, “Dux”, “Ds.”, etc.). Similarly the Histpop data prompted the addition of many new terms associated with place (e.g. “Borough”, “District”, “Soke”, “Ward”, “Diocese”, “M.B.”, “R.D.” etc.). The Stormont data contains regular patterns in which person names occur (e.g. “The MINISTER OF COMMERCE (Mr. Barbour):”) and rules were included to respond to these patterns.
- The data also prompted the inclusion of rules for names of saints as well as the tendency for possessive saint names to denote places (e.g. “St, Paul’s”).
- 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 *Slow Markett,* in the County of *Suffolke;*”). The output of 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 capitalised. The NER rules were therefore modified to take font information into account.

2.3 Georesolver

The output of the geotagger is an XML file (with extension *.geotagged.xml*). This file is input to the resolultion process which consists of two main stages, look-up of the place names in a gazetteer and resolution which ranks the resulting matches. For visualisation of the results there is a third stage which uses the Google Maps API to display the ranked locations.

2.3.1 Gazetteer look-up

This is implemented as a shell script named *geogaz*. First, the place names are extracted from the .geotagged.xml 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. In the example in the previous section, there were two place names, “Shepherd’s Bush” and “Kensington” so the file of extracted place names is this:

```
<placenames>
  <placename id="1" name="Shepherd's Bush"/>
  <placename id="2" name="Kensington"/>
</placenames>
```

The placenames are then passed to a gazetteer look-up script. These are named, for example, *gazlookup-xwalk*, *gazlookup-geonames*. The idea is that each gazetteer's script should do the look-up in whatever way is appropriate, but produce 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).
Since geonames queries are just URL fetches ("REST style"), the geonames script uses an XSLT stylesheet to generate and perform a query for each place name. This is simple but slow because it involves many transactions per document. For xwalk, on the other hand, a single query is generated listing all the place names, for example:

```
<?xml version="1.0" encoding="UTF-8"?>
<gazetteer-service xmlns="http://www.alexandria.ucsb.edu/gazetteer" version="1.1">
 <query-request>
  <gazetteer-query>
   <or>
    <name-query text="Shepherd's Bush" operator="equals"/>
    <name-query text="Kensington" operator="equals"/>
   </or>
  </gazetteer-query>
  <report-format>standard</report-format>
 </query-request>
</gazetteer-service>
```

This 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 <*placename*> element contains a number of <*place*> elements which are the candidate places from the gazetteer. These elements have attributes as follows:

- 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)
- type (feature type, see below)
- ctv (xwalk only, "city/town/village")

Each gazetteer has a large set of feature types. We reduce this to a small set for use by the disambiguation code:

- 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,
- other: other

Each gazetteer script is responsible for doing this mapping. After gazetteer look-up 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 look-up using xwalk for the current example is this:

```
<placenames>
 <placename id="1" name="Shepherd's Bush">
  <place long="-0.234769" lat="51.509433" ctv="village" type="ppl"
         gazref="xwalk:209396" name="Shepherd's Bush"/>
 </placename>
 <placename id="2" name="Kensington">
  <place long="-2.955419" lat="53.415832" ctv="village" type="ppl"
         gazref="xwalk:127825" name="Kensington"/>
  <place long="-0.189287" lat="51.500736" ctv="towns" type="ppl"
         gazref="xwalk:127826" name="Kensington"/>
  <place long="-0.20255" lat="51.5116515" type="other" gazref="xwalk:277827"
         name="Kensington"/>
  <place long="-2.937585" lat="53.4124255" type="civil" gazref="xwalk:283287"
         name="Kensington"/>
  <place long="0.058092" lat="51.542856" type="civil" gazref="xwalk:284660"
         name="Kensington"/>
 </placename>
</placenames>
```

From this it can be seen that the GeoCrossWalk gazetteer contains one entry for Shepherd's Bush and five for Kensington (three of which have grid references which place them in London, and two of which are in Liverpool).

The output from the geonames gazetteer look-up is this:

```
<placenames>
 <placename name="Shepherd's Bush" id="1"/>
 <placename name="Kensington" id="2">
  <place name="Kensington" gazref="geonames:2161608" type="ppl" lat="-33.9166667"
long="151.2166667" in-cc="AU"/>
  <place name="Kensington" gazref="geonames:2161609" type="civil" lat="-37.7833333"
long="144.9333333" in-cc="AU"/>
  <place name="Kensington" gazref="geonames:2645801" type="ppl" lat="51.5009423821754"
long="-0.191745758056641" in-cc="GB"/>
  <place name="Kensington" gazref="geonames:3489872" type="ppl" lat="18.3333333"
long="-77.3333333" in-cc="JM"/>
```

```
  <place name="Kensington" gazref="geonames:3489873" type="ppl" lat="18.1833333"
long="-76.5833333" in-cc="JM"/>
  <place name="Kensington" gazref="geonames:3489874" type="ppl" lat="17.9" long="-77.65"
in-cc="JM"/>
  <place name="Kensington" gazref="geonames:4203717" type="ppl" lat="33.4979141"
long="-82.1581753" in-cc="US"/>
  <place name="Kensington" gazref="geonames:4203718" type="ppl" lat="34.7742451"
long="-85.3727368" in-cc="US"/>
  <place name="Kensington" gazref="geonames:4273927" type="ppl" lat="39.7669559"
long="-99.0317506" in-cc="US"/>
  <place name="Kensington" gazref="geonames:4296911" type="ppl" lat="38.9036742"
long="-84.613277" in-cc="US"/>
  <place name="Kensington" gazref="geonames:4359759" type="ppl" lat="39.0831654"
long="-76.5799644" in-cc="US"/>
  <place name="Kensington" gazref="geonames:4359760" type="ppl" lat="39.0256651"
long="-77.0763669" in-cc="US"/>
  <place name="Kensington" gazref="geonames:4558950" type="ppl" lat="39.9859458"
long="-75.1318429" in-cc="US"/>
  <place name="Kensington" gazref="geonames:4837222" type="ppl" lat="41.6353769"
long="-72.7687083" in-cc="US"/>
  <place name="Kensington" gazref="geonames:5088311" type="ppl" lat="42.9270334"
long="-70.9439447" in-cc="US"/>
  <place name="Kensington" gazref="geonames:5123281" type="ppl" lat="40.7934345"
long="-73.7220756" in-cc="US"/>
  <place name="Kensington" gazref="geonames:5362849" type="ppl" lat="37.9104805"
long="-122.2802471" in-cc="US"/>
  <place name="Kensington" gazref="geonames:5991080" type="ppl" lat="46.433429917"
long="-63.648714891" in-cc="CA"/>
  <place name="Kensington" gazref="geonames:991368" type="ppl" lat="-31.25"
long="26.1833333" in-cc="ZA"/>
  <place name="Kensington" gazref="geonames:991370" type="rgn" lat="-28.8"
long="25.1833333" in-cc="ZA"/>
 </placename>
</placenames>
```

There are no entries for Shepherd's Bush and at least twenty for Kensington (the queries to the gazetteers place a maximum restriction on the number of possible responses, so only twenty are returned). Of the entries for Kensington, two are in Australia, one is in Great Britain, three are in Jamaica, one is in Canada, two are in South Africa and the remainder are in the U.S. The gazetteer look-up files are input to the resolution component described in Section 2.3.3.

2.3.2 Gazetteer look-up issues

The gazetteer look-up process appears to be performing adequately but there are a number of issues which should be explored further:

- The scripts would benefit from more tailoring to the different gazetteers: in particular details of the matching such as case-sensitivity. However, it's difficult to do a case-insensitive look-up if the gazetteer doesn't provide it; we would have to try all plausible case combinations and then eliminate duplicates. At present, because xwalk is case-sensitive, a look-up of "LONDON" will return no results.
- The xwalk gazetteer is confined to Great Britain. 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. This produces more issues; for example the case problem mentioned above results in "LONDON" being found as the town in Canada, but not as the capital of England.
- In creating the queries duplicate place names are reduced to a single representative. There is a problem with this, because we try to use some contextual information such

as container-contained relations. These relations are given by the "near", "contains", and "contained-by" attributes. It's possible that these attributes will be present on one or more of the instances of a place name. We just use the first occurrence of each attribute for a given place name. The same applies for the alternative names given by *altname* and *abbrev-for* attributes (NB no attempt is currently made to de-duplicate the alternative names).

2.3.3 Resolution (or disambiguation).

This is implemented as a shell script named *georesolve*. Before applying any of the heuristics described below, we try to augment the information about each candidate place. 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 it's the right one. The information added is:

- population
- containing country

A number of heuristics are used for resolution.

- 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: co-locations indicating containment and proximity ("London, England", "Leith near Edinburgh").
- A locality parameter from the user: *georesolve* can be called with an optional “-l” parameter which enables the user to specify the geographic focus of a document expressed as “-l lat long radius score” to give an additional “score” to places within “radius” of “lat”/”long”.
- Clustering: places in a document are often close together. Minimising the bounding box has problems: often a document mentions one or two remote places, so that the overall bounding box is most of the world, and it's hard to combine with other heuristics. 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 these heuristics is scaled to be in the range 0-1, using logarithmic scaling for the population and clustering. This scaling is not very principled and we don't have enough data to experiment properly. Some changes we ought to make are clear, e.g. facilities should be weighted even lower than they are.

The scaled values are combined to produce a single score for each candidate. Again the combination is unprincipled; it is specified by an XPath in *georesolve* and can be passed in as an argument. We can potentially change this formula in accordance with things we know about the text: perhaps weight population higher and clustering lower for news articles, for example.
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. In both the xwalk and the geonames look-up of our running example, the correct Kensington in London is ranked first.

3. Evaluation

3.1 Overview

When building a system such as the geoparser, it is important to be able to report the quality of its performance in concrete, quantifiable terms. In the field of Information Extraction it is usual practice to create 'gold standard' evaluation data against which system output can be compared. Since the data in the GeoDigRef and Embedding GeoCrossWalk projects is not standard data, it was necessary to create evaluation data specifically for the collections being processed. In the case of Histpop and Stormont, completely new evaluation data was created. For BOPCRIS there was a pre-existing evaluation set for named entity recognition which was created during the previous collaboration between LTG and BOPCRIS (see Appendix 1 for the original guidelines). This consisted of annotations on the output of an earlier OCR stage and was not directly reusable; however it was possible to semi-automatically transfer the previous annotations to the new OCR output. For ASR, there was no provision in the work plan for evaluation data to be generated, so none was created.

Since the geoparser is composed of two main stages, the geotagger and the georesolver, we created evaluation (test) data for both stages. The geotagger test sets were a selection of Histpop , BOPCRIS and Stormont texts manually annotated for person and location entities. The georesolver test sets were the same data where the human annotated location entities had been manually resolved twice, first using the xwalk gazetteer and then using the geonames gazetteer. The geotagging guidelines for Histpop and Stormont (see Appendix 2) were drawn up by Matthew Woollard, the georesolution guidelines for Stormont were generated by Richard Tobin and Vasilis Karaiskos (see Appendix 3) and were revised for Histpop by Matthew Woollard (see Appendi 4). Both stages of annotation of the Histpop data were completed by staff at Essex, while the BOPCRIS and Stormont collections were annotated by LTG staff in Edinburgh.

Table 1. shows information about the three geotagger test sets. The Histpop set is comprised of 500 documents, each corresponding to an OCRed page randomly selected from the complete Histpop collection. The BOPCRIS set contains 92 randomly selected pages from Volumes 14 and 50 of the Journals of the House of Lords. 46 of the documents are from Volume 50 and are updates of one of the test sets of previous work with BOPCRIS (Grover, Givon, Tobin and Ball, 2008); 46 are from Volume 14 and are updates of the second test set in the previous work. The previous work also included a devtest set but this was not updated. The Stormont set contains 12 randomly chosen documents but since each document represents a day of proceedings it can contain a lot of pages – the 12 documents contain a total of 471 pages. On average, a BOPCRIS page is twice the length of a Histpop page (1,118 tokens for BOPCRIS vs. 523 tokens for Histpop, where a token is a *w* element created by tokenisation – see previous section) and a Stormont document (15,459 tokens) is more than ten times the length of BOPCRIS page. Although there are differences in the number of documents per set, the difference in absolute corpus size not so large. Interestingly, 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 is comparatively sparsely populated with entities.

Collection	Documents	Sentences	Tokens	Entities
Histpop	500	9,329	261,676	6,400
BOPCRIS	92	5,486	102,851	5,990
Stormont	12	7,601	185,503	3,023

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

Tables 2-4 provide a breakdown of the entities in the two sets. The BOPCRIS data was originally annotated some years ago according to the guidelines attached as Appendix 1. At that time, only the broad categories of *person* and *location* were annotated. The previous annotations were transferred semi-automatically using the MMAX2 annotation tool (http://mmax2.sourceforge.net) to hand correct an automatic mapping. The Histpop and

Stormont test sets were created using the guidelines in Appendix 2 and again using the MMAX 2 tool. Here we took the opportunity to define a more fine-grained set of entities by sub-typing the basic *person* and *location* categories. Briefly, a *person* entity with subtype *name* is the name of a person while one with subtype *other* is some other use of person's name (e.g. "Hodgkins disease"). A *location* entity with subtype *geop* (geo-political) is the name of a populated place, one with subtype *feat* is a natural geographical feature (mountain, river etc) and one with subtype *other* is a very small place (like a street or building). A few instances are discontinuous (e.g. the strings "Sir Thomas" and "Dixon" in "Sir Thomas and Lady Dixon" make up one discontinuous entity while "Lady Dixon" is a standard entity). These discontinuous cases were marked up by the annotator (see the *-disc* entries in the tables below) but the mapping from the MMAX 2 format, which permits discontinuous entities, to the system format, which does not, is not straightforward and was done inconsistently and unsatisfactorily. However, the numbers are small and do not affect evaluation significantly.

From Tables 2-4 it can be seen that the three sets are quite different: Histpop entities are predominantly geo-political 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 of number of person names in BOPCRIS is the listing of all the lords present at the start of each day's proceedings.

Entity	Subtype	No. occurrences
person	*name*	245
	name-disc	2
	other	53
	subtotal	**300**
location	*geop*	5,780
	geop-disc	18
	other	230
	other-disc	7
	feat	92
	subtotal	**6,127**
Total		**6,427**

Table 2. Breakdown of entities in Histpop test set

Entity	Subtype	No. occurrences
person	*name*	1,590
	name-disc	8
	other	36
	subtotal	**1,634**
location	*geop*	1,169
	geop-disc	7
	other	188
	other-disc	17
	feat	40
	subtotal	**1,421**
Total		**3,055**

Table 3. Breakdown of entities in Stormont test set

Entity	No. occurrences
person	4,809
location	1,181
Total	**5,990**

Table 4. Breakdown of entities in BOPCRIS test set

3.2 Geotagger evaluation

To assess the performance of the geotagger, the system is run over the documents in the test sets and its output is compared to 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 since a very accurate but conservative system might obtain near perfect precision but very low recall while a system that is over-enthusiastic in making predictions might achieve very high recall but at the expense of precision. F-score is the harmonic mean of precision and recall (F=2 x (precision x recall)/(precision+recall)) and gives an idea of overall system performance.

The gold standard Histpop and Stormont test sets contain finer-grained entities if we take subtypes into account. However, the system itself does not currently distinguish between locations which are geo-political and ones which are features. For this reason we evaluate on the broader categories. Moreover, the system is not designed to recognise *subtype="other"* entities so we exclude these from this evaluation as well as from the georesolver evaluation reported in the next section. Tables 5-7 shows results for the Histpop, BOPCRIS and Stormont test sets respectively.

	Precision	**Recall**	**F-score**
location	82.09%	80.78%	81.43
person	53.07%	59.51%	56.11
Total	80.77%	79.93%	80.34

Table 5. Results for Histpop test set

	Precision	**Recall**	**F-score**
location	71.72%	74.67%	73.17
person	85.71%	88.55%	87.10
Total	79.64%	82.55%	81.07

Table 6. Results for Stormont test set

	Precision	**Recall**	**F-score**
location	55.92%	61.56%	58.61
person	81.83%	82.57%	82.20
Total	76.35%	78.43%	77.38

Table 7. Results for BOPCRIS test set

It can be seen from Tables 5-7 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. There are a number of factors which may contribute to the pattern of results:

1. The results may simply be affected by the frequencies of entities in the original texts;
2. The results may be affected negatively by mismatches between the extent of gold and system entities (and overlap in terms of places/people);
3. The results across samples may be affected by differences in use/interpretation of person/location entities.
4. The results are affected by incorrect gold standard mark up, i.e. annotation errors;

The first and second of these issues are related to the text, in the first case better results are being found when there are some very high frequencies of entities across the samples. The second issue is interesting and it may be that the BOPCRIS texts, especially the earlier ones, have the highest proportions of entities (or part entities) which can be identified as both

places and persons. This may *seem* to lead to a higher recall for people than places. The third and fourth are to do with the application of the gold standard.

To explore the first possible factor, 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 8 shows the top fifteen most frequent entities in each corpus with their counts. From this it can be seen that the fifteen most frequent entity types in the Stormont corpus account for approx 32% of the tokens, while for Histpop the proportion is 27% and, at the other extreme, for BOPCRIS it is 9%.

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%)**

Table 8. Entity frequencies in the corpora

The relative distributions of *location*/*person* entities in the top fifteen reflect the distributions shown in Tables 5-7: BOPCRIS has more *person* entities, Histpop has more *location* entities and Stormont is roughly balanced. The rarity of *person* entities in Histpop is reflected by the fact that the most frequent person name (*Dr. Price*) occurs only eight times.

The second factor concerns overlap and variability in the extent of entities. If a long gold standard entity is compared to a system entity which covers some of the same string but is shorter, then the system entity is penalised even if it is a plausible entity. For example, in one BOPCRIS document "Ann Dowager Baroness Southampton" is the gold standard entity while the system identified 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.

The third factor concerns systematic differences between the annotation and the way the system has been designed to behave. For example, in the Stormont data certain place names within larger names seem to have been systematically annotated, e.g. “Ulster” in “Royal Ulster Constabulary” even though the guidelines ruled against this kind of mark-up. The system has been designed not to recognise such examples and is therefore wrongly penalised in these cases. It would be interesting is to try and apply the rules for gold standard mark-up absolutely consistently across the files. The fact that slightly different rules have been applied probably reflects the understanding of the source material.

The fourth factor concerns the quality of gold standard mark up. Human error is unavoidable and there will be some gold entities which are wrongly marked-up. Unfortunately resources did not permit us to follow the common practice of having a sub-sample annotated by two independent annotators in order to calculate inter-annotator agreement. We can make some inferences about the likely level of annotator agreement by considering that the inter-annotator agreement for the original annotation of the BOPCRIS data report in Grover et. al. (2008) was 91.5%. Other published inter-annotation scores range from as much as 98% for modern newspaper text to 85% or lower for entities such as genes and proteins in biomedical text (Alex et. al. 2008). OCR quality also interacts with the issue of gold annotation: the

annotators marked up entities which 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 (sic=Comes) Rochester" should be a person and not a place but because the system didn't recognise "tomes" as a person title, it marked "Rochester" as a place.

3.3 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. To do the annotation Kate Byrne and Richard Tobin created a purpose-built web-based annotation tool. Figure 5 shows the tool in use with one of the Stormont Papers files. All of the test data for Histpop, BOPCRIS and the Stormont papers were annotated using this tool. The guidelines for the annotation task are included in Appendix 3.

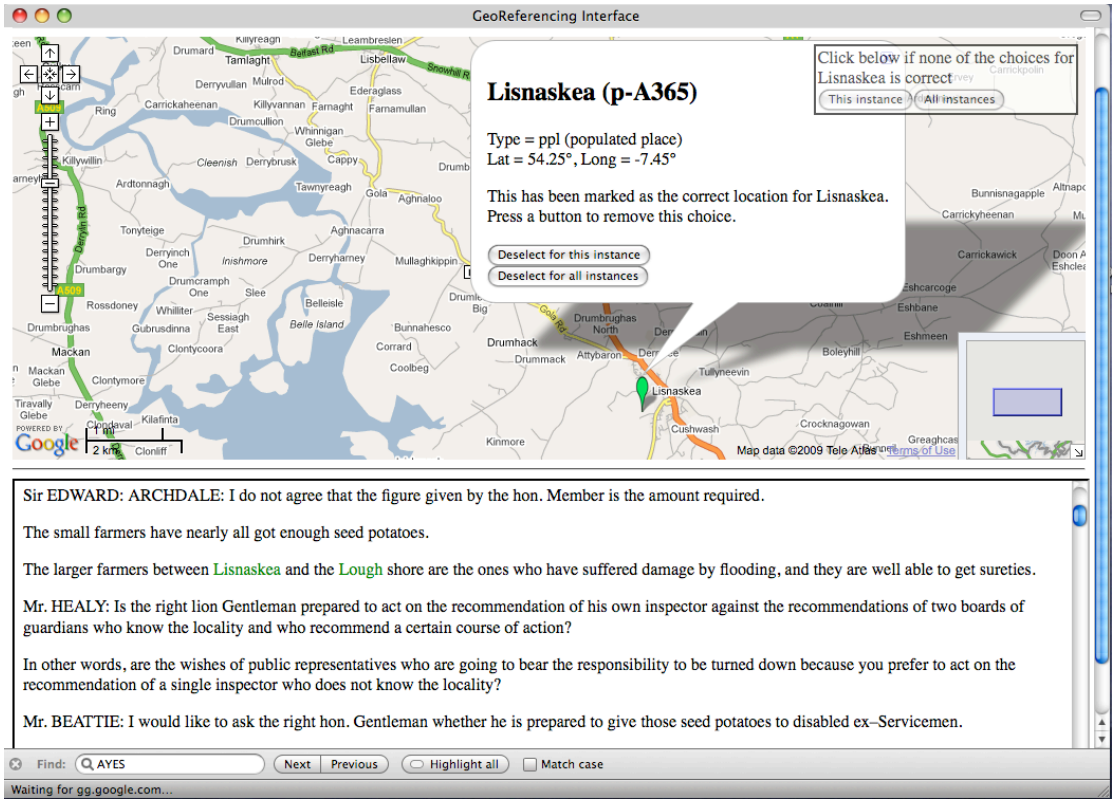

Figure 5. The georesolution annotation tool

The georesolver was evaluated against the test data under the following conditions:

- the input was the gold-standard entity mark-up to ensure evaluation of only the georesolver and not the full pipeline;
- the gazetteer entries for the resolver to rank were exactly the entries that were available to the human annotators – for Histpop and BOPCRIS there were two gold annotation sets, one using geonames and one using xwalk while Stormont was annotated only for geonames;
- the gold standard entity mark-up was augmented with linguistic context features (e.g. the contains/contained-by and abbrev-for attributes) in order to provide all the information that the georesolver would normally work with;
- georesolution was tested both with the user supplied locality parameter switched off and with it switched on. For Histpop and BOPCRIS the setting was "-l 55.45 -5.2 655 .3" (to cover the British Isles) and for Stormont the setting was "-l 54.6 -6.8 92 .3" (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 5km matching where gold and system choices would be counted the same if their grid references were within 5km 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:

1. no gazetteer entry was found during gazetteer lookup;
2. entries were found but the human annotator considered that there was no correct entry (they selected 'none');
3. entries were found but the human annotator neither chose one nor selected 'none' – we consider these to be annotation errors;
4. entries were found and the human annotator selected one of them as correct.

In the tables below we exclude cases 1 to 3 from the evaluation though we indicate the numbers of each of the cases. We exclude the second case as the system will always choose one of the entries because it is not designed to make a 'none of the above' judgement. The vast majority of instances fall under case 4 and here we look at the first ranked entry from the system. If the first ranked entry is the same entry as the gold then it is correct in the strict sense; if it falls within 5km of the gold entry then it is correct in a looser sense. The tables below also show 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.

	geonames	**xwalk**
documents	499[2]	500
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 5km without locality	3875 (76.1%)	4110 (92.7%)
correct within 5km with locality	4177 (82.0%)	4112 (92.7%)

Table 9. Georesolver evaluation on Histpop test sets

[2] One of the files was inadvertently omitted from the geonames annotation.

	geonames	xwalk
documents	93	93
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 5km without locality	523 (71.2%)	592 (91.1%)
correct within 5km with locality	598 (81.4%)	593 (91.2%)

Table 10. Georesolver evaluation on BOPCRIS test sets

	geonames
documents	12
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 5km without locality	855 (86.8%)
correct within 5km with locality	907 (92.1%)

Table 11. Georesolver evaluation on Stormont test set

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 xwalk, a result that reflects the difference suggested by the baselines. In both cases, however, the system achieves good results which are significantly better than the baselines. As would be expected, the strict measure of success gives rise to lower scores than the 'within 5km' measure. When using geonames, the use of the locality parameter results in higher scores. Its use with xwalk makes little difference since it provides the British Isles as the locality and xwalk 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 of these are OCR errors (e.g. "County of [flits") or possibly older spellings (e.g. "Materdale").

4. References

Beatrice Alex, Claire Grover, Barry Haddow, Mijail Kabadjov, Ewan Klein, Michael Matthews, Richard Tobin, and Xinglong Wang. 2008. The ITI TXM corpora: Tissue expressions and protein-protein interactions. In *Proceedings of the Workshop on Building and Evaluating Resources for Biomedical Text Mining at the 6th International Conference on Language Resources and Evaluation (LREC 2008)*.

James Curran and Stephen Clark (2003). 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)*.

Claire Grover, Sharon Givon, Richard Tobin, and Julian Ball. 2008. Named entity recognition for digitised historical texts. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*.

Ben Hachey, Beatrice Alex, and Markus Becker. 2005. Investigating the effects of selective sampling on the annotation task. In *Proceedings of the 9th Conference on Computational Natural Language Learning (CoNLL-2005)*.

Jochen L. Leidner. 2007. *Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names.* Ph.D. thesis, School of Informatics, University of Edinburgh.

Inderjeet Mani, Janet Hitzeman, Justin Richer, Dave Harris, Rob Quimby and Ben Wellner. 2008. SpatialML: Annotation scheme, corpora and tools. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*.

Mitchell P. Marcus, Beatrice Santorini and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: the Penn Treebank. Computational Linguistics 19(2).

Guido Minnen, John Carroll and Darren Pearce. 2000. Robust, applied morphological generation. In *Proceedings of INLG*.

Christoph Müller. 2006. Representing and Accessing Multi-Level Annotations in MMAX2. In *Proceedings of NLPXML-2006 (Multi-dimensional Markup in Natural Language Processing).*

Appendix 1. Original annotation guidelines for BOPCRIS

BOPCRIS Annotation Guidelines (Version 1.4.1)

Ian Hughson*

February 26, 2007

1 Persons

1. The names of individuals, including their titles, if given, should be annotated with the "Person" label. Neither *Gentleman* nor *M.P.* are titles, but *Esquire* is.

   e.g. *Gilbert; Cornelius O'Leary; Isaac Bernal the Younger; Thomas Johnes Esquire; Mr. Hugh Smith; Reverend Thomas Lawrence Dundas; Major General William Robertson; Sir John Charles Hamilton Baronet; John Duke of Atholl.*

2. The names of monarchs should include the ordinal.

   e.g. *Georgii Tertii; King William the 3d.*

3. Noble titles given without personal names should be annotated, whether they are in Latin or English, omitting any leading "the".

   e.g. *Comes Harrowby; March. Buckingham; Ds. Bolton; Viscount Sidmouth; Marquis of Donegall; Earl of Minto.*

4. Anaphoric uses of titles alone should **not** be annotated.

   e.g. *the said Earl.*

5. Titles attached to an office or position, rather than a person, should **not** be annotated.

   e.g. *Cancellarius; Praeses.; Lord Chancellor; Lord Chief Justice of the Court of King's Bench; Clerk of the Parliaments.*

6. For present purposes, item 1.1 above includes the name *George, Prince of Wales* and item 1.3 above includes the title *Prince of Wales*, while item 1.5 above includes the title *Prince Regent.*

*With help from Clair Grover and Laurence Durnan.

7. Honorifics should **not** be annotated, whether attached to a name or standing alone, with the exception of *Reverend, Very Reverend, Right Reverend, etc.* which should be treated under item 1.1 above as a title.

   e.g. *Honorable; Right Honorable; Most Honorable; Most Noble; His Grace; His Royal Highness.*

8. Names occurring as premodifiers or possessives should be annotated, excluding the possessive "'" or "'s", where present.

   e.g. *Francis Lee's Wife; Mrs. Jenkins's house.*

9. The names of individuals occurring as parts of company names should be annotated.

   e.g. *William Anderson and Company.*

10. The names of individuals occurring as parts of the names of legal cases should be annotated.

    e.g. *"Powles against Powles".*

11. Non-referential uses, i.e. where the name itself is being referred to rather than the person, should **not** be annotated.

    e.g. *William Cadell alias MacDonald; the Surname of Bootle Wilbraham in lieu of Wilbraham Bootle; Charles Newdigate Parker, calling himself Charles Newdigate Newdegate; "I asked for her by the Name of Mrs. Blaquiere."; "What was her Maiden Name?" "Louisa Chambers."*

2 Locations

1. The names of places should be annotated with the "Location" label..

   e.g. *Belhaven; Kilmarnock; Wales; Ireland; East Indies; United States.*

2. Addresses should be annotated as single entities, unless there is intervening material.

   e.g. *Buccleugh Street, Edinburgh* is one entity but *Lambeth Place in the County of Surrey* is two.

3. The names of rivers, counties, etc. should include the relevant prefix, if given, but omit any leading "the".

   e.g. *River Thames; Parish of Headon cum Upton; Ward of Bishopsgate; Tything of East Woodhay; Precinct of Saint Katherine; Lands of East Garravagh; Estate of Neidpath; Town of Saltcoats; Port of London; City of Bath; County of Hereford; West Riding of the County of York; Presidency of Bombay.*

4. The names of canals, bridges, and forests should be annotated.

   e.g. *Caledonian Canal; Stratford upon Avon Canal Navigation; Elland Bridge; Mytholm Royd Bridge; Saint John's Bridge; Countess Wear Bridge; Forest of Brecknock*

5. Names occurring as premodifiers or possessives should be annotated, excluding the possessive "" ' or "'s", where present.

   e.g. *East India Goods; Greenland Fisheries.*

6. The names of places occurring as parts of the names of organisations or buildings, other than canals and bridges, should be annotated.

   e.g. *Bank of Ireland; East India Company; Burslem Church.*

7. The names of organisations or buildings not covered by item 2.6 above should **not** be annotated.

   e.g. *British Linen Company; Gas Light and Coke Company; Swan Inn.*

8. Alternative names for places should be annotated.

   e.g. *Burnby, otherwise Burmby; Burnby, alias Burmby.*

3 Combinations

1. Co-ordinated person or place names should be annotated as separate entities, with any title or prefix attached to the one to which it is adjacent.

   e.g. *Messieurs Barclay, Perkins, and Company; John La Touche and Peter La Touche Esquires; Counties of Hereford and Surrey; Worcester and Birmingham Canal Navigation.*

2. Names with co-ordinated premodifiers or titles should be annotated with the head noun as a single entity.

   e.g. *Mr. and Mrs. Lee; East and West Teignmouth; Town and Borough of Deal; Town and County of the Town of Nottingham.*

3. Item 3.2 above combines with item 2.8 above as per the following example.

   e.g. *Forest of South otherwise East Bere, otherwise Bier.*

4. The names of places occurring as parts of personal titles should **not** be annotated separately as places.

   e.g. *Duke of Atholl; Marquis of Donegall; Prince of Wales.*

5. The names of persons occurring as parts of place names should **not** be annotated separately as persons.

   e.g. *Earl of Bridgewater's Canal; Countess Wear Bridge.*

6. The names of places intervening between personal names and titles should be annotated separately as places, while titles dislocated by such intervening material should be ignored.

   e.g. *John Porter of Lambeth Place, in the County of Surrey, Esquire.*

4 Extraneous Material and Interruptions

1. The names of persons or places occurring in marginal notes should **not** be annotated.

2. Names or places interrupted only by extraneous material should be annotated as single entities, with the extraneous material also independently annotated as "Interrupt".

   Extraneous material consists of (a) marginal notes and/or (b) quotation marks repeated at the start of a new line. Both hyphenation and additional or erroneous characters resulting from the OCR processing should be treated as normal text, except where they fall under (a) or (b).

   e.g. *Charles Mar-Q^^r*
   " quefs of ^ueenfberry
   should be annotated whole as a person, with only
   Q^^r
   "
   annotated also as "Interrupt".

3. The names of persons or places interrupted by non-extraneous material, such as text from another column mistakenly inserted by the OCR process, should **not** be annotated.

   e.g. In the following series of mistakenly interwoven questions and answers, only *Mrs. Jenkins* should be annotated, not *Mr. James* or *Christie*:

   "And did you see Mr. James
   "To Mrs. Jenkins's house."
   "Christie there?"
   "And how long did you stay there?"

4. The names of persons or places rendered incomplete by quotation in amendments to parliamentary bills or other documents should **not** be annotated.

   e.g. *Leave out ("Frances") and insert ("Ann"); Leave out ("Thomas") and insert ("John Mallows"); After ("Croke") insert ("Doctor of Laws").*

Change log

v1.4:
- Improved cross-referencing.
- Included forests as locations to annotate.
- Inserted item explicitly accounting for locational premodifers.
- Inserted item clarifiying interaction of co-ordination with alternative names.
- Various alterations and additions to examples.

v1.3:
- Reversed policy on personal names in company names.
- Inserted items including place names in company and building names and excluding other company and building names.
- Inserted items accounting for legal case names, alternative names, and non-refereential uses.
- Inserted item excluding personal names within place names.
- Inserted item including canals and bridges.
- Inserted item accounting for co-ordinated pre-modifiers.
- Inserted item accounting for non-extraneous interruptions and incomplete references.
- Various alterations and additions to examples.

v1.2:
- Made exclusion of "Gentleman" explicit.
- Excluded "M.P."
- Clarified item on titles without personal names and excluded anaphoric uses.
- Made treatment of Reverend explicit.
- Inserted item explicitly accounting for personal premodifers.
- Inserted item excluding place names within personal names.
- Inserted item accounting for Interrupts.
- Various alterations and additions to examples.

Appendix 2. Geotagging guidelines for Histpop and Stormont

Rules for annotation

Matthew Woollard

March 4, 2009

Version 2

All items below in italics denote the extent of the annotation.

1 Places

1.1 Geo-political places

All geo-political place names should be annotated as a `location-geop`. When place names appear on their own in the text — either in full or abbreviated — they should be tagged alone; when place names are reported with a descriptive form of the unit, the whole of the unit should be marked up.

- **continents**: *Africa*.
- **countries**: *England*.
- **counties**: *Hertfordshire*; *Herts*; *County of Bedford*; *Shire of Aberdeen*.
- **islands**: *Guernsey*; *Islay*; *Isle of Mann*.
- **sub-county units**: *Lindsey*; *Milford*; *Lonsdale Hundred*; *Milford Tything*; *Wapentake of Staincliffe and Ewcross*; *West Riding*; *West Riding of Yorkshire*; *Parish of St. Bartholomew*.
- **other sub-units**: *South Wales* (but beware some contexts: south *Wales*).
- **definable groups of any of the above**: *Highlands*, *Northern Division*.
- **cities, towns, villages, hamlets, etc.**: *London*, *Newport Pagnell*, *Caerphilly*, *Hutton le Hole*.

Note that occasionally two place names may be mentioned in the text in an overlapping form, e.g., East and West Sussex. There are special rules for annotating these as two separate places, e.g., *East Sussex* and *West Sussex*. See the Instructions for Annotation for details on how to annotate these.

Note also that there may be occasions where one place name may be best annotated within another one. For example *St. Agnes, Truro* is the "official" name of a parish, but the geographic references which will be used later in the project may only reference *Truro*, so for the purposes of annotation this place should be marked up as three places, e.g., *St. Agnes, Truro*, *St. Agnes* and *Truro*. For clarification, some administrative units and places will include compass points. These are usually capitalised (see the example above). However, over the whole of the century there are some typographical changes, and in some documents, generic areas are referred to with capitalisation, e.g., the North of England, or North England. The context is paramount for the tagging of these places.

1.2 Geographical features

Geographic names and features which are not "geo-political" should be marked up as `location-feat`. This includes:

- **rivers, canals, etc.**: *river Thames*; *Firth of Forth*; *Forth and Clyde Canal*.
- **mountains**: *Rocky Mountains*.

1.3 Otherplaces

1.3.1 Streets, roads, etc.

Streets, roads and other similar geographical information appearing in the text should be marked-up as `location-other`. Care is needed with these "other" features as some things which look like street names may actually be the name of a parish or an other administrative unit, e.g., *Glasshouse Yard* is a parish in London, *Liverpool Howard-street*, is a Registration District (named after a street) in Liverpool. Some street names also form an element of a parish name, e.g., *St Mary Magdalen Old Fish Street*. Furthermore, some places use the -street element without necessarily also referring to a street, e.g., *Charnham-Street* which is not a street in itself. Addresses should also be marked-up in their entirety.

- **streets, roads, etc.**: *Camomile-Street*.
- **addresses**: *13, Great Queen Street, W.C.*

1.3.2 Physical structures

Physical structures per se, appear very occasionally. Sometimes these can be mistaken for places, e.g., *Hebden Bridge* or *St Nicholas Cole Abbey* (both parishes). It is suggested that all such physical structures are marked-up for the removal of ambiguity as a `location-other`.

- **bridges, etc.**: *London Bridge*; *Port of London*.
- **significant landmarks**: *Houses of Parliament*.

1.3.3 Possessives

Place names acting as possessives or pre-modifiers should be tagged to whichever category they fall into. Thus *Australian Colonies* is a `location-geop`. However, adjectival forms of place names acting in the same sense, e.g., Australian aborigines should not be tagged. Similarly, German empire, Indian empire are not to be tagged. Note that there are places used to describe stuff, e.g., 'York regents' (a type of potato), 'Berlin wool repository keeper' and 'Barbadoes leg'. These should not be tagged.[1]

- **possessives, etc.**: *East India* government, Church of *England*.

1.3.4 Interruptions

Places described as interruptions caused by hyphenation in the original text or poor OCRing should be tagged as though they were a place.

- **Interruptions**: *Metro- polis*.

1.4 Words

Words which are used as place names which are used in a different context as other words should not be tagged, e.g., **china** maker, **cork** cutter, **flint** dealer, and so **forth**.

1.5 Additions made during tagging process

1.5.1 Company names

Words which are used as place names within company / society names have not been tagged, e.g., City of London Company, Meteorological Society of Scotland, etc.

[1] Berlin wool repository keeper, etc., is perhaps best described as a middleman for the sale of embroidery. He'd commission and buy the embroidery work, which would then be sold on to other retailers or to individuals. It seems to be named from Berlin as the place where the patterns were first made. Barbados leg is "a species of elephantiasis incident to hot climates" – caused by worms!

1.5.2 Places in Legislation

Words which are used as place names within Acts of Parliament have not been tagged, e.g., Dublin Corporation Act.

1.5.3 Non-contiguous places and collective places

The General Register Office frequently used the collective noun Islands in the British Seas to refer to the Isle of Man and the (eight inhabited) Channel Islands. This phrase has not been tagged.

The term "the Continent" appears occasionally. Though this is usually the 'European' continent it has not been tagged, neither has the word Kingdom when it appears on its own.

2 People

All personal names shall be tagged. Two different attributes are used. The first, `person-name` is for when people are being referred to; the second `person-other` is to be used to mark words or phrases which refer to people, but are not people themselves. The context will generally make the distinction clear. For example, *John Menzies* and *T. Fisher Unwin* may be tagged either as `person-name` and `person-other` depending on whether or not the text is referring to a person or a (publishing) company which is being referred to.

All diseases which include a personal name as a premodifier, e.g., Bright's disease (named after Joseph Bright) should be tagged as `person-other`, as should industrial processes named after a person, e.g., *Betty Dodd Mangler*.

2.1 Names

- **Names**: *John Cox*; *A. Jacob, M.D.*; *Mr. C.A. Whitmore, M.P.*; *Ramazzini*; *King Alfred*; *Constantine the Great*; *Bishop of Rome*.

2.2 Other nominal information

- **Names**: *Hodgkin*'s disease; *John Menzies*; *Bright*'s disease.

2.3 Additions made during tagging process

2.3.1 Other nominal information

Diseases which include a non-English language premodified, i.e., Gaelic have not been tagged (e.g., Teinas Phoil (St. Paul's disease)).

Personal names which occur within the context of parliamentary legislation, e.g., "23 Vict. c.24" have not been tagged.

Non-specific job titles: 'the King', 'the Pope', 'the Emperor of China' have not been tagged.

Appendix 3. Georesolution guidelines for Stormont

Rules for georeferencing

Vasilis Karaiskos
vkaraisk@staffmail.ed.ac.uk

March 12, 2009

In this task, placenames (geo-political ones, as well as geographical features) found in the Stormont papers are linked to their corresponding points on a map.

1 The tool

The annotation tool itself can be accessed at:
http://www.cogsci.ed.ac.uk/ richard/geo/stormont/geonames/

At the top of the page, there is a drop-down list of the documents available for annotation. In order to load a document, you can select it on the list, and click **Load**. The document will appear at the bottom half of the browser window.

The document itself will have three kinds of different highlighting:

Red : placenames that need to be found on the map.
Green : placenames that have already been found on the map.
Blue: placenames that have been marked up in the original annotation, but which have not been found in the gazetteer(s) being used.

In order to link marked placenames in the document with points on the map, you need to click on a red item in the document. A number of red markers are going to appear on the map. They represent the candidate points to be linked to the placename.

Click on the marker of your choice, and then click on **Select for this instance** or **Select for all instances**, depending on whether you think that this placename may have different referents on the map or not. The marker will turn green.

Similarly, if you change your mind, you can click on a green marker and deselect your original choice. The marker will turn back to red.

Sometimes, none of the markers on the map correspond to the placename mentioned in the document. In those cases, you need to click on either of the two buttons at the top right corner of the map, under the heading *Click below if none of the choices for <placename> is correct*. You should choose **This instance**, if it is just the current mention of the placename that has not been found on the map; **All instances** for cases where this is also the case for any subsequent mention of the same placename.

The placenames highlighted in blue in the document can be ignored.

The new annotation needs to be saved by clicking on the button at the top of the browser window.

2 Ambiguity resolution

This is, generally, a very straightforward annotation. The few ambiguities that arise have been resolved as follows:

administrative division vs. populated place

Some placenames, in particular cities, towns etc. may have multiple markers on the map for the same placename. Typically, one marker will correspond to some *administrative division* (council etc.), and another to the place itself (*populated place*). In those cases, select the *populated place* marker.

[...]an incompatibility of interest between the Minister of Finance and a merchant bank operating in ***Belfast***

However, a county name and a city name may be used interchangeably to refer to either. In those cases, you should select *populated place* when the referent is the city/town, or it is unclear what the referent is; select *administrative division* (or *region*) when the referent is the couny.

the lands recently bought by the Ministry near ***Londonderry***. (populated place)

Counties. Sergeants and Constables.
[...]
Londonderry *. 198* (region)

country vs. populated place vs. administrative division vs. region

The above types of markers (usually a subset thereof) may appear as referents on the map for the same placename. Is such cases, the order of preference should be as follows:

1. country
2. populated place
3. region
4. administrative division

I happen to be pretty familiar with the development of ***New Zealand*** (region)

I think it is a very deplorable thing that in this small area, which is neither ***Ulster*** *nor the Free State, neither* ***England*** *nor* ***Scotland****, nor any other part of the world, an area which has no individuality of its own, there should be such a tale of woe and depression at this stage of its history.* (region, region, administrative division, respectively)

multiple markers of the same type

In some cases, there will be more than one marker of the same type referring to the same placename. In those cases, and if the context does not help the choice in any way, choose the most northern one.

The hon. Member wants me to deal with the question of the ***Belleek*** *potteries.* (two markers for *populated place* appear on the map)

Appendix 4. Georesolution guidelines for Histpop

Georeferencing rules and instructions

Matthew Woollard

March 23, 2009

1 The annotation tool

The annotation tool is accessed at the following URLs. [1]

http://www.cogsci.ed.ac.uk/~richard/geo/histpop/geonames-000/
http://www.cogsci.ed.ac.uk/~richard/geo/histpop/geonames-100/
http://www.cogsci.ed.ac.uk/~richard/geo/histpop/geonames-200/
http://www.cogsci.ed.ac.uk/~richard/geo/histpop/geonames-300/
http://www.cogsci.ed.ac.uk/~richard/geo/histpop/geonames-400/
http://www.cogsci.ed.ac.uk/~richard/geo/histpop/geonames-500/
http://www.cogsci.ed.ac.uk/~richard/geo/histpop/geonames-600/
http://www.cogsci.ed.ac.uk/~richard/geo/histpop/geonames-800/
http://www.cogsci.ed.ac.uk/~richard/geo/histpop/geonames-900/
http://www.cogsci.ed.ac.uk/~richard/geo/histpop/xwalk-000/
http://www.cogsci.ed.ac.uk/~richard/geo/histpop/xwalk-100/
http://www.cogsci.ed.ac.uk/~richard/geo/histpop/xwalk-200/
http://www.cogsci.ed.ac.uk/~richard/geo/histpop/xwalk-300/
http://www.cogsci.ed.ac.uk/~richard/geo/histpop/xwalk-400/
http://www.cogsci.ed.ac.uk/~richard/geo/histpop/xwalk-500/
http://www.cogsci.ed.ac.uk/~richard/geo/histpop/xwalk-600/
http://www.cogsci.ed.ac.uk/~richard/geo/histpop/xwalk-800/
http://www.cogsci.ed.ac.uk/~richard/geo/histpop/xwalk-900/

Each of these URL's link to a series of pages from the histpop website. The sub-directories xwalk and geonames represent the two different gazetteers which were used to georeferencing the marked up place names. You should work through each directory from gro

At the top of each page, there is a drop-down list of the documents available for annotation. To load a document (e.g., 27_62_a1) you can select it on the list, and click **Load**. The document will appear at the bottom half of the browser window.

The document itself will have three kinds of different highlighting:

[1] This document is based on a guide 'Rules for Georeferencing' by Vasilis Karaiskos, 12 March 2009.

- Red : place names that need to be found on the map;
- Green : place names that have already been found on the map;
- Blue: place names that have been marked up in the original annotation, but which have not been found in the gazetteer(s) being used;

In order to link marked place names in the document with points on the map, you need to click on a red item in the document. This will result in a number of red markers appearing on the map. These represent the *candidate* points to be linked to the place name. Clicking on the marker will pull up an information box which will give the type of place (see below) and two options. You can either:

- **Select for this instance** or
- **Select for all instances**.

If you think that the same place name may have multiple referents on the map *in this particular page* then **Select for this instance** otherwise **Select for all instances**. The marker will turn green.

If you change your mind, you can click on a green marker and de-select your original choice. The marker will turn back to red. Sometimes, none of the markers on the map correspond to the place name mentioned in the document. In those cases, you need to click on either of the two buttons at the top right corner of the map, under the heading **Click below if none of the choices for <place name> is correct**. You should choose **This instance**, if it is just the current occurrence of the place name that has not been found on the map or **All instances** for cases where this is also the case for any subsequent appearance of the same place name.

The place names highlighted in blue in the document can be ignored. It is important to explicitly save once you have finished a page, before loading the next page. If you use the back button it will also lose any annotations you have made since you last saved.

2 Resolving ambiguity

Ambiguity will occur when a place name occurs more than once in a gazetteer. London may, for example, refer to a place in Canada or a place in England. In most cases ambiguities will be easily visible on the map.

2.1 Administrative place vs populated place

Some place names, in particular cities, towns etc. may have multiple markers on the map for the same placename. Typically, one marker will correspond

to some *administrative division* and another to the place itself (*populated place*). In the histpop material most places will be *administrative divisions*, so the place defined as such should be selected unless the text is explicitly about a populated place.

County names and city names are often used interchangeably (e.g., the county of *Lincoln*, the town of *Lincoln*. Again context should be used to make a decision, but by default *administrative division* should be selected.

2.2 Other possibilities

Other types of marker may also appear on the map for the same place name. The order of preference should be as follows:

1. Country;
2. Administrative division;
3. Region
4. Populated place

Again, context should be the best guide for these.

2.3 Multiple markers of same type

It may occur that there will be more than one marker of the same type referring to the same place name. This will most generally occur with *other* types. In case of ambiguity for what are really populated places, e.g., towns, cities, etc. select the place nearest to what looks like the centre on the map. In cases of administrative units select the most northern possibility.