The current issue and full text archive of this journal is available at
www.emeraldinsight.com/0001-253X.htm

Linking archival data to location: a case study at the UK National Archives

Paul Clough and Jiayu Tang
Information School, The University of Sheffield, Sheffield, UK, and
Mark M. Hall and Amy Warner
The National Archives, Kew Gardens, London, UK

Received 15 September 2010
Revised 18 November 2010
Accepted 15 January 2011

Abstract

Purpose – The National Archives (TNA) is the UK Government's official archive. It stores and maintains records spanning over a 1,000 years in both physical and digital form. Much of the information held by TNA includes references to place and frequently user queries to TNA's online catalogue involve searches for location. The purpose of this paper is to illustrate how TNA have extracted the geographic references in their historic data to improve access to the archives.

Design/methodology/approach – To be able to quickly enhance the existing archival data with geographic information, existing technologies from Natural Language Processing (NLP) and Geographical Information Retrieval (GIR) have been utilised and adapted to historical archives.

Findings – Enhancing the archival records with geographic information has enabled TNA to quickly develop a number of case studies highlighting how geographic information can improve access to large-scale archival collections. The use of existing methods from the GIR domain and technologies, such as OpenLayers, enabled one to quickly implement this process in a way that is easily transferable to other institutions.

Practical implications – The methods and technologies described in this paper can be adapted, by other archives, to similarly enhance access to their historic data. Also the data-sharing methods described can be used to enable the integration of knowledge held at different archival institutions.

Originality/value – Place is one of the core dimensions for TNA's archival data. Many of the records which are held make reference to place data (wills, legislation, court cases), and approximately one fifth of users' searches involve place names. However, there are still a number of open questions regarding the adaptation of existing GIR methods to the history domain. This paper presents an overview over available GIR methods and the challenges in applying them to historical data.

Keywords Historical periods, Archives management, Knowledge management, Geographic Information Systems

Paper type General review

1. Introduction

Geo-referencing, relating information to geographic location, is an important consideration for information access systems (Hill, 2006; Chapman and Wieczorek, 2006). This is due to the ubiquity of location (particularly place names) within much of the information we encounter on a regular basis. For example, many documents on the Web contain geographical identifiers, including place names (or toponyms), addresses (and address fragments), postal codes, and hyperlinks (Ding *et al.*, 2000; McCurley, 2001; Himmelstein, 2005). In addition, Petras (2004) showed that over half of a set of

Aslib Proceedings: New Information Perspectives
Vol. 63 No. 2/3, 2011
pp. 127-147

0001-253X
DOI 10.1108/00012531111135628

five million library catalogue records of the University of California contained one or more place-related subject headings or codes.

However, documents not only contain geo-references, but users of search systems also query based on location. For example, a study by Zhang *et al.* (2006) found that 12.7 per cent of queries submitted to a web search engine contained a place name; Sanderson and Han (2007) also showed that queries for city names were repeated more frequently than for other names (e.g. country and state). This information can be exploited and used to provide spatial awareness to information systems (see, for example, Buckland *et al.*, 2007; Purves *et al.*, 2007). Zong *et al.* (2005) discuss how the ability to perform query by location can be an important and useful addition to a digital library and Buckland *et al.* (2007, p. 376) concur with this in saying "libraries have a broad need to support geographic search".

This paper presents an overview and case study of the challenges and goals of linking archival data to location. Managed archives, particularly those operating on a large scale, such as national archives, often hold large and extremely diverse collections of historic data, making it difficult for users to successfully retrieve the historic documents they are interested in. Linking historical documents and records to place allows synthesised, seamless access across heterogeneous archival data sets and facilitates novel ways of being able to search and browse large-scale archival collections. This paper describes the initiation of a project to explore place-based access to historical data at The National Archives[1] (TNA), the UK Government's official archive. Section 2 discusses past work in exploiting location for enhancing information access to libraries and archives; section 3 introduces the UK National Archives, the data and motivations for the proposed project on space exploration; section 4 describes areas involved in geo-referencing information; section 5 outlines our planned project work and section 6 concludes the paper.

2. Related work

Geography provides an important facet for information seeking in many contexts, including historical, cultural, libraries and archives. Several studies have examined the nature of search queries in the cultural heritage field, both for general information (Cunningham *et al.*, 2004) and for images (Pask, 2005; Choi and Rasmussen, 2003; Collins, 1998; Chen, 2001). These all had similar findings: namely that people, *places,* time periods, and subjects were popular topics of search in this domain. This is also true of library users (Buckland *et al.*, 2007; Zong *et al.*, 2005) and utilising place (and time) for information access has been investigated in a range of past projects.

Probably one of the most widely cited research projects that made use of geo-referencing was the Alexandria Digital Library (ADL) project (Frew *et al.*, 1998; Goodchild, 2004) which aimed "to develop a user-friendly digital library system that provides a comprehensive range of services to collections of maps, images and spatially-referenced information".

Moritz (1999) describes the geo-referencing of narrative descriptions of specimens in natural history museums written by collectors in the field. Reid *et al.* (2004) discuss some of the projects funded by EDINA, a National Data Centre in Edinburgh, including Go-Geo! and Geo-X-Walk. These projects facilitate the discovery of geo-referenced information from distributed collections in the UK.

Buckland and Lancaster (2004) describe combining place, time and topic in the Electronic Cultural Atlas Initiative[2], designed to foster a research community interested in geo-temporally encoded data and create a catalogue of geo-referenced online resources. As a part of this initiative, Buckland *et al.* (2007) overview the Going Places in the Catalog project, an initiative to offer seamless searching of educational and scholarly numeric and textual resources. This includes extracting geo-spatial information from library catalog records to enhance the user's search experience.

Johnson (2004, 2005) describes the indexing and delivery of historical maps online from the National Library of Australia using the TimeMap[3] tool (a mapping Java applet which generates complete interactive maps with a few simple lines of html code). With a similar goal, Martins *et al.* (2007) describe the EU co-funded DIGMAP project aimed at making collections of digitised historical maps spatially aware, through the use of text mining techniques and Geographic Information Retrieval (GIR). This project deals specifically with some of the issues involved in using digitised versions of historical documents (e.g. old font styles, incorrectly OCR'd words, etc.).

Mostern and Johnson (2008) describe the construction of a historical event gazetteer (these are records of historical events that describe the existence of named places) and the use of spatio-temporal visualisation to view events and relationships from the Heurist[4] collaborative database. Clough *et al.* (2008, 2009) describe a study in which semantic access to cultural heritage information from Tate Online, a large online art collection, is investigated. This includes the development of a prototype system to demonstrate advanced search and browse functionalities, including faceted browsing, timelines and maps, based on semantic enrichment of data from Tate Online.

Finally, the Vision of Britain[5] web site brings together census data, election results, historical maps, gazetteers and travel writing between 1901 and 2001 into a single point of access. This is part of the Great Britain Historical Geographical Information System (GBHGIS) project managed by the University of Portsmouth (Gregory and Southall, 2003).

3. The National Archives (TNA)

3.1 The role of TNA

The National Archives (TNA) is the UK Government's official archive. The documents that it holds reflect almost 1,000 years of history, with records ranging from parchment and paper scrolls through to digital files and archived web sites[6]. The National Archives also provides advice and guidance to the public and private sectors about caring for archives, and records the location of archival collections throughout the UK. The National Archives brings together the Public Record Office (founded in 1838), Historical Manuscripts Commission (founded 1869), the Office for Public Sector Information, and Her Majesty's Stationery Office (founded 1786). Through HMSO, it publishes all UK legislation, while OPSI advises on and encourages the re-use of public sector information.

3.2 Datasets held by TNA

The National Archives' history, and its central role in information management, means that the datasets it holds are extensive, varied and rich. These fall into four main groups:

(1) *Original documents*, in both paper and electronic format (found in Electronic Records Online – ERO) and digitised copies of original documents made available online (through Documents Online). However, only a very small subset of the original records have been scanned and digitised, and only for an even smaller proportion has the original text been extracted from the scanned documents. In light of this the majority of the work described in this paper is applied to the data stored in Catalogues, which is described next.

(2) *Catalogues* describe the content of The National Archives collections and record the location of archival records held throughout the UK. These range from the general, in particular TNA's Catalogue of its collections and the National Register of Archives, to the specific, with the E179 Taxation database.

(3) *Published information*, in particular the *London Gazette*, official newspaper of record and the Statute Law Database. These data are currently maintained and hosted separately[7].

(4) *A wide variety of other datasets*, including the ARCHON Directory of archive contact details, and Your Archives, a wiki, which allows users to contribute content about archival records.

Most of the data are provided as unstructured free-text and methods for extracting place information from free-text will be described later in this paper (section 4). In some cases, geographic data have been at least partially structured, for example the digital Domesday data explicitly provided mark-up for place and historical name information, and we will present an approach that can take advantage of this structuring to improve the place information extraction methods (section 5.1.2). Geographic information can be found throughout TNA's databases, varying from the postal addresses of every archive in the UK (found in the ARCHON directory); a comprehensive directory of medieval places, with related Ordnance Survey[8] (OS) grid references and alternative names (drawn from the E179 tax database); and a list of parishes which relate to the National Farm Survey carried out during World War Two (found in the Catalogue, reference MAF 32).

3.3 Motivations to support geographical search

Allowing users to access TNA data by reference to location is high priority, because of the frequency with which users issue search requests that involve place names. To quantify this, around 3,000 queries were extracted for January and July 2009 from transaction logs for the online catalogue. The queries were manually analysed and search terms classified as belonging to one or more pre-defined categories, including named entities (person, place, organisation, date) and query intent (navigational or informational). The analysis showed that queries involving person names were the most frequent type of search, followed by those including a place name reference (see examples of place names in Table I). Queries involving a place name were found to account for approximately 20 per cent of all queries, which is substantially higher than the 12 per cent of place-related queries reported for general web search (see, e.g. Zhang *et al.*, 2006).

Enabling TNA's data to be navigated and explored by place would provide an effective means of helping at least a quarter of TNA's users to find a way through the

January 2009		July 2009	
Place name	Frequency	Place name	Frequency
Palestine	195	Aston	219
Pannal	159	Fermanagh	196
Knaresborough Forest	129	Wigley	186
Lockton	97	Palestine	154
Ayelsford, Kent	96	Gibraltar	145
Churton	94	Donegal	132
Monks Eleigh	86	Ninfield	132
Wicklow	82	Yarmouth	114
Clapham Common	75	India	111
Tuxford	69	Gold Coast	111

Table I.
Top 10 most frequent place names in January 2009 and July 2009

mass of information held by the archives – currently about 32 million records in the online databases and documents. In particular this would allow:

- Visualisation of TNA data through geographical interfaces.
- Search by explicit place name, without the results being mixed with people's names that are also place-names. An explicit place-based search for "Wellington" would not return documents relating to the "Duke of Wellington", unless they were related to the actual place.
- Recommendation of other documents of potential interest, based on spatial proximity (e.g. present to the user documents which contain references to places near to a given location).

Exploiting place information in TNA databases, using the methods described later (section 4), could unlock their potential, allowing the information to be interrogated in a way, which is not possible with the current, independent databases. Complementary information in different databases could be highlighted, enhancing the richness of the data. For example, researchers interested in the history of Godalming Parish[9] could, after searching for "Godalming", be presented with links to:

- Tithe records in *The Catalogue*.
- Details on a seventeenth century Vicar of Godalming, who failed to pay his tax, found on the E179 database.
- The records of Godalming parish, held at Surrey History Centre, recorded on the National Register of Archives.
- The location of records relating to the manor of Godalming, found on the Manorial Documents Register.

The user's resource discovery experience could potentially be enriched further by links to related external, non-archival resources, for example:

- The page on Godalming on the Exploring Surrey's past web site[10].
- The entry for Godalming in the Victoria County History, found on the British History Online web site[11].
- The Wikipedia page on Godalming[12].

The fragmented nature of TNA's databases can make resource discovery challenging for users, and the vision of a single catalogue interface is something which is being actively pursued and the identification and exploitation of place data in TNA's resources will contribute to this.

4. Geo-referencing

Relating information to geographic location involves linking between informal means of referring to locations using place names (or toponyms) and formal representations based on spatial referencing systems (e.g. longitude and latitude or the British National Grid system). Much of the geographical information contained in libraries and archives is based on place names, however these must be geospatially referenced (grounded) if they are to be used to their greatest potential for enriching information access and knowledge discovery. Locations are also identifiable in other forms than just place names, including addresses, address fragments, postcodes, phone numbers, URLs and IP addresses.

Gazetteers are typically used to provide the link between informal (place names) and formal (corresponding spatial references) representations (section 4.1 presents some example gazetteers). A key task for providing geospatial information access is the recognition (or extraction) of potential geo-references and their geospatial referencing (discussed further in section 4.2). Geographic Information Retrieval (GIR) deals with the indexing of information to enable searching over it to find information that is potentially relevant to some information need (discussed in section 4.3).

4.1 Sources of geographical knowledge

To explicitly geo-reference information relies on having access to existing geographical knowledge. The most commonly used resource is the gazetteer: at minimum a list of place names, feature types and spatial positions (Hill, 2006; Mostern and Johnson, 2008). Sources of gazetteers commonly include (Goodchild and Hill, 2008, p. 1039):

- Gazetteers of "official" toponyms (e.g. the Ordnance Survey 1:50,000 Scale Gazetteer).
- Indexes accompanying published atlases (e.g. Multimap.com).
- Place identifier tables accompanying GIS datasets.
- Place authority files (e.g. Vision of Britain) and rules (e.g. NCA Rules for the Construction of Personal, Place and Corporate Names[13]) used for cataloguing and indexing.
- Historical printed gazetteers and encyclopedias (e.g. Gazetteer of Great British Place Names).
- General online resources (e.g. Wikipedia.com).

Researchers have investigated how to combine multiple sources to create more comprehensive resources (see, e.g. Manguinhas *et al.*, 2008; Leidner, 2008). Example sources which could be used and are being considered within TNA's programme of work on location-based access to the archives include the following[14] (these are predominantly UK-biased):

- *Getty Thesaurus of Geographic Names*[15]: contains more than 1 million names and other information about places across all continents (including political and historical places). The locations are organized into a hierarchical structure (e.g. World > Europe > United Kingdom > Scotland) which can be useful for computational text analysis.
- *Ordnance Survey 1:50,000 Scale Gazetteer (OS50k)*: contains 260,000 names in Great Britain from the current and previous years' 1:50,000 Scale gazetteers. Locations are represented by points (National Grid squares) with additional feature type.
- *Ordnance Survey Code-Point*: provides a National Grid reference for each unit postcode in Great Britain. The postcode data are sourced from, among other resources, Address Point, which contains 26 million addresses recorded in the Royal Mail Postcode Address List (PAF).
- *GeoNames*[16]: is a (growing) database of over 8 million place names (and postcodes) from countries worldwide. Locations are represented by points and contain a range of feature types (e.g. country, region, city, body of water, etc.). The database can be downloaded and used free of charge. The data are based on several sources including the GEOnet Names Server, the US Geological Survey Geographic Names Information System and Wikipedia.
- *Gazetteer of British Place Names*[17]: contains over 50,000 entries (points – National Grid references) referring to places in Great Britain. Each place name is related to its historic county (administrative areas) and variants of place names are also included.
- *UK Placename Finder*[18]: is a database of more than 31,000 UK place names (available for purchase on CD-ROM). Places are referenced to the UK National Grid.
- *Seamless Administrative Boundaries of Europe*: is a pan-European dataset of administrative units compiled from national contributions (containing the geometry and semantics of the administrative hierarchies of 30 European countries). For the UK, the dataset consists of around 22,000 place names.
- *Alexandria Gazetteer*: contains around 5.9 million place names, with feature type information for all entries. The gazetteer is mainly US-based (US place names from the US Geological Survey's GNIS database are used), but also contains around 4 million non-US place names from the GEOnet Names Server.

In addition to these resources, geographic data can be obtained or extracted from various online resources and databases and used to supplement the data found in existing gazetteers. For example, the following resources and services may provide useful data (see also Chen and Nottveit, 2010, for a review of available resources in the development of an application for searching and browsing historical photos):

- *Wikipedia*: contains a wealth of place names, which can be automatically extracted and compiled into useful gazetteer resources (see, e.g. Overell and Rüger, 2006).
- *Geo-X-Walk*[19]: is available for use through EDINA and provides APIs for UK name, postcode and identifier lookup.

- *Geograph*[20]: is a project, which aims to collect photographs and information for every square km of Great Britain. The project provides an Application Programming Interface (API), which can be used to interact with the data. Currently there are 1,359,305 images for Great Britain, which include place name metadata and a UK Grid Reference.
- *Yahoo Geocoding API*[21]: is a freely available web service (limited to 5,000 requests per day) that provides coordinates for a given address or place name. (The GeoNames service uses this as a backup resource if the place name cannot be found in its own database.)

A more sophisticated form of geographic knowledge representation is an ontology, which typically organises locations into hierarchical structures and provides topological relations between locations (see, e.g. Fu *et al.*, 2005). Recent work has focused on areas such as spatio-temporal modelling (Mostern and Johnson, 2008), interoperability (Janowicz and Kessler, 2008) and automatic gazetteer construction (Nadeau *et al.*, 2006; Popescu *et al.*, 2008).

4.2 Identifying and resolving geo-references

A basic task in geo-referencing information is to identify candidate geo-references. For place names (or toponyms), this is commonly referred to as geo-parsing (Larson, 1996) or toponym recognition (Leidner, 2008) and can be likened to the Information Extraction task of Named Entity Recognition and Classification (or NERC): the process of assigning every word or group of words to a set of pre-defined categories (including "not an entity"). These categories commonly include entities such as location, person and organization.

Most NER systems consist of at least three basic components:

(1) a tokeniser (software that splits text into sentences and sentences into individual words);

(2) gazetteer lists (lists of place names, person names, company names, etc.); and

(3) a Named Entity (NE) grammar (syntactic rules that take into account features that describe the surrounding context).

Rules for NER can be generated entirely by hand (knowledge-based) or automatically, using machine learning (or statistical) techniques on previously classified texts (see, for example, Zhou and Su, 2002; Curran and Clark, 2003). The most common approach is to use gazetteers together with syntactic rules. For example, Curran and Clark (2003) use a Maximum Entropy model to perform NER using features such as Part-Of-Speech tags, a history of preceding and following NEs, orthographic information and gazetteers. Their approach was used on historical texts in the Geo-X-Walk project (Nissim *et al.*, 2004).

Once recognised, a candidate place name must be resolved, i.e. given a spatial reference (Larson (1996) calls this geo-coding; Leidner (2008) refers to this as toponym resolution). This may involve disambiguating the place. There are two main ambiguities in geo-references:

(1) *Referent* ambiguity.

(2) *Reference* ambiguity.

The former occurs when the same place name is used to refer to more than one location (e.g. "Chapeltown" refers to a location in South Yorkshire (UK), Lancashire (UK), Kent County (USA) and Panola County (USA)). This also includes a geographical reference being used for other entities (e.g. the name of a person or company), which is called referent class ambiguity. (Amitay *et al.* (2004) separate ambiguity into geo/non-geo and geo/geo.)

Reference ambiguity occurs when the same location can have more than one name, for example, due to the historical deviation of a location name over time (Smith and Mann, 2003), transliteration (Kwok and Deng, 2003) and implicit formats and character encoding (Axelrod, 2003). An extensive summary can be found in Leidner (2008).

4.3 Geographical IR (GIR)

Geographic Information Retrieval (GIR) is concerned with finding documents relevant to queries, which include some geographical context (Larson, 1996). Information is assigned one or more footprints and documents are retrieved and ranked according to thematic and spatial relevance (results can also be visualised using maps).

Research in this field has addressed a wide range of relevant areas such as the building of geographical ontologies, spatial indexing and storage of documents, geographical relevance ranking, the extraction and resolution of geographical references, determining the geographical scope (or focus) of web documents, user interfaces and visualization, and methods for the formulation of spatial queries and interaction with the results of geographic search (see, e.g. Hill, 2006, pp. 185-214; Purves *et al.*, 2007).

5. Supporting geographical searches at TNA

The geo-spatial work at TNA is still at an early stage. A number of challenges and potential approaches to handling historic, geographic data have been identified. These will now be discussed and then three case studies will be presented that illustrate the benefits of adding explicit spatial information to historic data.

5.1 Challenges

In section 3, an overview of the general challenges in geo-referencing documents was presented and these also apply to the archival domain. When dealing with historical data, as commonly found in archives, there are a number of domain-specific challenges, which need to be overcome, which include the following.

5.1.1 Integrating space and time. For historic documents, when the document was created, or what the document talks about, is central. This also translates into the geographic data: new places come into existence (e.g. "South Yorkshire" in the UK was created on 1 April 1974 as a result of the Local Government Act of 1972); places cease to exist or people simply do not refer to them anymore; places are referred to by different names and their spelling may change over time. This creates an interesting problem because gazetteers tend to store only the current official place name (see, e.g. Buckland and Lancaster, 2004; Mostern and Johnson, 2008). To support this temporal aspect, TNA's geo-data storage is set up to annotate all spatial data with temporal annotation as shown in Figure 1. When documents are geocoded, each piece of information that is added to the geo-data storage is annotated with the date of the document from which it was extracted. Therefore, if a place in *Domesday Book* is geocoded, then the old

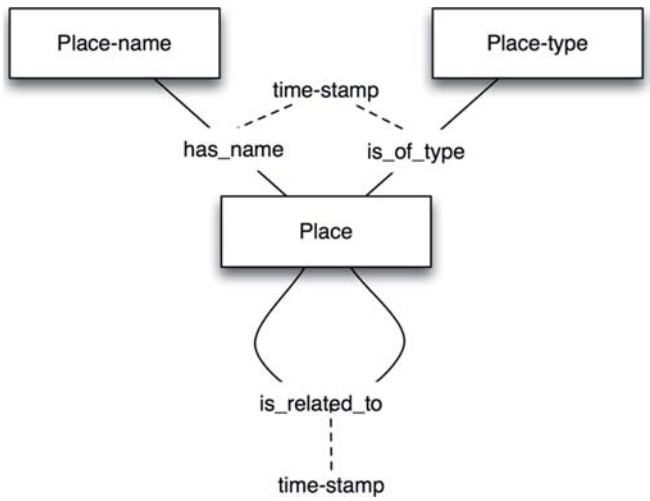


Figure 1.
The basic structure of the TNA gazetteer, with the three main objects "Place, Place name, and "Place type" and all relations between these objects annotated with a time-stamp

spellings of the place name are all added with a time-stamp of 1086, in addition to the current name, which is time-stamped as 2010. This detailed annotation with temporal information is also performed for all spatial relations and feature types.

5.1.2 Historical places. The problem with detecting historical place names is that inevitably spellings will have changed over time, for example due to changes in orthography (Crane, 2004). However, gazetteers mainly hold the current, official spelling. To overcome this, we have taken an iterative approach to derive variant spellings. Initially only documents where the place names use modern spellings are geocoded. From these documents, it is possible to learn potential historical spelling variants. These spellings are stored in the TNA gazetteer and are used in a second iteration to geocode those documents where only a historic spelling is used and potentially find further spellings. This process is repeated until no further documents can be geocoded and no further spellings are learned. To make the maximum use of this information the initial geo-coding will focus on record series that have both historic and modern spellings, allowing us to quickly build up a set of historic to modern name mappings.

In addition to this iterative process, a number of simple rules have been developed to increase the amount of overlap between the spelling in the documents and in the gazetteers. Some gazetteers use a set of abbreviations that are usually not found in the documents themselves. For example, in the OS 50K gazetteer all farm names have the abbreviation "Fm" added to their names (e.g. "Abbey Fm"). If in the document the place name is spelled out as "Abbey Farm" then no match will be found, if an exact match technique is used to identify gazetteer entries. Similarly the OS50K gazetteer adds the suffix "Manor" to all manor houses, while in historic documents they are usually just described using their name. To support these differences a number of transformation rules have been defined, some specific to gazetteers or historic data sources, some generic that create alternate spellings of the name to increase the number of places that can be detected. If such transformation rules lead to the place name being

geo-referenced, then the geo-referenced place is annotated with the knowledge that the name had to be transformed to create a match.

5.1.3 Heterogeneous data-sources. TNA's various archives and databases have developed organically over time and, thus, they are very different in structure, the geographic information is represented differently in each database, and the type of geographic information also varies from database to database. To handle this challenge a series of custom-built "wrappers" are needed that know, for each database, where and how to extract the spatial information. The knowledge about how the data are structured is encoded in a series of hints that are passed to the NER and geo-referencing components. For example, in the Domesday Book data the place names are already marked-up in the source data. The wrapper will, thus, add a hint that no NER is required and will also hint the geo-referencing component that the data are UK only, which simplifies the disambiguation process. On the other hand, the Dixon-Scott photograph collection contains image captions from which the place names have to be first extracted via NER. The advantage of using such a structure is that the NER and geo-referencing components can be very generic, but at the same time can adapt to the specific aspects of each data-source.

5.1.4 Ambiguity. As mentioned earlier, the ambiguity of place names provides a major challenge when geo-referencing both historic and modern geographic data. When processing archival data, the primary issue is with referent ambiguity: the same name is used to refer to different places. When geo-referencing an identified place name, the first step is to find all possible places in the gazetteers that have the same name, which are called potential groundings of the place name. From these potential groundings, one location has to be chosen in a process called disambiguation. A number of simple heuristics for disambiguation have been adapted to the history domain from existing work (Clough, 2005):

- Initial disambiguation is performed using information about the hierarchical organisation of locations. Most gazetteers contain information about the administrative regions a place is contained in (e.g. "Sheffield" is a city contained in "South Yorkshire"). If similar information is available from the document, then the amount of overlap between the hierarchy of the place name to geo-reference and the hierarchies of the potential groundings can be used to rank the potential groundings and then choose the highest-ranked potential grounding as the final grounding. The problem with this approach is that the documents describe historic data and thus the containment hierarchies extracted are also historic. These might no longer overlap with the current hierarchy (Middlesex no longer exists, the boundary between Oxfordshire and Hertfordshire has moved) and thus the ranking would either not find any overlap or be unable to disambiguate correctly. To counter this, a set of qualitative mappings describing which administrative regions have been renamed into which other regions or where the boundaries have changed has been built-up. If no direct hierarchy match can be found, then the transformation mappings are applied to see if this adds the necessary knowledge to be able to successfully disambiguate between the potential groundings. While currently this has focused on the changes of UK county boundaries, the system is set up to be extended to country boundary changes such as "Ceylon" becoming "Sri Lanka".

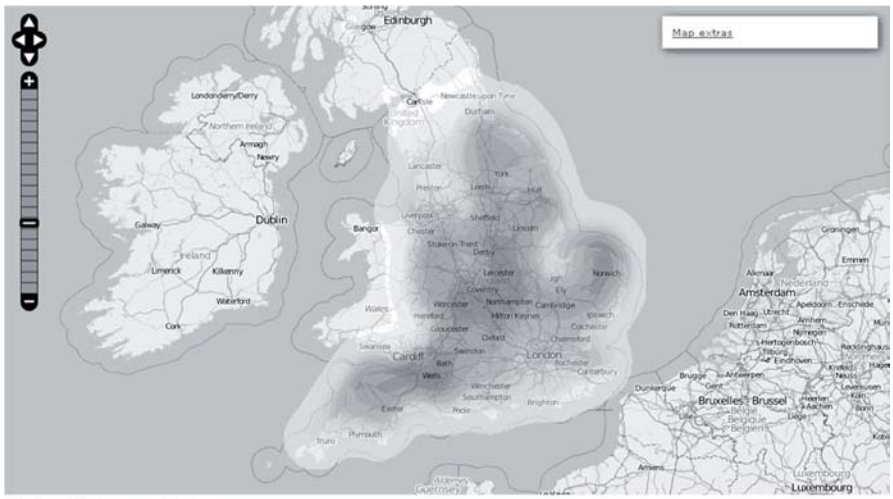
- To the basic hierarchical disambiguation approach, heuristics based on the place type are added. If both the place name to geocode and the potential groundings have type information then those groundings where the place-type information does not match can be discarded. If the place name to be geocoded does not have type information, then the potential groundings' type information can still be used if the data source wrapper provides hints as to the preferred place-type ranking. For example, in data from *Domesday Book*, the wrapper specifies that groundings are to be ranked in the following order: Populated places, Farms, Historical sites, everything else. This makes it easier to handle the case where there is a populated place and a hill of the same name.

Using these two metrics on a very clean data-source, such as *Domesday Book*, where all places use modern spellings and pre-1970 county names are used for hierarchy information, it is possible to automatically geo-reference approximately 81 per cent of the data. The remaining 19 per cent consist of places that do not appear at all in the gazetteers, place names where spelling mistakes could not be corrected automatically, and thus the places were not found in the gazetteer, and places that could not be disambiguated, usually because there are two or more places with the same name in the same county, which the algorithms simply cannot disambiguate. While some of the un-referenced places could potentially be geo-referenced automatically if the algorithms were further tuned, most of them are simply out of reach for automatic approaches.

5.1.5 Scalability. Scalability is a problem that partially affects geocoding historic records. Geocoding a few million documents is a slow process; however the historic datasets do not change very frequently and thus speed is not a concern. Nevertheless, on the front-end scalability is a major concern. There is a limit to how much information can be displayed concurrently on a map. Two approaches to this have been tested. The first is clustering points on the client side, where those points that lie very close together are combined into a single point. This reduces the number of points displayed on the map; however when the user clicks on one of the clusters it is necessary to display some kind of interface where the user can then decide which of the points to actually see. The second approach is to create a different representation on the server side when there is too much information to display as points. To enable this "heatmaps" are employed, which indicate where there is more information and where less (Hill, 2006, p. 211). Figure 2 shows a heatmap of place names mentioned in *Domesday Book*. The user can then use this to guide them to parts of the map where there is more information. When the number of points is low enough to display on the map, the heatmaps are switched off and the point-based display switched on (see Figure 3).

5.2 Case studies

In the process of investigating the use of spatial information to access data from TNA a number of applications have been developed. We have selected three of these as case studies to demonstrate some of the potential benefits of exploiting geo-referenced data within the national archives. Some of these are available via TNA's experimental site called The National Archives Labs[22].



Note: Darker = denser

Figure 2.
Heatmap showing the density of places mentioned in *Domesday Book* (darker = denser)

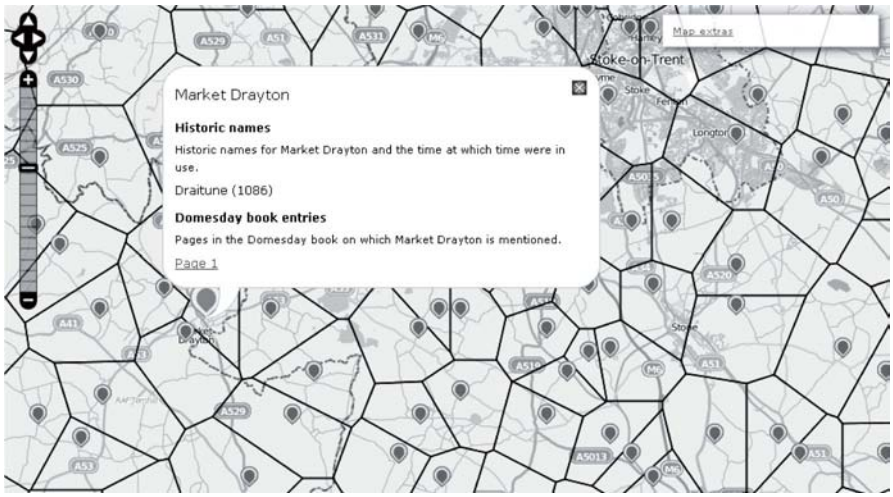


Figure 3.
Point-based display showing individual entries from *Domesday Book*

5.2.1 Geographical search and browse. After geo-referencing the data, an obvious application is to exploit the geo-referencing for information access by providing online users with functionalities to search and navigate the archives through place. A "light-weight" prototype for this kind of functionality was developed using existing tools including the Edina Unlock service (http://unlock.edina.ac.uk) for extracting place names and assigning geographic coordinates based on resources including the Ordnance Survey 50k gazetteer and the publicly-accessible Geonames gazetteer.

The underlying search engine is built on Apache Solr[23], which is "an open source enterprise search server based on the Java search library Lucene[24]". Apache Solr provides a HTTP interface for issuing queries and supports different types of response,

formats such as XML and JSON. Geographic search is realised by LocalLucene/LocalSolr. These are geographical extensions to Lucene and Solr. The LocalLucene search engine augments Lucene with the support of a single geographical query: search for documents within a certain distance from X (expressed as a point using longitude and latitude). For example, it is possible to formulate the following kind of search: "find documents about X within ten miles from [51.507778, −0.128056]". In order to avoid redundant distance calculations, LocalLucene uses a "Cartesian Grid" approach to indexing documents spatially[25].

An interactive map has been used based on OpenLayers[26], an open source JavaScript library for displaying map information from a variety of sources. Different kinds of information are organised as layers, for example in our prototype system there are two layers: the map layer and result markers layer. We have chosen OpenStreetMap[27] as the source of map to be used in OpenLayers. The map supports basic interactions such as zoom in/out and pan.

The prototype enables two main forms of interaction. First the user can simply browse over the archives, the prototype displaying the archival records available for the area visible on the map. The second mode of interaction combines the map browse with the ability to search by keyword. The user can enter a keyword ("cricket" in the example in Figure 4) and the database is searched for documents with that keyword, but the results are also restricted to the area that is visible in the map (see Figure 4). As the user browses around the map the search results are continuously updated so that the user only sees results that relate to the area currently visible in the map. This ability to navigate along both the content dimension, using the search functionality, and the geographic dimension, by interacting with the map, provides a very powerful method for interrogating large archival datasets.

The system also enabled us to investigate methods for dealing with the large amount of data available in archives. When large numbers of documents are displayed

Figure 4.
Example result from the prototype user interface

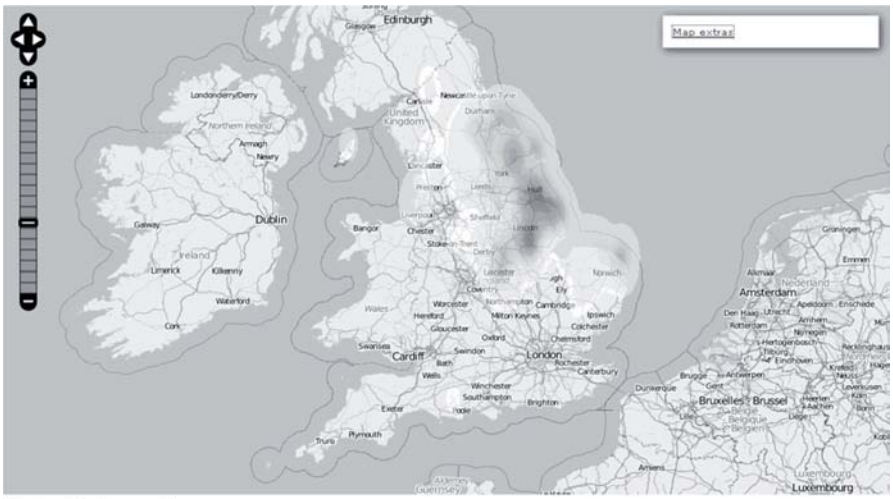
on a map it is highly likely that some of the points representing the documents' locations will overlap. To avoid this, the automatic clustering functionality available in OpenLayers is used to group together documents that are "too close" to each other. While the individual documents are represented as a coloured disc, the markers for the document clusters are further annotated with the number of documents that have been clustered. This information enables the user to quickly identify for which areas there is more data and they can then zoom into that area. As the user zooms in there will be fewer documents that lie within the area shown on the map and these will also be spread out more across the map. The clusters will thus break apart, first into smaller clusters and then, when the user has zoomed in enough, into individual documents. This enables us to provide access to a large amount of information via the map, without overloading the map and the user.

This initial prototype is currently not available via the TNA labs site, as work is underway to integrate it more closely with the existing systems at TNA and to provide advanced functionality such as providing search assistance to the user by using a gazetteer or geographical ontology to suggest locations spatially related to the current search position (Purves *et al.*, 2007).

5.2.2 Bespoke series-specific interface. While the generic search/browse is very useful for searching and browsing in large data collections, it is often useful to provide bespoke interfaces for specific historical data collections. One such collection is *Domesday Book*. Geo-referencing places mentioned in it provides a new way into the data: users can freely browse around the *Domesday* entries of the area they are interested in and no longer be restricted to viewing a single record at a time. The addition of spatial information also enables the display of analysis results. For example, when the user zooms out, then the individual points can be replaced by a heatmap to illustrate the density of *Domesday* places (see Figure 3) to allow the user to focus on which part of England they want to zoom into to see the detailed data. The same heatmap technique is also used to show which of those areas of England where many place names are of Viking origin (see Figure 5), identified by common suffixes such as "-thorpe" or "-by". While it is impossible to build such bespoke interfaces for each record series, it does provide a method for highlighting the information contained in certain record series that are especially spatial.

5.3.3 Developing a shared historical gazetteer. The push towards more open data in the UK Government also applies to TNA. To underpin geographical search and browse across the entire archives a common geographical reference is required. Therefore as a part of the Space Exploration project a gazetteer containing historical place names is being developed that can be shared across applications and made freely available. To support this, a Linked Data representation of the gazetteer has been created. Linked Data are the principle that underlies the idea of the Semantic Web and at its most basic is a method for describing information in such a way that it can be parsed and reasoned with automatically.

The basic element of the Linked Data representation is the Uniform Resource Identifier (URI) that identifies a specific resource. For example, "http://labs.nationalarchives.gov.uk/linked-gazetteer/place/id/10" identifies a resource about a place called "Loughton" in TNA's gazetteer. The document available via that URI then defines the properties of the resource, such as its name (foaf:name), location (tna:geometry), and also that it is the same as (owl:sameAs) the resource at "http://data.



Note: Darker = denser

Figure 5.
Heatmap showing the location of *Domesday Book* places that are of Viking origin (darker = denser)

ordnancesurvey.co.uk/doc/50kGazetteer/150033". Each property consists of a namespace (foaf, tna, owl) and the property name. Using the namespaces it is possible to use properties defined by other organisations, which provides a shared vocabulary and a shared interpretation of what the property means. Figure 6 shows an example of a gazetteer entry based on Linked Data that is expressed using a standardised way of expressing entities – the Resource Description Language (RDF). In the example which follows, the property owl:sameAs is used to link to a resource owned by a different organisation and it is this ability to link between resources owned by different organisations that makes Linked Data so powerful, as an automatic system can gather the individual resources together and automatically draw conclusions that could not be drawn from any of the individual resources.

The principle of using the Semantic Web to integrate cultural heritage collections, has been most recently expounded by Jankowski *et al.* (2009) who discusses the benefits available to cultural heritage institutions if they design their applications according to the Linked Data principles. Hardman *et al.* (2009) also identified benefits to cultural heritage "... information tasks that can be supported using linked data by making non-obvious connections among related pieces of information explicit, such as exploratory tasks or topic search". Such work explores technologies to satisfy user needs, especially those of cultural heritage experts whose search tasks often involve

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:foaf="http://xmlns.com/foaf/0.1/"
    xmlns:tna="http://labs.nationalarchives.gov.uk/linked-gazetteer/"
    xmlns:owl="http://www.w3.org/2002/07/owl#">
      <rdf:Description rdf:about=" http://labs.nationalarchives.gov.uk/linked-
      gazetteer/place/id/10">
            <foaf:name>Loughton</foaf:name>
            <tna:geometry>POINT (7.3333 51.648333)</tna:geometry>
            <owl:sameAs
            rdf:resource="http://data.ordnancesurvey.co.uk/doc/50kGazetteer/150033
            "/>
      </rdf:Description>
</rdf:RDF>
```

Figure 6.
Example RDF document describing the place "Loughton" in TNA's linked-data historic gazetteer

relatively complex information gathering and use, combining, as mentioned previously, results from multiple sources (Amin *et al.*, 2008).

6. Conclusions

In this paper we have discussed the role of geo-referencing within archives. In particular, we have discussed a programme of work at The National Archives, the UK Government's official archive, on linking archival data to location. The diversity and vastness of content at TNA currently presents a real challenge for the archival staff and end-users of the archive, such as scholars and the general public. By exploiting geo-references from the content, material can be linked and aggregated to provide enhanced forms of information access and knowledge discovery across discreet data sources. It would also highlight complementary information in TNA's databases, and in doing so enhance the richness of the data. Place data are also one of the most popular topics for searching the archives from TNA's web site and Geographical Information Retrieval (GIR) technologies could help improve the current search paradigms. By applying (and exploring) existing technologies from areas such as GIR and text mining, we can geo-enable the archive and prepare it for future access.

Notes

1. www.nationalarchives.gov.uk/
2. www.ecai.org/
3. www.timemap.net
4. http://heuristscholar.org/heurist/
5. www.visionofbritain.org.uk
6. www.nationalarchives.gov.uk/about/whowhathow.htm? source = ddmenu_about1
7. www.gazettes-online.co.uk/, www.statutelaw.gov.uk/
8. Ordnance Survey (OS) is the UK's national mapping agency: www.ordnancesurvey.com/
9. Goldalming parish is near Guildford, Surrey (UK) – www.british-history.ac.uk/report.aspx?compid = 42924
10. www.exploringsurreyspast.org.uk/themes/places/surrey/waverley/godalming
11. www.british-history.ac.uk/report.aspx?compid = 42924&strquery = godalming
12. http://en.wikipedia.org/wiki/Godalming
13. www.nca.org.uk/materials/namingrules.pdf
14. A comprehensive survey can be found from EDINA: http://edina.ac.uk/projects/crosswalk/
15. www.getty.edu
16. www.geonames.org/
17. www.gazetteer.co.uk/
18. www.digital-documents.co.uk/archi/placename.htm
19. www.geoxwalk.ac.uk
20. www.geograph.org.uk/
21. http://developer.yahoo.com/maps/rest/V1/geocode.html

22. http://labs.nationalarchives.gov.uk/
23. http://lucene.apache.org/solr/
24. http://lucene.apache.org/
25. www.nsshutdown.com/projects/lucene/whitepaper/locallucene_v2.html
26. http://openlayers.org/
27. www.openstreetmap.org/

References

Amin, A., van Ossenbruggen, J., Hardman, L. and van Nispen, A. (2008), "Understanding cultural heritage experts' information seeking needs", *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries, Pittsburgh, PA, 16-20 June*, ACM, New York, NY, pp. 39-47.

Amitay, E., Har'El, N., Sivan, R. and Soffer, A. (2004), "Web-a-where: geotagging web content", *Proceedings of the 27th Annual International Conference on Research and Development in Information Retrieval (SIGIR04), Sheffield*, pp. 273-80.

Axelrod, A.E. (2003), "On building a high performance gazetteer database", in Kornai, A. and Sundheim, B. (Eds), *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References*, ACL, Alberta, pp. 63-8.

Buckland, M. and Lancaster, M. (2004), "Combining place, time and topic: the electronic cultural atlas initiative", *D-Lib Magazine*, Vol. 10 No. 5, available at: www.dlib.org/dlib/may04/buckland/05buckland.html (accessed November 2010).

Buckland, M., Chen, A., Gey, F.C., Larson, R.R., Mostern, R. and Petras, V. (2007), "Geographic search: catalogs, gazetteers, and maps", *College and Research Libraries*, Vol. 68 No. 5, pp. 376-87.

Chapman, A.D. and Wieczorek, J. (2006), *Guide to Best Practices for Georeferencing*, Global Biodiversity Information Facility, Copenhagen.

Chen, H. (2001), "An analysis of image queries in the field of art history", *JASIST*, Vol. 52 No. 3, pp. 260-73.

Chen, W. and Nottveit, T. (2010), "Digital map application for historical photos", in Chowdhury, G., Koo, C. and Hunter, J. (Eds), *The Role of Digital Libraries in a Time of Global Change*, Vol. 6102, Chapter 20, pp. 158-67.

Choi, Y. and Rasmussen, E.M. (2003), "Searching for images: the analysis of users' queries for image retrieval in American history", *JASIST*, Vol. 54 No. 6, pp. 498-511.

Clough, P. (2005), "Extracting metadata for spatially-aware information retrieval on the internet", *Proceedings of Workshop on Geographic Information Retrieval (GIR'05), in conjunction with CIKM2005, Bremen*, pp. 25-30.

Clough, P., Ireson, N. and Marlow, J. (2009), "Extending domain-specific resources to enable semantic access to cultural heritage data", *Journal of Digital Information: Special Issue on Information Access to Cultural Heritage*, Vol. 10 No. 6, available at: https://journals.tdl.org/jodi/issue/view/89 (accessed January 2011)

Clough, P., Marlow, J. and Ireson, N. (2008), "Enabling semantic access to cultural heritage: a case study of Tate Online", in Larson, M., Fernie, K., Oomen, J. and Cigarran, J. (Eds), *Proceedings of the ECDL 2008 Workshop on Information Access to Cultural Heritage, Aarhus, September 18*, available at: http://ilps.science.uva.nl/IACH2008/papers/Clough_etal_TateOnline_IACH2008.pdf (accessed January 2011).

Collins, K. (1998), "Providing subject access to images: a study of user queries", *American Archivist*, Vol. 61 No. 1, pp. 36-55.

Crane, H. (2004), "Georeferencing in historical collections", *D-Lib Magazine*, Vol. 10 No. 5, available at: www.dlib.org/dlib/may04/crane/05crane.html (accessed November 2010).

Cunningham, S.J., Bainbridge, D. and Masoodian, M. (2004), "How people describe their image information needs: a grounded theory analysis of visual arts queries", *Proceedings of Joint Conference on Digital Libraries (JCDL) 2004, Tucson, AZ, June 7-11*, pp. 47-8.

Curran, J.R. and Clark, S. (2003), "Language independent NER using a Maximum Entropy Tagger", *Proceedings of CoNLL-2003, Edmonton*, pp. 164-7.

Ding, J., Gravano, L. and Shivakumar, N. (2000), "Computing geographical scopes of web resources", *Proceedings of 26th International Conference on Very Large Databases, VLDB 2000, Cairo, September*, pp. 10-14.

Frew, J., Freeston, M., Freitas, N., Hill, L.L., Janee, G., Lovette, K., Nideffer, R., Smith, T.R. and Zheng, Q. (1998), "The Alexandria Digital Library architecture", in Nikolaou, C. and Stephanidis, C. (Eds), *Lecture Notes in Computer Science*, Springer, Berlin, pp. 61-73.

Fu, F., Jones, C.B. and Abdelmoty, A.I. (2005), "Building a geographical ontology for intelligent spatial search on the web", *Proceedings of IASTED International Conference on Databases and Applications, DBA 2005, Innsbruck, February*, pp. 167-72.

Goodchild, M.F. (2004), "The Alexandria Digital Library Project: review, assessments and prospects", *D-Lib Magazine*, Vol. 10 No. 5, available at: www.dlib.org/dlib/may04/goodchild/05goodchild.html (accessed November 2010).

Goodchild, M.F. and Hill, L.L. (2008), "Introduction to digital gazetteer research", *International Journal of Geographical Information Science*, Vol. 22 No. 10, pp. 1039-44.

Gregory, I. and Southall, H. (2003), *The Great Britain Historical GIS*, University of Portsmouth, Portsmouth, available at: www.port.ac.uk/research/ gbhgis/aboutthegbhistoricalgis/documentation/ (accessed January 2011).

Hardman, L., van Ossenbruggen, J., Troncy, R., Amin, A. and Hildebrand, H. (2009), "Interactive information access on the Web of Data", *Proceedings of the WebSci'09: Society On-Line, Athens, 18-20 March*, available at: http://journal.webscience.org/212/ (accessed 1 December 2009).

Hill, L.L. (2006), *Georeferencing: The Geographic Associations of Information*, The MIT Press, Cambridge, MA.

Himmelstein, M. (2005), "Local search: the internet is the *Yellow Pages*", *IEEE Computer Society Journal*, Vol. 38 No. 2, pp. 26-34.

Jankowski, J., Cobos, Y., Hausenblas, M. and Decker, S. (2009), "Accessing cultural heritage using the Web of Data", *Proceedings of 10th International Symposium on Virtual Reality, Archaeology and Cultural Heritage (VAST), St Julians*, available at: www.grey-eminence.org/articles/VAST2009-CHoWDer.pdf (accessed January 2011).

Janowicz, K. and Kessler, C. (2008), "The role of ontology in improving gazetteer interaction", *International Journal of Geographical Information Science*, Vol. 22 No. 10, pp. 1129-57.

Johnson, I. (2004), "Putting time on the map: using TimeMap for map animation and web delivery", *GeoInformatics*, July/August, pp. 26-9.

Johnson, I. (2005), "Indexing and delivery of historical maps online using TimeMap™", *National Library of Australia News*, Vol. XV No. 4, pp. 6-9.

Kwok, K.L. and Deng, Q. (2003), "GeoName: a system for back-transliterating Pinyin place names", in Kornai, A. and Sundheim, B. (Eds), *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References, ACL, Alberta*, pp. 26-30.

Larson, R.R. (1996), "Geographic information retrieval and spatial browsing", in Smith, L. and Gluck, M. (Eds), *GIS and Libraries: Patrons, Maps and Spatial Information*, University of Illinois, Urbana-Champaign, IL.

Leidner, J.L. (2008), *Toponym Resolution in Text*, Universal Publishers, Boca Raton, FL.

McCurley, S.K. (2001), "Geospatial mapping and navigation of the web", *Proceedings of the 10th International WWW Conference, Hong Kong, 1-5 May*, pp. 221-9.

Manguinhas, H., Martins, B. and Borbinha, J. (2008), "A geo-temporal web gazetteer integrating data from multiple sources", *Proceedings of the 3rd IEEE International Conference on Digital Information Management, London, November 13-16*, pp. 146-53.

Martins, B., Borbinha, J., Pedrosa, G., Gil, J. and Freire, N. (2007), "Geographically-aware information retrieval for collections of digitized historical maps", *Proceedings of the 4th ACM Workshop on Geographical information Retrieval, Lisbon, November 6-10*, ACM, New York, NY, pp. 39-42.

Moritz, T. (1999), "Geo-referencing the natural and cultural world, past and present: towards building a distributed, peer-reviewed gazetteer system", *Digital Gazetteer Information Exchange Workshop, October 13-14*, available at: www.alexandria.ucsb.edu/~lhill/dgie/DGIE_web site/session1/moritz.htm (accessed January 2011).

Mostern, R. and Johnson, I. (2008), "From named place to naming event: creating gazetteers for history", *International Journal of Geographical Information Science*, Vol. 22 No. 10, pp. 1091-108.

Nadeau, D., Turney, P.D. and Matwin, S. (2006), "Unsupervised named-entity recognition: generating gazetteers and resolving ambiguity", *Proceedings of 19th Canadian Conference on Artificial Intelligence, AI 2006, Québec City, Québec, June 7-9*, pp. 266-77.

Nissim, M., Matheson, C. and Reid, J. (2004), "Recognising geographical entities in Scottish historical documents", *Proceedings of the Workshop on Geographic Information Retrieval, SIGIR 2004, Sheffield*, available at: www.geo.unizh.ch/~rsp/gir/abstracts/nissim.pdf (accessed January 2011).

Overell, S.E. and Rüger, S. (2006), "Identifying and grounding descriptions of places", *Proceedings of Workshop on Geographic Information Retrieval, GIR 2006, Seattle, WA, August 10*, available at: www.geo.uzh.ch/~rsp/gir06/papers/individual/overell.pdf (accessed January 2011).

Pask, A. (2005), "Art historians' use of digital images: a usability test of ARTstor", dissertation, University of North Carolina at Chapel Hill, Chapel Hill, NC.

Petras, V. (2004), *Statistical Analysis of Geographic and Language Clues in the MARC Record*, technical report, University of California at Berkeley, Berkeley, CA.

Popescu, A., Grefenstette, G. and Moëllic, P.A. (2008), "Gazetiki: automatic creation of a geographical gazetteer", *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL 2008, Pittsburgh, PA, June 16-19*, pp. 85-93.

Purves, R.S., Clough, P., Jones, C.B., Arampatzis, A., Bucher, B., Finch, D., Fu, G., Joho, H., Khirini, A.S., Vaid, S. and Yang, B. (2007), "The design and implementation of SPIRIT: a spatially-aware search engine for information retrieval on the internet", *International Journal Geographic Information Systems (IJGIS)*, Vol. 21 No. 7, pp. 717-45.

Reid, J.S., Higgins, C., Medyckyj-Scott, D. and Robson, A. (2004), "Spatial data infrastructures and digital libraries: paths to convergence", *D-Lib Magazine*, Vol. 10 No. 5, available at: www.dlib.org/dlib/may04/reid/05reid.html (accessed January 2011).

Sanderson, M. and Han, Y. (2007), "Search words and geography", *Proceedings of the 4th ACM Workshop on Geographical Information Retrieval, GIR*, ACM, New York, NY, pp. 13-14.

Smith, D.A. and Mann, G.S. (2003), "Bootstrapping toponym classifiers", in Kornai, A. and Sundheim, B. (Eds), *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References, ACL, Alberta*, pp. 45-9.

Zhang, V.W., Rey, B., Stipp, E. and Jones, R. (2006), "Geomodification in query rewriting", *Proceedings of the 2006 Workshop on Geographic Information Retrieval, Seattle, WA*, pp. 23-7.

Zhou, G. and Su, J. (2002), "Named entity recognition using an HMM-based Chunk Tagger", *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL2002), Philadelphia, PA, July*, pp. 473-80.

Zong, W., Wu, D., Sun, A., Lim, E. and Goh, D.H. (2005), "On assigning place names to geography-related web pages", *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries, Denver, CO, June 7-11*, pp. 354-62.

Corresponding author
Paul Clough can be contacted at: p.d.clough@sheffield.ac.uk