Extracting Metadata for Spatially-Aware Information Retrieval on the Internet

Paul Clough
University of Sheffield
Western Bank
Sheffield, UK
+44 (0)114 2222664
p.d.clough@sheffield.ac.uk

ABSTRACT

This paper presents methods used to extract geospatial information from web pages for use in SPIRIT, a new Geographic Information Retrieval (GIR) system for the web. The resulting geospatial markup tools have been used to annotate around 900,000 web pages taken from a 1TB web crawl, focused on regions in the UK, France, Germany and Switzerland. This paper discusses a versatile geo-parsing tool for extracting spatial metadata based upon the GATE Information Extraction (IE) system, and a simple geo-coding program based on default sense to assign spatial coordinates to extracted locations. A preliminary analysis of markup accuracy for geo-parsing and geo-coding is provided, and an initial statistical and geographical analysis of the SPIRIT collection presented.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *search process, retrieval models, information filtering*; H.3.5 [**Information Storage and Retrieval**]: Online Information Services – *Web-based services*

General Terms

Algorithms, Experimentation.

Keywords

Spatial markup, Geographic Information Retrieval (GIR).

1. INTRODUCTION

Many documents on the web contain geospatial information including addresses, postal codes, hyperlinks and geographic references [1][2]. This information can be exploited and used to provide spatial awareness to information systems. These include transport timetables, routing systems for motorists, map-based web sites and location-based services (e.g. Google Local and Yellow Pages). A key part of providing such services is the extraction and use of geospatial information. In this paper we discuss approaches used in the Spatially-Aware Information Retrieval on the Internet (SPIRIT) project (http://www.geo-spirit.org/) to generate a sample web document collection for prototyping a working GIR system [3]. Extracting geospatial references from documents involves two main tasks: identifying geographic references and assigning them spatial coordinates. These are commonly referred to as geo-parsing and geo-coding respectively [4]. The final approach adopted for each task has been influenced by the following criteria:

- **Speed –** fast execution of geo-parsing and geo-coding programs to allow the processing of large collections (e.g. the SPIRIT 1TByte web collection [6]) within feasible timescales.
- **Reliability** – robust processing on typical web data with error-recovery strategies to ensure minimal manual intervention.
- **Flexibility** – enable control over the geo-parsing process including the addition of custom gazetteer lists and creation of grammars for context matching.
- **Multilingualism** – to be able to process texts written in a variety of languages other than English.

Given these constraints, geo-parsing and geo-coding methods used in the SPIRIT project are based on simple approaches: for geo-parsing, gazetteer lookup is supplemented with context rules to filter out common-usage words and personal names; for geo-coding a default sense is used based on information provided by geographic resources available in the SPIRIT project. These simple approaches provide a baseline against which more advanced methods can be compared. Geospatial information extracted from these pages is used to enable the retrieval of documents which are not only topically relevant to a query, but also match spatial criteria (e.g. "stone circles near Sheffield"). The SPIRIT system supports this by indexing web pages based on extracted spatial footprints and textual keywords.

2. SPATIAL MARKUP

2.1 Sources of Spatial Data

Several sources of geographic information were available in the SPIRIT project (summarised in Table 1). Each resource contains both place names and spatial information which are used in the geo-parsing and geo-coding stages respectively. Resources differ in granularity of place names (e.g. OS contains villages and points

GIR'05, November 4, 2005, Bremen, Germany.

of interest), quality of spatial reference (e.g. OS is more accurate than TGN), scope (e.g. global versus national) and type of spatial representation (e.g. point versus polygon). The main resources used were: (1) the SABE[1] (Seamless Administrative Boundaries of Europe) dataset, (2) the Ordnance Survey 1:50,000 Scale Gazetteer[2] for the UK and (3) the Getty Thesaurus of Geographic Names (TGN[3]). The SABE data was integrated into ontological form for use in the SPIRIT project [5]. The resources in Table 1 also differ in the metadata they provide about each location.

For example, in addition to the place name and spatial coordinates, TGN and SABE both provide hierarchical information pertaining to regions which encompass the location; whereas the OS gazetteer provides only localized information (e.g. Hillsborough is in Sheffield), but does provide useful information such as feature type (e.g. city, town, city, water).

	Scope	Footprint	Total	Unique
TGN	Global	Point	1,063,259	630,427
OS	UK	Point	258,797	202,112
SABE	UK	Polygon	22,370	19,951
SABE	France	Polygon	54,889	44,392
SABE	Germany	Polygon	20,427	15,723
SABE	Switzerland	Polygon	3,716	3,344

Table 1 Geographic resources used in SPIRIT

2.2 SPIRIT 1TByte Web Collection

The SPIRIT web collection [6] consists of 94,552,870 web pages with an approximate size of 1TByte. The crawl was undertaken in Mid-2001 by the University of Waterloo, seeded by a set of educational websites. Based on domain name, approximately 9.6 million pages (approximately 10.24% of the entire collection) were from European domains; the rest mainly from US sites. A total of 22 European domains were found in the top 50 most frequent domains, of which nearly half are from the .uk and .de domains. To facilitate storage and processing, pages were distributed across a cluster of 25 machines.

Based on this collection, all pages for the four focus regions: UK, France, Germany and Switzerland were extracted based on top domain name. In total, 1,759,681 UK, 1,556,585 German, 505,023 French and 270,715 Swiss web pages were extracted for geo-parsing. To reduce the amount of text to process and simplify the annotation process, the `UNIX lynx` command was used to extract plaintext. This was also found to help to reduce false hits, e.g. names within HTML tags, although at the cost of removing potentially useful contextual evidence.

2.3 Geo-Parsing

Geo-parsing is closely related to the more general problem of Named-Entity Recognition or NER [7]. Most NER algorithms combine lists of known locations, organisations and people (called gazetteers) with rules, which capture elements of the surrounding context. Although simple list lookup for place names can perform well [8], the approach fails when list entries are not used in a geographic sense (e.g. names of people, businesses or common language use) or variants of names are used (e.g. historical or vernacular forms). Somewhere in the text there is likely to be external (or contextual) evidence, which makes clear what type of entity a word or phrase is (e.g. lexical or structural linguistic clues).

In SPIRIT, due to constraints on execution speed, gazetteer lookup was first used to identify candidate place names. Following this, contextual evidence was used to filter out locations not being used in a geographic sense (an approach similar to [9][10]). Although most NER approaches attempt to use context rules to find locations not found in the gazetteer lists, in SPIRIT this was not a concern because only locations for which we had the necessary spatial information were assigned co-ordinates. Gazetteer lookup also has the benefits of being language independent and robust, which is very important for web pages which are often ungrammatical and contain limited context. Upon initial investigation, a large degree of ambiguity was found to derive from entries in the gazetteer being used as common words (e.g. Bath, Battle and Derby) and entries being used as part of a proper name (e.g. as the surname of a person).

The current implementation of the SPIRIT geo-parser is built using the General Architecture for Text Engineering (GATE) system which provides a Collection of REusable Objects for Language Engineering (CREOLE), a set of resources integrated into GATE [11]. GATE provides a grammar called JAPE[4], which is compiled into a finite state transducer for pattern matching (similar to `flex` and `lex`). Rules can be defined within terms of entities (or annotations) identified within GATE, which may or may not depend on previous stages in the IE process. Using GATE as a framework enables experimentation with using full IE versus gazetteer lookup for NER (as well as creating custom context rules).

The final approach filters candidate locations using context rules to remove stopwords, references to people and organizations, and links to emails/URLs (e.g. "Mr. Sheffield" is filtered out as a non-geographic reference using the context rule "<title><location> → null" where <title> and <location> are placeholders for entries from the gazetteer lists). Rules requiring minimal language-processing have been manually created thereby enabling operation between languages. Language-dependent gazetteer lists have been translated using the Systran MT system (e.g. "Mr." translated into German as "Herr."). Although a simple and limited approach, this enabled multilingual geo-parsing in SPIRIT to deal with the selected focus regions.

Department of Information Studies, University of
<location>Sheffield</location>, Regent Court, 211 Portobello St,
<location>Sheffield</location>, <location>S1 4DP</location>,
<location>UK</location>

Figure 1. Example of current address markup

[1] http://www.eurogeographics.org/eng/03_projects_sabe.asp

[2] http://www.ordnancesurvey.co.uk/oswebsite/products/50kgazetteer/

[3] http://www.getty.edu/research/conducting_research/vocabularies/tgn/

[4] Java Patterns Annotations Engine (JAPE).

Place names occurring at the start of certain organizations (e.g. universities) were kept as found to be generally useful geographical information. Address and address fragments are not dealt with formally but instead buildings and streets are ignored. From addresses, place names and postcodes are extracted and then grounded (Figure. 1).

2.3.1 Filtering out commonly used words

All geographical resources used in SPIRIT have entries for places which are more commonly used in a non-geographical sense (e.g. "New", "More", "Read", "Guide" and "Old" all appear as entries in the OS gazetteer). A stopword list for each geographic resource was computed by calculating document frequency from each entry in the gazetteer based on the SPIRIT 1TByte collection. The top 1000 place names (ranked in descending order of document frequency) were manually assessed for any well-known locations (e.g. "Bath" and "Derby" in the UK) and removed from the stopword list. Next, gazetteer entries also matching dictionary words were filtered out (assuming to be common words of language use). Entries in `UNIX dict` beginning with a lowercase letter were extracted and compared against the gazetteer lists filtering out any matches (e.g. we found 3,985 dictionary terms in TGN and 538 in OS). Finally, lists of stopwords from the Snowball stemmer project[5] were used for each language (includes names of common words, e.g. determiners, conjunctions etc).

2.3.2 Using person name lists

In addition to using lists of common words, GATE was extended with lists of person names (both forenames and surnames) to deal with ambiguity resulting from locations being part of a personal name (e.g. "John Sheffield"). Sources for the names are provided in Appendix 1 including English, French and German names. These were used in conjunction with the context rules to help reduce false hits.

2.4 Geo-Coding

After extracting candidate locations, the second stage involves assigning them spatial co-ordinates (through gazetteer matching). This involves disambiguating place names with multiple spatial references (referent ambiguity [12]). For example, the name "Chapeltown" refers to a location in South Yorkshire (UK), Lancashire (UK), Kent County (USA) and Panola County (USA). Table 1 indicates the degree of ambiguity for each spatial resource used in SPIRIT. For example, in OS approximately 8% of place names are ambiguous. TGN exhibits the most ambiguity (59%) due to ambiguity between countries as well as within a single country.

The simplest (and often most effective) method to resolve referent ambiguity is to assign ambiguous places a default sense (or position). This can be decided by, for example, the most commonly occurring place [12], by population of the place name [13] or by semi-automatic extraction from the Web [9]. Based on available metadata provided by resources used in SPIRIT, the following attributes were used to establish a default sense: (1) the length of hierarchy containing a location for SABE and TGN data (e.g. World>North and Central America>United States>New York has length 4), and (2) the feature type provided by OS data. These are used with the assumption that places with shorter hierarchies or certain feature type (in the order of city → town → village) are more likely the location being referred to. To generate a hierarchy for the OS data and provide a spatial footprint useable in the SPIRIT system (bounding box or polygon), a spatial join was computed between the SABE UK and OS resources. This mapped the OS names to an entry in SABE. Hierarchies between TGN and SABE were normalised to enable comparison between them.

Although TGN was not used to ground places, having a global resource was found beneficial in situations when UK names are used to refer to non-UK places. For example, using only the UK resources to ground "New York" would map the location to North Tyneside (Newcastle) which is a small region. However, in practice it is more likely that this refers to the city New York (provided by TGN). In addition to using the default sense, the overlap (matching words) between place names in the hierarchy of an ambiguous location and those found within *n* words either side of the name (from empirical study we have found a good choice for *n* to be 2 words left and 8 words right) was also computed. This helps to deal with cases when the local context qualifies the location (e.g. "Lancaster, UK" and "Lancaster, PA") and provides good evidence for sense selection.

A further feature of the geo-coding method used in SPIRIT is the use of preferences to rank senses: (1) senses from resources in the following order are preferred (due to the quality and type of spatial reference): SABE [3] → OS [2] → TGN [1], and (2) senses matching a command-line option specifying the country currently being processed (e.g. "UK" or "Germany" – the preferred country is given a value 1; the rest a value 0) are preferred. The algorithm is implemented in Perl and senses are matched by sorting resource matches based on overlap, hierarchy depth, resource and country preference. In summary, the geo-coding algorithm is as follows:

```
For each ambiguous location
        Get matching locations from resource [TGN,SABE,OS]
        Foreach matching location (sense)
                Compute overlap score between hierarchies and
                local context
                Compute hierarchy depth
                Sort senses in following order
                  By overlap score (matching words)
                  By depth of hierarchy
                  By resource preference
                  By country preference
        End
        For equal ranking senses select SABE → OS → TGN
        Assign sense ID
End
```

3. EVALUATING SPATIAL MARKUP

3.1 Experimental Procedure

From the final annotated collection of 885,502 documents (described in section 4.1), 169,442 documents were initially selected containing between 5 and 10 unique footprints (to make evaluation feasible). From these, 10% were randomly selected (without replacement) and documents with all footprints assigned to the UK (we only evaluated UK markup due to available language resources for assessment) selected. This resulted in 130 documents for evaluation. Using the GUI provided by GATE, *all* geographic names (1864 in total) were manually identified and stored as a benchmark for further comparison.

[5] http://snowball.tartarus.org/

3.2 Geo-Parsing Results

Using the benchmark data, different geo-parsing were evaluated: the default version of ANNIE from GATE, gazetteer lookup (GAZ) and the SPIRIT geo-parser which uses the additional context rules and stopword lists (SPIRIT). Experiments using each geographical resource were carried out and evaluated using the GATE AnnotationDIFF tool to compare the benchmark annotations (key-set) with those generated by the system (response-set). AnnotationDiff creates several measures of annotation overlap, including: correct[6] (C), missing (M), false hits (S), precision (P), recall (R) and F_1-measure (F1). The precision, recall and F_1-measure are computed from the number of annotations found to match correctly, the number of annotations missing from the key-set, and the number of false.

	C	M	S	P	R	F1
GATE	772	1101	239	0.7563	0.4135	0.5347
GATE+SABE	923	897	1088	0.4863	0.5046	0.4518
GATE+SABE+OS	1084	729	1760	0.3848	0.5933	0.4668
GAZ+SABE	1199	570	2065	0.3725	0.6654	0.4776
GAZ+SABE+OS	1423	372	5380	0.2137	0.7763	0.3351
GAZ+SABE+OS+TGN	1550	223	8380	0.1928	0.8776	0.3008
SPIRIT+SABE+OS	**1340**	**479**	**596**	**0.6966**	**0.7820**	**0.7148**
SPIRIT+SABE+OS+TGN	1385	423	798	0.6535	0.7899	0.6916

Table 2. Geo-parsing results using different systems

From the annotations in the response-set, automatically generated, precision measures the proportion of these matching the manually assigned annotations. From the annotations defined manually, recall measures the proportion of these which are also correctly identified by the geo-parser. The F1 score is a single-valued summary of both precision and recall and enables much simpler comparison of different geo-parsing methods.

Observations about the geo-parsing from Table 1 include the following. First, comparing the gazetteer lookup results only, it appears that by adding additional geographic resources (starting with SABE UK), more locations are correctly found but at the cost of many more false hits (S). Using TGN obtains more correct matches because some pages contain global names incorrectly ground to the UK. Second, compared with using a more sophisticated IE approach (GATE), the gazetteer lookup with SABE data fairs favorably and the best GATE result is using GATE in the default setup. This is because the additional geographical resources add more ambiguous names which are not removed because stopword removal is not included in default GATE. Thirdly, the highest F1 score is obtained using the SPIRIT geo-parser with SABE and OS resources. Adding TGN increases the number of correct, but again at the cost of more false matches. Upon inspection, many of the missing are due to variations in spelling (e.g. “South Yorks” vs. “South Yorkshire”), phrases such as “North Scotland” where only Scotland is located in the gazetteer and names written in all uppercase letters.

Table 3 shows the contribution of each heuristic to geo-parsing results based on the SPIRIT geo-parser with SABE and OS (highest F1 score from Table 1 which is baseline in Table 2). The main observation is the effect of not removing the stopwords: the F1 score drops from 0.7148 to 0.5433 (24% decrease) indicating the importance of identifying and removing stopwords. Not filtering out gazetteer entries which are proper names decreases the baseline by about 6%. Similarly, ignoring the identification of postcodes decreases the baseline by about 5%.

	C	M	S	P	R	F1
Baseline	**1340**	**479**	**596**	**0.6966**	**0.7820**	**0.7148**
Without stopwords	1398	401	2022	0.4114	0.7621	0.5433
Without persons	1412	403	1101	0.6086	0.8220	0.6732
Without streets	1340	479	596	0.6966	0.7821	0.7148
Without emails/urls	1340	479	596	0.6966	0.7821	0.7148
Without postcode	1236	578	594	0.6849	0.7324	0.6836

Table 3. Contribution of heuristics to SPIRIT geo-parser

3.2.1 Execution time

To gain an indication of speed, the average execution time over 5 runs of the geo-parser on 100 randomly selected documents from the 1TByte collection was computed using a Viglen machine (Intel Pentium 4 CPU 2GHz with 1GB RAM). The average set-up time was 28.5s using UK lists (SABE and OS) and 15.4s using default lists only. The average time spent on writing out the annotations (UK only) was: all annotations 40.3s (37% of total time) and locations only (2.9s – 1.6% of total time). On average execution time as 147 seconds. For 10,000 documents using all resources on web pages from the SPIRIT collection, using the SPIRIT geo-parser required approximately 3 hours of processing time.

3.3 Geo-Coding Results

Based upon the 1864 place names *manually* identified in the evaluation data (which includes around 100 postcodes), 1021 can be found in TGN of which approximately 68% are ambiguous. In

[6] AnnotationDIFF also computes partially correct (PC) which is not shown in Table 1.

comparison, 942 locations can be found in the UK SABE data of which approximately 11% are ambiguous. This corresponds with Table 1 where TGN contains more ambiguity as a global resource than the other country-specific data.

Using all resources to geo-code, 1581 are matched of which 1283 have more than one unique identifier (UID). These come from both ambiguous names and the same location being found in more than one resource. In total 1137 places (89%) are correctly assigned a UID: these are locations which are not only assigned to the correct geographic sense, but also the correct resource order (SABE → OS → TGN). Places assigned to TGN are effectively ignored. The geo-coding algorithm is implemented in Perl and on the same set-up described in Section 3.2.1, ran in approximately 13 seconds over the benchmark data. The use of default sense would appear successful in the majority of cases because typically ambiguous locations are referring to the largest or most popular place. The use of multiple geographical resources and our ranking method for sense selection appears to provide useful results. For example, the use of TGN as a global source of geographic knowledge appears useful in cases such as:

>>>> Europe <<<<
TGN1000003 (World>)
CH3851 (World>Europe>Switzerland>Zurich >)
SELECTED: TGN1000003 (World>)

In this example, without TGN Europe would be assigned to a place in Switzerland. The preference of resources also appears to work:

>>>> Histon <<<<]
TGN1029574 (World>Europe>United Kingdom>England>Cambridgeshire>)
OS138421 (World>Europe>United Kingdom>England>Cambridgeshire>South Cambridgeshire>)
UK1519 (World>Europe>United Kingdom>England>Cambridgeshire>South Cambridgeshire>)
SELECTED: UK1519 (World>Europe>United Kingdom>England>Cambridgeshire>South Cambridgeshire>)

In this example, without the resource preference ranking, the TGN sense would have been selected because of the shorter hierarchy length.

	# Files	Unique UIDs	Unique UID occurrences	Avg. UIDs/file
UK	339,819	25,841	1,541,442	3.97
France	363,183	7,504	959,104	2.61
Germany	79,491	2,648	321,362	2.85
Switzerland	87,009	5,832	258,188	3.1

Table 4. Summary of footprints (UIDs) for SPIRIT collection

4. SPIRIT COLLECTION

Based on the spatial markup 885,502 web pages were finally included in the SPIRIT collection. This is less than total number described in Section 2.2 because many files either contained no locations, or contained locations unable to be grounded.

4.1 Collection Statistics

Table 4 provides some analysis of the SPIRIT collection based on the number of spatial footprints (or UIDs) extracted from the processed web pages. The number of unique UIDs provides the number of unique UIDs counted once for each documents; unique UID occurrences provides the total number of times each UID appears on the web page. For the UK, on average each place occurs around 60 times, but of course this is not proportional as some locations (e.g. names of countries or cities) appear more times than others.

4.2 Geographical Analysis

Figure 2 shows a density plot of footprints in the SPIRIT collection for the UK with a document frequency of greater than 50. Labels for some locations are also shown. The red and yellow areas (regions surrounding the points) indicate the regions of highest density and as expected these occur at major populated areas such as cities (e.g. London, Liverpool and Nottingham). The plot does highlight the effect of not removing high frequency words which are commonly not used in a geographic sense (e.g. "Fans" and "Stand") in Scotland. Also the location "Moscow" is most likely wrong and results from incorrect geo-coding. Despite the false hits, the general coverage of the SPIRIT UK collection appears reasonable.

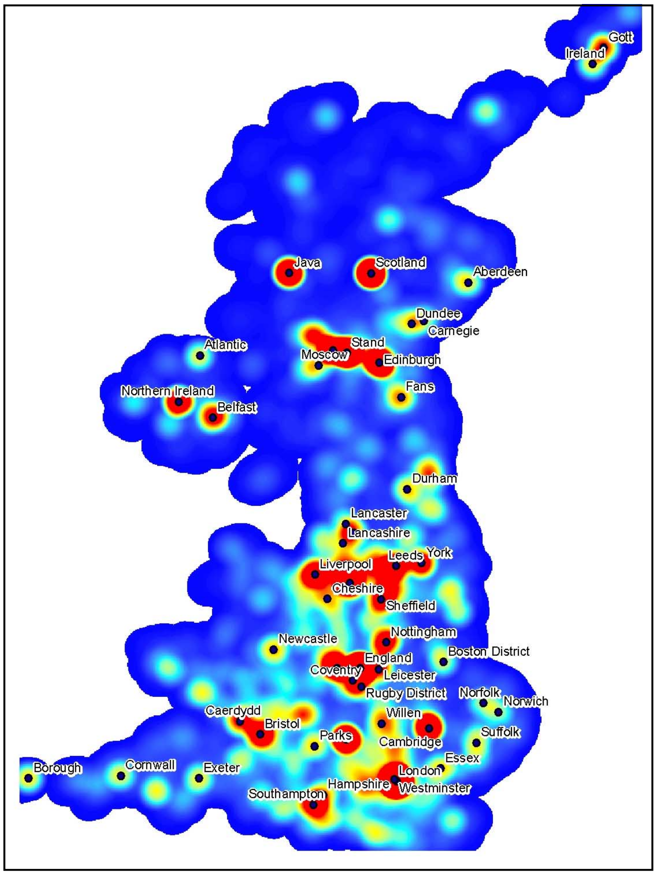

Figure 2. Density plot of UK SPIRIT collection based on places with a document frequency > 50.

5. CONCLUSIONS

This paper presents work from the SPIRIT project on extracting spatial metadata from web pages for prototyping a working GIR system [3]. Simple methods for geo-parsing and geo-coding have been presented which address specific constraints including execution time and language independence. Methods used are relatively simple and provide a baseline upon which to construct more complex approaches. In particular, the geo-parsing method used has been based on the GATE system providing a versatile framework in which to develop custom tools. The use of gazetteer lookup provides an F1 score of 0.7148 when used with removal of commonly occurring words (stopwords) and other entities including person names to address cases when gazetteer entries are used in a non-geographical sense. The geo-coding method is also simple based on default sense and matches between metadata provided by the geographical resources and local context. Using a small set of manually annotated data, 89% of locations are correctly assigned a unique identifier (UID). In particular, the grounding method also deals with resource selection in addition to referent ambiguity. Whether this accuracy of markup is sufficient in practice is still being investigated, but both user and system evaluation of the SPIRIT prototype based on this markup have shown promising results [14]. In the case of SPIRIT, further methods for ranking results also help to reduce the effects of incorrect markup.

The current geo-parsing method could be improved by enhancing the gazetteer matching method and the filtering of non-geographic senses of gazetteer entries. In the former, supplementing the gazetteer list with variant forms could improve gazetteer lookup alone. For the latter, generating better lists of stopwords and using further context rules could reduce false hits. In particular, using machine learning methods to generate further context rules based on features derived from local and more global evidence would alleviate the need for creating rules manually. Dealing more effectively with addresses and address fragments would also improve markup.

For geo-coding investigating better methods for combining various geographical resources (e.g. different gazetteer lists and spatial resources) could improve performance. In addition, generating further metadata for each sense against which to match the local context will be investigated, e.g. Sheffield in the UK will tend to co-occur with terms such as “steel” and “Blades” (a local football team). Identifying the extent of a reference (rather than relying on a fixed window) could also help to limit erroneous matches due to incorrect local context. This could be as simple as segmenting the text into passages, or using more sophisticated methods to identify blocks of geo-references such as HTML markup (e.g. table cell delimiters).

6. ACKNOWLEDGMENTS

This research was supported by the EU-IST Project No. IST-2001-35047 (SPIRIT). The author wishes to thank Hideo Joho and Ross Purves with their support in this work.

7. REFERENCES

[1] Junyan Ding, Luis Gravano, and Narayanan Shivakumar. Computing geographical scopes of web resources. In *26th International Conference on Very Large Databases*, VLDB 2000, Cairo, Egypt, September 10--14

[2] McCurley, S.K. (2001) Geospatial mapping and navigation of the web. *In Proceedings of the Tenth International WWW Conference*, Hong Kong, 1-5 May, 221-229.

[3] Jones C.B., R. Purves, A. Ruas, M. Sanderson, M. Sester, M.J. van Kreveld, R. Weibel (2002). Spatial information retrieval and geographical ontologies an overview of the SPIRIT project. *SIGIR 2002: In SIGI’02*, Tampere, Finland, 387-388.

[4] Larson, R.R. (1996) Geographic Information Retrieval and Spatial Browsing. In *GIS and Libraries: Patrons, Maps and Spatial Information*, *Linda Smith and Myke Gluck, Eds.*, University of Illinois.

[5] Jones C.B., A.I. Abdelmoty and G. Fu (2003) Maintaining ontologies for geographical information retrieval on the web, *In Meersman, R.; Tari, Z.; Schmidt, D. C. (Eds.) On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE Ontologies, Databases and Applications of Semantics*, ODBASE’03, Catania, Italy.

[6] Joho, H. and Sanderson, M. (2004) The SPIRIT collection: an overview of a large web collection. *In SIGIR Forum*, 38(2), 57-61.

[7] Cowie, J. and Lehnert, W. (1996) Information Extraction. *Communications of the ACM*, 39(1), 80-91.

[8] Mikheev A., Moens M. and Grover C. (1999) Named Entity recognition without gazetteers. In *Proceedings of the Annual Meeting of the European Association for Computational Linguistics EACL'99*, Bergen, Norway, 1-8.

[9] Li, H., et al. (2003) InfoXtract location normalization: a hybrid approach to geographic references in information extraction. In: Kornai, A. and Sundheim, B. (eds.) *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References*, Alberta, Canada: ACL, 39-44.

[10] E. Amitay, N. Har’El, R. Sivan, and A. Soffer. (2004) Web-a-where: geotagging web content. *In Proceedings of the 27th SIGIR*, pages 273–280, 2004

[11] Cunningham, H., Maynard, D., Bontcheva, K., and Tablan, V. (2002) GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings of ACL'02*. Philadelphia, July 2002.

[12] 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*, Alberta, Canada: ACL, 45-49.

[13] Rauch, E., et al. (2003) A confidence-based framework for disambiguating geographic terms. In: Kornai, A. and Sundheim, B. (eds.) *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References*, Alberta, Canada: ACL, 50-54.

[14] Bucher, B., Clough, P., Joho, H., Purves, R., and Syed, A. K. (2005) Geographic IR Systems: Requirements and Evaluation. *In: Proceedings of the 22nd International Cartographic Conference,* A Coruña, Spain, CD-ROM.

Appendix 1

- Personal names - http://en.wikipedia.org/wiki/Personal_name
- French boy’s names - http://french.about.com/library/travel/bl-fr-names-m.htm
- French surnames - http://genealogy.about.com/cs/surname/a/french_surnames.htm
- German surnames - http://www.last-names.net/origincat.asp?origincat=German
- Identifying German names - http://www-lib.iupui.edu/kade/nameword/apend-a.html
- U.S. census name files from 1990 - http://www.census.gov/genealogy/names/names_files.html
- Names from the Bible - http://www.behindthename.com/nmc/bibl.html