Defining and identifying the roles of geographic references within text: Examples from the Great Britain Historical GIS project

Humphrey Southall
Department of Geography
University of Portsmouth
Buckingham Building
Lion Terrace
Portsmouth PO1 3HE
ENGLAND
Humphrey.Southall@port.ac.uk

Abstract

Reliably recognizing, disambiguating, normalizing, storing, and displaying geographic names poses many challenges. However, associating each name with a geographical point location cannot be the final stage. We also need to understand each name's **role** within the document, and its association with adjacent text. The paper develops these points through a discussion of two different types of historical texts, both rich in geographic names: descriptive gazetteer entries and travellers' narratives. It concludes by discussing the limitations of existing mark-up systems in this area.

1 The Great Britain Historical GIS

The Great Britain Historical GIS is a very large assembly of historical information about Britain, all in some sense tied to particular places. The earliest data in the system was computerised in the late 1970s, and it was established as a relational database in 1989-91. Until recently, however, almost our entire content was either statistical or locational: by now, we have computerised or acquired from collaborators a substantial fraction of the information published in the reports of the Censuses of Population for England and Wales, and for Scotland; and of the information published in vital registration reports for the same areas since the 1840s.

In general, our coverage ends in the early 1970s, when the relevant information began to be published in digital form. Our statistical database by now comprises over 33m. data values, and is closely linked to digital mapping containing the changing boundaries of the various statistical reporting units: counties, various types of district, and approaching 20,000 parishes. This material has formed the basis for studies of demographic, economic and social change.

However, our largest source of current funding has a different focus. A grant from part of the UK National Lottery is turning the GBH GIS into an on-line resource for 'life-long learners', which in practice means the general public. Our system is not a conventional on-line GIS as our most obvious audience are people interested in local history: the most basic functionality of our site allows users to specify a location by a place-name or, preferably, a postal code which in the UK identifies a group of maybe ten houses, and therefore a fairly precise location; they will then be taken to a page providing information on how their current local authority – there are 408 in Great Britain – has changed over the last 200 years, with the option to drill deeper by accessing information for the various past administrative units which contained the location they specified. An initial system, limited to the above functions for the modern units, should go live during May 2003:

www.VisionOfBritain.org.uk

Later versions of the site will contain vastly greater content. We will provide access to data for the original historical units, using 'point-in-polygon' searching of a spatial database to identify relevant units, having first converted the users' postal code into a geographic coordinate. The site will draw on both vector mapping of historic boundaries and two complete set of geo-referenced image scans of historic maps of Great Britain at one inch-to-the-mile scale: Firstly, we are scanning two complete editions of Ordnance Survey One Inch-to-the-Mile maps of Great Britain: the New Popular Edition, published in the late 1940s and the first to include National Grid lines, simplifying geo-referencing all the scanned maps; and the nineteenth century First Series. Geo-referencing the latter will be challenging, and require extensive 'rubber sheeting'.

We are computerising two existing inventories of historical administrative units covering England and Wales, and constructing an equivalent digital resource covering Scotland. These inventories are not gazetteers but systematic lists of all the counties, parishes and various kinds of districts that ever existed, with their hierarchic relationships and some information on variant names. This information provides the absolute core of our system, structured as an ontology rather than as a strictly hierarchic thesaurus. The core ontology does not require locational information for units, but if available locations are stored as polygons representing the boundaries, with dates of creation and abolition.

In our final system, we will be able to offer 'home pages' not just for the 408 modern districts but for over 20,000 historic units, including the parishes which were the lowest level of administration until recently, and generally correspond with individual villages. Each home page will contain a map showing the overall location within Britain and a short description generated from the database and highlighting key statistics. From the home page, users will be able to access a more local map showing the unit's boundary, a range of statistics mostly presented graphically, and information on the unit's history including boundary changes and hierarchic relationships.

Relative to the overall size and scope of the site, our own capacity to author descriptive and explanatory text is limited. We will concentrate on the text to accompany maps showing national patterns. However, a site that was largely limited to statistics, even presented as maps and graphs, would be pretty boring and we are therefore computerising a large quantity of text from existing publications. This text forms the main subject of this paper.

There are in fact three types of text. Firstly, we are computerising the introductions to all the census reports between 1801 and 1851, to provide a description of the country as a whole. This aspect of the project is not further discussed here.

Secondly, we are computerising three descriptive gazetteers published in the late nineteenth century, totalling over 4,000 pages and containing about 5m. words:

- John Bartholomew's Gazetteer of the British Isles (Edinburgh, 1887). This covers the whole British Isles, including Ireland.
- John Goring's Imperial Gazetteer of England and Wales (Edinburgh, 6 vols., 1870-72)
- Frances Groome's Ordnance Gazetteer of Scotland, (Edinburgh, 6 vols., 1882-85). Our work on this is a collaboration with the Gazetteer for Scotland project.

Even with the gazetteer text, it will be very easy for users to locate information about very specific places, but much harder for them to move around the system to explore the relationships between places, and form a 'vision of Britain through time' as a whole. This justifies our third and final type of new content: narratives describing historical journeys around Britain. We are computerising four well known accounts, as well as some shorter accounts written by radical agitators as they moved around in the mid-19th century:

- **William Cobbett**, *Rural Rides* (London, 1830).
- **Daniel Defoe**, *A Journey through the whole island of Britain divided into circuits or journies* (London, 1724-7).
- **Celia Fiennes**, *Through England on a side saddle in the time of William and Mary, being the diary of Celia Fiennes* (London, 1888).
- **Arthur Young**, *Tours in England and Wales, Selected from the Annals of agriculture* (London, 1784-98).

This list has been deliberately kept short, as we almost certainly have the capacity to digitise more books via our Optical Character Recognition system but not necessarily to mark them up. Three obvious additions would be the journals of John Wesley, the Torrington Diaries and Boswell and Johnson's tour of the Hebrides.

2 Descriptive Gazetteers

The descriptive gazetteers form a very large body of text, but fortunately they are highly structured, making automated parsing feasible. The parsing software runs within our Oracle database and is written in SQL and PL/SQL. I have no doubt that it would be both more efficient as well as more effective if it were written in, say, Perl. Little more will be said about the software, other than to note that its relative effectiveness is mainly evidence of the vital importance of having a large database of placenames already built.

Although we are working with three different books, all are written to a broadly similar formula:

- Each consists of alphabetically arranged entries; each entry begins with a **head-word**, i.e. the place-name usually in bold or upper case letters.

- The head-word is followed by an indication of the **feature type** (‘a parish’, ‘a river’, etc).
- Third comes some indication of where the feature is, which almost always indicates a county, sometimes a relative location (‘9 miles SW of Worcester’) and never an absolute location such as latitude and longitude.

The main differences between the books is that Bartholomew’s consists of a very large number of short entries while the *Imperial Gazetteer* and Groome’s provide longer entries, those for major cities and counties covering several pages. Mostly, however, we focus on the first sentence as outlined above. Here are some samples, firstly from the very beginning of Bartholomew’s:

- **A'an,** or **Avon,** lake, S. Banffshire, among the Cairngorm mountains, 1_ mile long, at alt. of 2250 ft.; it is the head-water of river Avon: which see.
- **Aasleagh,** place, co. Mayo, 16 m. S. of Westport; P.O.
- **Abbas and Temple Combe,** par., mid. Somerset, 4miles S. of Wincanton sta., 1850 ac., pop. 590.
- **Abbenhall.** See ABENHALL.
- **Abberley,** par. and seat, W. Worcestershire, 4 miles SW. of Stourport sta., 2636 ac., pop. 605; P.O.
- **Abbert**, seat, 10 miles NE. of Athenry, co. Galway.
- **Abbertoft,** hamlet, Willoughby par., mid. Lincolnshire, 2 miles SE. of Alford.
- **Abberton.**—par., E. Essex, on Roman road, 4 milesS. of Colchester, 1068 ac., pop. 244; P.O.—**2. Abberton,** par. and seat, E. Worcestershire, on river Piddle, 4miles NE. of Pershore sta., 1001 ac., pop. 92.
- **Abberwick,** township, Edlingham par., N. Northumberland, on river Alne, 3 miles W. of Alnwick, 1680 ac., pop. 109.
- **Abbess Roding.** See ABBOTS ROOTHING.
- **Abbethune,** seat, 1 m. from Inverkeilor sta., Forfarsh.

Secondly, from the **Imperial Gazetteer**:

- **AFTON**, a village 2 miles S of Yarmouth, Isle of Wight. Afton House adjoins it, on a pleasant slope toward the Yar. Afton Down rises in the south-eastern neighbourhood, overhangs the English Channel, has an altitude of about 500 feet, and is crowned by tumuli.
- **BINSTEAD**, a small village and a parish in the Isle of Wight. The village stands on the coast of the Solent, amid charming environs, 1_ mile W by N of Ryde. The parish comprises 1,140 acres of land and 335 of water; and its post-town is Ryde. Real property, £2,775. Pop., 486. Houses, 105. The manor belonged, at the Conquest, to William Fitz-Stur; and passed to the Bishops of Winchester. Several picturesque villas, one of them belonging to Lord Downes, stand near the village and on the coast. Quarr Abbey House is the seat of Admiral Sir Thomas J. Cochrane. Remains of a Cistertian Abbey, called Quarr Abbey, founded in 1132, by Baldwin de Redvers, afterwards Earl of Devon, stand at a farmstead, 5 furlongs west of the village; and, though fragmentary and mutilated, show some interesting features. A siliceous limestone, containing many fossils, and well suited for building, has been extensively quarried since at least the time of William Rufus. The living is a rectory in the diocese of Winchester. Value, £80.* Patron, the Bishop of Winchester. The church was rebuilt in 1842; is in the early English style; and embodies some sculptured stones of a previous Norman edifice.
- **BRAMBLE CHINE**, a small ravine on the NW coast of the Isle of Wight; at Colwell bay, 2 miles SW of Yarmouth. A thick bed of oyster shells, in a fossil state, is here; the shells in the same position as in life, but entirely decomposed.
- **CALBOURNE**, a village, a parish, and a sub-district in the Isle of Wight. The village stands 5 miles WSW of Newport; and-has a post-office under Newport. The parish includes also Newtown borough; and extends from Brixton Down to the Solent. Acres, 6,397; of which 265 are water. Real property, £4,471. Pop., 728. Houses, 145. The property is divided among a few. Westover manor belonged to the Esturs; passed to the Lisles and the Holmeses; and belongs now to the eldest son of Lord Heytesbury, in right of his wife, the daughter of the late Sir Leonard W. Holmes. The house on it is modern; and the grounds are tasteful. Calbourne Bottom, 1_ mile SSW of the village, is a depression between Brixton and Moltestone downs. The living is a rectory, united with the p. curacy of Newtown, in the diocese of Winchester. Value, £675.* Patron, the Bishop of Winchester. The church is early English, much modernized; and has a brass of 1480.—The sub-district contains eight parishes. Acres, 25,050. Pop., 5,417. Houses, 1,071.
- **WIGHT (Isle of)**, an island in Hants; bounded, on the N, by the Solent,–on the other sides, by the English channel. Its outline is irregularly rhomboidal, and has been compared to that of a turbot, and to that of a bird with expanded wings. Its length from E to

> W, from Bembridge Point to the Needles, is nearly 23 miles; its greatest breadth from N to S, from West Cowes to St. Catherine's Point, is 13_ miles; its circuit is about 56 miles; and its area, inclusive of foreshore, is 99,746 acres. The general surface has a considerable elevation above sea-level. The coast, along the N, is low; around the W angle, is rocky, broken, precipitous, and romantic; and along the SW, the S, and the SE, breaks down in a richly varied series of cliffs, often abrupt or mural, extensively terraced and lofty, including all the magnificent range known as the Undercliff, and everywhere replete with scenic interest. The water-shed uniformly follows the trending of the S coast; and is distant from it never more than 2_ miles, generally less than 1 mile. A range of downs extends about 6 miles from St. Catherine's Hill to Dunnose; rises from the shore, with excessive steepness, to a height of nearly 800 feet; and is marked, along its steep sea-front, with the picturesque terraces of the Undercliff. A diversified range of downs extends about 22 miles, from the Needles on the W to Culver cliff on the E; commences in grand cliffs about 600 feet high; runs 9 miles nearly due east, in a single, sharp, steep ridge, to Mottiston; attains there its highest altitude, at 662 feet above sea-level; makes several debouches in its subsequent progress; suffers repeated cleaving and disseverment, in the form of gaps or depressions; assumes, for some distance, in the neighbourhood of Carisbrooke, the character of a double or a triple range; is, in some parts of its course, saddle-shaped and slender,–in other parts, broad-based and moundish; and divides the island into two pretty nearly equal sections. A transverse ridge, about 400 feet high, extends about 3 miles in the contiguous to the river Yar; and another transverse ridge, tame in feature, but sometimes of considerable height, extends between the Medina and the Brading. The rest of the surface is either undulating or gently sloping, and has little or no claim to be called picturesque. The chief streams are the Yar, the Newton, the Medina, the Wooton, and the Main or Brading. The geognostic structure comprises chiefly lower greensand in most of the S, chalk in part of the centre, and upper eocene in most of the N; but includes many details, possesses deep interest, and may advantageously be studied with the aid of Mantell's and Martin's manuals. [just the first paragraph of a long entry]

The second example from the *Imperial Gazetteer* demonstrates the main reason we need to do a limited amount of work on the whole entry, not just the first sentence. For Binstead, the feature type clause is 'a small village and a parish', meaning that the place-name is associated with more than one entity. In extreme cases, a single entry covers four distinct entities, so for example Ledbury in Herefordshire was described as being 'a small town, a parish, a sub-district and a district'. The last three of these terms are all distinct entities within our ontology, and the entries for such places are in fact divided into a series of sections, each beginning with the type of sub-entry. For Binstead, the first part begins 'The village' and is just a single sentence; the second part begins 'The parish'.

The texts begin by being scanned in by a specialised Optical Character Recognition system optimised for historic materials, operated by our team based with the Centre for Data Digitisation and Analysis at the Queen's University Belfast. The OCR output is then visually scanned and tidied up by Information Technology trainees there, and delivered to the project's main team as Microsoft Word files replicating the source documents as closely as possible.

The first stage in our work is breaking the text down into the individual entries, each of which is then loaded as a separate record into our database. The way the text so clearly divides into such discrete sections greatly simplifies how we handle it. One consequence is that, for now, we are not applying any mark-up system to the text itself, other than basic HTML tags to preserve basic formatting, such as bold and italics. Instead, additional structure and search facilities are provided by adding additional columns to the table.

What follows describes the parsing process for entries from Bartholomew's:

- Firstly, the end of the head word is identified simply from it being in bold face, and the head word is copied into another column. NB with the gazetteers, identifying the most important geographical name within the text is fairly trivial.

- Entries which are cross-references are identified from their containing specific phrases immediately after the head word: 'See', 'also called XXX, which see', 'another name of XXX, which see' and, at present, 'Welsh name of XXX, which see'. The system then searches for the cross-referenced name elsewhere in the table.

- The system then tries to identify the feature type by brute force methods: all the strings immediately following the headword in the first two thousand entries were extracted and sorted, and the section identifying the feature type isolated to give 410 distinct type strings. Each of these was then marked up firstly with a version of itself in which all abbreviations were expanded, and secondly with three flags indicating whether the type indicated the entry was for a county, a parish or a borough. Longer term, these 'original' feature types will be mapped onto the Alexandria Digital Library Gazetteer Feature Type thesaurus.

- The next major stage is the identification of the county, which almost every entry includes. Our core ontology contains a complete list of all counties in the British Isles, together with some variant names. We **know** that one of these names must appear in each entry, so brute force methods are used to find them, starting with the first clause of the first sentence, then looking at the second clause and so on up to the eighth.
- A similar method is used to find parish names, looking for various text strings which indicate a reference to a parish, such as 'and forming part of XXX' or 'the hamlet is in XXX'. NB as we have already identified the county, the system matches only parish names in the correct county.

What current procedures do not do is systematically associate each entry with a point location. This may not be necessary. Almost every single entry is associated with a county defined within our ontology, and the county will be associated with a polygon. Further, a great many entries are also associated with parishes, or actually are entries for a parish: each parish covers a few square miles, which is sufficiently precise for many purposes. However, we are examining methods for linking each entry with a single grid reference. Three approaches are possible:

- In principle, it would be possible to parse the data on relative locations within the entries, such as '16 m. S. of Westport' or '2 miles SW of Yarmouth'.
- In practice, we will first attempt to add locations by running the entries against a modern gazetteer. One possibility is to use the Geo-X-walk gazetteer constructed by collaborators at EDINA and query it via the ADL gazetteer service protocol, using our county polygons to specify the area to be searched.
- As always, neither method is likely to be totally effective, and a small number of entries could be located manually.

So far, we are not trying to systematically extract other information from the entries. Although parish and district entries contain certain items of statistical data in a largely predictable way, this is almost all taken from census reports which we have separately computerised. For example, the district entries always including data on numbers attending each type of church, but this comes from the report of the 1851 census of religion, which we also hold.

We have built a demonstration system containing a variety of information for the Isle of Wight, an island and small county just off the south coast of England. This includes all relevant entries from the Imperial Gazetteer, and the system also includes the first 1,720 entries from *Batholomew's*. Of these, 54 are cross-references to other entries, and of the remaining 1,666 the parsing software has associated 1,609 (97%) with counties. This system can be accessed on-line at:

`http://www.gbhgis.org/demo_gaz.htm`

Although this prototype currently lacks mapping capabilities, it does show the high level of integration we have been able to achieve between the descriptive gazetteers and other content.

Summing up this discussion of our work with descriptive gazetteers, the formulaic nature of the original texts is of great assistance in automated parsing. For any given type of entry, we know what kinds of geographical names to expect:

- **Subject**: the name of the place itself. Easily identified as it is the head word.
- **Containers**: higher level units that contain the subject. We always look for a county, and sometimes look for a parish. In either case, our existing database provides an authoritative list of valid units.
- **Relatives**: these are the generally larger settlements that are mentioned in relative locations. Identifying them from the surrounding phrasing should be straightforward, and it may prove helpful in identifying 'larger' places that our database also contains a mass of population statistics.

3 Travellers' Accounts

The travellers' accounts are much less formulaic than the descriptive gazetteers and constitute a smaller body of text. While work on the gazetteers is well advanced, methodologies for the travel narratives are still being explored and our expectation is that we will use semi-manual methods, identifying place-names and other elements by reading through the text but using software to assist in adding mark-up. The paper's concern is not with this process, but with what features we should be identifying and how best to mark them up.

Celia Fiennes' travels in the 1690s are perhaps the purest example of journeys both undertaken and described for purely personal reasons. They were 'begun to regain my health by variety and change of aire and exercise' (p. ix), and her account was not published for nearly two hundred years. Here is her account of visiting Stonhenge:

> Thence 6 miles to Blandford, thence 18 to Salsebury and 8 mile to Newtontony which stands in y^{e} midst of y^{e} downs 8 mile from Andover a market town in Hampshire and y^{e} roade to London. It lyes 15 mile from Winchester–it is three mile from Amesbury and 2 mile more to Stoneage that stands on Salsebury plaine–eminent for many battles being faught there–this Stoneage is reckon'd one of the wonders of England how such prodigeous stone should be brought there, as no such Stone is seen in y^{e} Country nearer than 20 mile. They are placed on the side of a hill in a rude jregullar form–two stones stands up and one laid on their tops with morteses into each other and thus are severall in a round like a wall with spaces between, but some are fallen down, so spoyle the order or breach in the temple, as some think it was in the heathen tymes; others thinke it the Trophy of some victory wone by one Ambrosious, and thence the town by it has its name of. Amsebury. There is severall rows of lesser stones within the others set up in the same forme of 2 upright and one lies on the top like a gateway. How they were brought thither or whether they are a made stone is not resolved– they are very hard yet I have seen some of them scraped– the weather seemes not to penetrate them. To increase the wonder of the story is that none Can Count them twice alike–they stand confused and some single stones at a distance but I have told them often, and bring their number to 91. This Country is most Champion and open, pleasant for recreations–its husbandry is mostly Corn and sheep, the Downs though short grass y^{e} feed is sweet, producing the finest wooll and sweet meat though but small. (pp. 9-10)

We are also including narratives written by radical agitators. These reflect a personal interest which began with tramping artisans and moved onto working class autobiographies written by men who used the tramping system (see Southall 1991a, 1991b, 1996). Autobiographies generally described mobility in early adulthood, artisans being encouraged to travel to widen their experience after they completed their apprenticeships but before they married. Narratives written long after the events described tend to describe journeys very vaguely: ‘for sixteen months I tramped through the principal towns of Middlesex, Lancashire, and Yorkshire’ or, unhelpfully, ‘I need not follow my wanderings for some years, as my life at that time was of the ordinary kind.’ In contrast, the texts described here as “agitators’ narratives’ were written immediately after the events described, and themselves formed part of the agitation: they typically appeared in the relevant movement’s weekly or monthly newspapers. For now, we are concerned with three such narratives, although a number of others have been examined:

The movements of Feargus O’Connor, arguably the principal leader of the Chartists movement of the 1830s and 1840s, were extensively reported in the *Northern Star*, which he edited; it also sold signed engravings of him. Scattered through the *Star* are a number of shorter narratives. For example, the issue of January 19th 1839 contained a 4,500 word letter describing an eight day tour of Scotland.

Secondly, ‘The Life and Rambles of Henry Vincent, written by himself’, which appeared as a series of articles in *The Western Vindicator*, a newspaper owned and edited by Vincent. The articles total c. 30,000 words, appeared in issues 3 (9th March 1839) to 13 (18th May 1839) and cover the period between February 26th and May 10th 1839. Vincent, the ‘Demosthenes of the West’, was arguably the leading figure in the Chartist movement in the west of England and South Wales. The narrative ends with Vincent in Monmouth gaol, and his release was the supposed aim of the Newport uprising in October 1839.

Finally, a series of articles about and fairly clearly by Edwin Russell, an organiser employed by the National Agricultural Labourers’ Union in the 1870s. Between September 1872 and February 1873 a series of articles in the *Labourers’ Union* Chronicle, totalling about 10,000 words, described Russell’s travels, mainly in Herefordshire and Gloucestershire, in the autumn of 1872.

All narratives will reflect the biases of the author, and both Cobbett and Young had well-defined agendas, but the agitators’ narratives are distinguished by the purpose not of their writing but of the journeys themselves. On occasion they adopted the conventions of orthodox travel writing; here, for example is Vincent’s description of his journey to Ledbury:

> Took coach for Ledbury, in Herefordshire. The scenery along the road is very beautiful. On the right, within four miles of Ledbury, stands **Eastnor Castle**, the residence of Lord Somers. The castle is a fine building, situate on a piece of rising ground, surrounded by a sheet of water, and bounded on either side by extensive plantations. The entrance into Ledbury is exceedingly pleasing. To the right, amidst a profusion of trees and shrubs, is Ledbury church.

However, their basic goal was not to describe places but to change them, by creating new local branches of their organisations and establishing a sense of solidarity spanning places. This was very much a job of work: For March 27th 1839 Vincent reported:

A meeting was called in Newport for seven. Just before the meeting commenced the dark clouds rolled away — the rain ceased — and the silver moon looked smilingly upon us. We had above 4000 persons present. Edward Thomas took the chair. I delivered a thrilling oration to the people, which produced a pleasing effect. I felt in excellent spirits and tone notwithstanding my continued travelling and speaking; for I find, on calculating, *I have spoke about two hours a day for thirteen months, and travelled six thousand and seventy-one miles.* The Newport boys are advancing bravely.

Here is Russell:

I enrolled the 65 members' names on the Madley branch book, filled up a lot of members' cards, and put them in thorough good working order, and I think now they will be able to go on, but they do want such a lot of leading and guiding. I then walked on to Hereford, six or seven miles, in the pouring rain, it pouring all the way, and, as a consequence, I got wet though. I had a meeting planned for tonight at Wellington, but on account of the heavy rains which had fallen the rivers Lugg and Wye have overflowed their bank to such an extent that the roads are impassable, and I could not go to Wellington, but as I have got to arrange for a fortnight's meeting on the other side of the county, I shall not be able this afternoon.

4 Marking-Up Travel Narratives

Given these sources, how can we organise them to make them more accessible, and in particular to make it possible for, firstly, users to access information on specific places and, secondly, to generate maps of where travellers went? By now, this almost inevitably means some kind of mark-up system, i.e. adding tags to the text. Both common sense and the requirements of the program funding our work mean we should as far as possible follow existing standards rather than invent our own. The obvious starting point is the work of the Text Encoding Initiative (TEI). The TEI's *Guidelines for Electronic Text Encoding and Interchange* were first published in April 1994 and were initially based on Standard Generalized Markup Language (SGML). The current version, TEI P4, is also compatible with XML (Extensible Markup Language), a subset of SGML which is now far more widely used than its parent (Sperberg-McQueen and Burnard, 2002). The full *Guidelines* can be downloaded from:

```
http://www.tei-c.org
```

The *Guidelines* are just that, and any specific project will need a more specific mark-up system. Before discussing the facilities within the *Guidelines* for identifying locations and geographic terms, it should be noted that some key TEI-compatible mark-up schemes make no use of them:

- 'TEI Lite' is defined by the TEI itself as 'a manageable subset of the full TEI encoding scheme', but the only place-name tagging it provides for is place of publication.
- The American Memory DTD (Document Type Definition) was developed by the Library of Congress to mark up historical texts included within its very diverse American Memory system (http://lcweb2.loc.gov/ammem). This is of particular interest to us as our project is a major component within a consortium led by the British Library aiming to create a UK equivalent to American Memory. However, the AMS DTD again excludes geographical and place-name tags, and in fact the introduction to the DTD specifically states 'it is too expensive to identify geographic names'.
- The Minnesota "Women's Travel Writing, 1830-1930" project have developed a WTW DTD based on the TEI. They use a TEI-defined mechanism for adding interpretative information on four themes: ethnicity, gender marking, transportation and women's occupations. However, they do not tag place-names or geographic features. They note that issues have arisen in three main areas, gender, language and geography, and that 'our work [on geography] has not progressed as far as in the first two categories' (Remnek et al).

This cannot pretend to be a systematic survey of TEI-based DTDs, but it is still surprising that so little use is made of the available facilities for geographical description. One project that is using these features is the Perseus Project, primarily concerned with classical literature. Their work is described in Crane *et al*, 2001, but the remainder of this section discusses the facilities provided by the overall TEI Guidelines for geographical mark-up, and their suitability for travellers' tales.

Firstly, the Guideline allow simple tagging of place-names such as:

```
I went from <placeName>New
York</placeName> to
<placeName>Boston</placeName>.
```

Additional attributes can be added to simplify processing: 'key' provides an alternative identifier for the object being named, such as a database record key, while 'reg' can provide a regularized form of the name used so part of the earlier excerpt from Fiennes becomes:

```
Thence 6 miles to <placeName
reg=”Blandford Forum”> Blandford
</placeName>, thence 18 to <placeName
reg=”Salisbury”>Salsebury </placeName>.
```

Secondly, they allow a degree of hierarchy to be identified, which is often essential to eliminate ambiguity in common place-names, so an excerpt from Vincent would become:

```
Took coach for
<placeName>
  <settlement type="town">
Ledbury</settlement>, in <region
type="county">Herefordshire</region>
</placeName>.
```

Thirdly, they provide quite a sophisticated mechanism for identifying relative locations via ‘offsets’, so that Vincent’s statements that ‘On the right, within four miles of Ledbury, stands **Eastnor Castle**, the residence of Lord Somers’ becomes:

```
On the right, within <placeName>
   <distance>four miles</distance>
   <offset>of</offset>
   <settlement type="town">
      Ledbury</settlement>
   stands
   <name>Eastnor Castle</name>,
   </placeName>
the residence of Lord Somers’.
```

Finally, the <geog> tag can be used to identify types of geographical features, so that an excerpt from Russell would become:

```
… on account of the heavy rains which
had fallen the
<geogName>
   <geog>rivers</geog>
   <name>Lugg</name> and
   <name>Wye</name>
</geogName>
rivers Lugg and Wye
have overflowed their bank to such an
extent that the roads are impassable.
```

Overall, the TEI *Guidelines* try very hard to capture locational information embedded within otherwise unstructured text, although writing software to make systematic use of some of these tag structures might be complex. However, enabling conversion of locational information into spatial co-ordinates is only one part of the problem of marking up travel narratives, and arguably the least interesting. Equally important is understanding the role a reference to a place plays within the text. Consider this excerpt from Defoe:

> From Lemster it is ten miles to Hereford, the chief city, not of this county only, but of all the counties west of Severn: 'Tis a large and a populous city, and in the time of the late Rebellion, was very strong, and being well fortify’d, and as well defended, supported a tedious and very severe siege; for besides the Parliament's Forces, who could never reduce it, the Scots army was call’d to the work, who lay before it, 'till they laid above 4000 of their bones there, and at last, it was rather taken by the fate of the war, than by the attack of the besiegers.
>
> Not far from Lidbury, is Colwal; near which, upon the waste, as a countryman was digging a ditch about his cottage, he found a crown or a coronet of gold, with gems set deep in it. It was of a size large enough to be drawn over the arm, with the sleeve. The stones of it are said to have been so valuable' as to be sold by a jeweller for fifteen hundred pounds.
>
> It is truly an old, mean built, and very dirty city, lying low, and on the bank of Wye, which sometimes incommodes them very much, by the violent freshes that come down from the mountains of Wales; for all the rivers of this county, except the Driffin-Doe, come out of Wales.

Defoe is describing a journey from north to south from Shropshire through Herefordshire into Monmouth, but his text refers to many places off the line of his route: the river Severn, about 30 miles east of Hereford; the town of Ledbury, ten miles to the east and the village of Colwall, fifteen miles east; the mountains of Wales to the west, where the rivers come from. It should be immediately obvious that while a computer program could be written to map the locations named in the text, and draw a series of lines between them following their order of appearance, this would not be a remotely accurate map of his route. Further, the relationship between the geographical names and the text relating to them is complex. The anecdote about Colwall is both preceded and followed by information about the city of Hereford; and to some extent all of this text should be seen as information about the county of Herefordshire.

The problems become still larger when we examine the agitators’ narratives, where placed appear regularly in their speeches as well as in descriptions of their journeys, and geography provides an important part of their rhetoric. Consider this from a report of a speech by O’Connor in 1838:

> He could not conclude without expressing his delight at having thus perfected the great chain between London and Edinburgh and Glasgow. All the links were now perfect. London, Newcastle, Carlisle, Glasgow, and Edinburgh had now become forged as it were together, and although the wages of corruption were taken from the provinces to support the idle in the metropolis, yet a spirit was now growing up which nothing but justice could put down.

The same basic metaphor, but used less geographically, appears again here:

> You cannot move partially, because you are as one link in the great chain. (Cheers.) There is an end to sectional agitation; you are each answerable to the other for the manner in which you should handle this cause. (Cheers.) Society is as a great chain. On one end is a rotten link which represents the Whigs, and on the other a rotten link which represents the Tories (cheers and laughter); while you, the people, form the bulk of the chain, which is made stronger by the loss of the rotten links. (Cheers.)

Sometimes, of course, geographical names themselves are used as metaphors. Curiously, the TEI *Guidelines* include this example:

```
After spending some time in our
<placeName key="NY1">modern
<placeName key="BA1">
Babylon</placeName></placeName>,
<placeName key="NY1">New
York</placeName>,
I have proceeded to the
<placeName key="PH1">City of Brotherly
Love</placeName>
```

In this case New York and Babylon are named together in such a way that it is quite clear that the 'modern Babylon' refers to New York. However, the mark-up does not really make this relationship clear, and one wonders how the TEI would encode these excerpts from early nineteenth century autobiographers:

> A month's stay in modern Babylon was quite sufficient for me, and, gasping like a fish out of water, I set my face towards the open country.
>
> On entering the "Modern Babylon" I was almost confounded by its "huge uproar" together with the incessant bustle and hurry in which I found myself suddenly involved.
>
> Spent a night on the road then at about a quarter to eleven I entered the modern Babylon. The din and bustle and the never ceasing stream of vehicles of every description rattling backwards and forwards here completely bewildered me.

In all three cases, it is entirely clear from the context that 'Babylon' refers to London, and it is the use of one placename as code for another that gives these excerpts their interest.

The TEI *Guidelines* are designed to be extended, and it is not too hard to envisage a set of additional tags or keywords designed to bring out *how* a geographical name is being used within the text of a travel narrative. The main possibilities would seem to be:

- As the author's current location.
- As a reference back to a previous location.
- As a reference forward to a planned location.
- As a reference to a place which is not part of the author's route.
- As a metaphor for another place.

As important is a mechanism for associating sections of text with specific place-names, and this needs to embody some notion of hierarchy: even in the most conventional linear narrative, a series of descriptions of villages passed through also form a description of some larger area.

5 Conclusion

This paper is an initial discussion of a large topic, mainly concerned with practical examples and problems.

Although the descriptive gazetteers and travellers' accounts are very different kinds of material, the second far less structured than the first, in both cases analysis requires not simply the identification of place-names but an identification of their roles.

With the gazetteer entries, identifying the actual subject of each entry is trivial, as it always appears at the very beginnings of each entry. However, to discover the geographical location of each entry we need to extract additional 'containing' place-names. This is also relatively straightforward provided these names appear in our existing ontology/gazetteer. This is well illustrated by one of the few mistakes made by the existing parsers: the following entry is placed in Dorset, a mainland county, because the Channel Islands did not appear in the gazetteer:

- **Alderney,** one of the Channel Islands, about 7 miles from the coast of Normandy and 50.miles SE. from Portland Bill, Dorset; separated from Cape La Hague in France by the Race of Alderney.

Roles in the travellers' tales are clearly more diverse, and harder to define, but some notion of **subjects**, **containers** and **relatives** still seems appropriate, along with the rarer but more interesting **metaphors**. Work is clearly needed to extend and possibly to substantially change current mark-up systems to capture such information.

References

Gregory Crane, Clifford E. Wulfman, David A. Smith. 2001. Building a Hypertextual Digital Library in the Humanities: A Case Study on London. Paper presented at JCDL 2001, Roanoke, Virginia, June 2001. (www.perseus.tufts.edu/Articles/jcdl01.pdf)

Remnek, M.B., Skemp, B. and Wadsworth, S., ‘Technologizing Women's Travel Writing: Issues & Implications’, available on-line from WTW site.

Sperberg-McQueen, C.M.. and Burnard, L., eds. 2002. *TEI P4: Guidelines for Electronic Text Encoding and Interchange*. Text Encoding Initiative Consortium. XML Version: Oxford, Providence, Charlottesville, Bergen.

Humphrey Southall. 1991. The tramping artisan revisits: labour mobility and economic distress in early Victorian England *Economic History Review*, 2nd ser., 44: 272-96.

Humphrey Southall. 1991. Mobility, the Artisan Community, and Popular Politics in early nineteenth century England. In G. Kearns and C.W. Withers (eds.) *Urbanising Britain: class and community in the nineteenth century*. Cambridge University Press, Cambridge, UK.

Humphrey Southall. 1996. Agitate! Agitate! Organise!: Political travellers and the construction of a national politics, 1839-1880. *Transactions of the Institute of British Geographers*, NS, 21: 177-93.