

# **The Design and Implementation of SPIRIT: a Spatially-Aware Search Engine for Information Retrieval on the Internet**

ROSS S. PURVES<sup>\*1</sup>, PAUL CLOUGH<sup>2</sup>, CHRISTOPHER B. JONES<sup>3</sup>,  
AVI ARAMPATZIS<sup>4</sup>, BENEDICTE BUCHER<sup>5</sup>, DAVID FINCH<sup>3</sup>, GAIHUA FU<sup>3</sup>,  
HIDEO JOHO<sup>2</sup>, AWASE KHIRNI SYED<sup>1</sup>, SUBODH VAID<sup>3</sup>, AND BISHENG YANG<sup>1</sup>

<sup>1</sup>\*Department of Geography, University of Zurich, Switzerland

<sup>2</sup>Department of Information Studies, University of Sheffield, UK

<sup>3</sup>School of Computer Science, Cardiff University, UK

<sup>4</sup>Institute of Information and Computing Sciences, Utrecht University, Netherlands

<sup>5</sup>Laboratoire COGIT - Institut Géographique National, France

\*Corresponding author ([rsp@geo.unizh.ch](mailto:rsp@geo.unizh.ch); +41 44 635 6531)

# **The Design and Implementation of SPIRIT: a Spatially-Aware Search Engine for Information Retrieval on the Internet**

**Abstract:** Much of the information stored on the web contains geographical context, but current search engines treat such context in the same way as all other content. In this paper the design, implementation and evaluation of a spatially-aware search engine are described which is capable of handling queries in the form of the triplet of <theme><spatial relationship><location>. The process of identifying geographic references in documents and assigning appropriate footprints to documents, to be stored together with document terms in an appropriate indexing structure allowing real-time search is described. Methods allowing users to query and explore results which have been relevance ranked in terms of both thematic and spatial relevance have been implanted and a usability study indicates that users are happy with the range of spatial relationships available and intuitively understand how to use such a search engine. Normalised precision for 38 queries, containing four types of spatial relationships is significantly higher ( $p < 0.001$ ) for search exploiting spatial information than pure text search.

**Keywords:** Geographical Information Retrieval, Geographical Ontology, Spatial Indexing, Spatial Relevance Ranking, Information Retrieval.

## 1 Introduction

As the web has grown, so has the realisation that locating information through traditional search engines is inadequate since the semantics of information are, in general, discarded and instead treated as a “bag of words” with meaning attached to individual terms through, for example, their scarcity, but not their context. Thus a “man biting a dog” is no more or less special than a “dog biting a man”. The semantic web (Berners-Lee *et al.*, 2001) aims to address this limitation by allowing machine-readable tagging to attach semantics to terms. In the paper from Berners-Lee *et al.* it is notable that many of the examples of such semantics relate to locations and relationships between them, for example the notion that “a Cornell University address, being in Ithaca, must be in New York State which is in the U.S.,....”.

This is perhaps unsurprising, especially to GIScientists, since almost everything that we do can be regarded as having some form of geographical context, and therefore many information resources refer in some way to location. This importance of geographic context is reflected by resources on the web, many of which contain some reference to location, through for example a place name, address, or less directly a telephone number. McCurley (2001) estimated that around 10% of web documents referred to a US Zip code or telephone number, with a considerably larger number referring in some other way to place – for instance in this study, from a sample of some 20,000 web pages identified as having being related to the UK on the basis of IP address, around 85% contained a reference to a UK place name. In a similar study, Himmelstein (2005) found at least 20 percent of web

pages included one or more easily recognisable and unambiguous geographic identifiers, such as a postal address. Since location is a common element of information on the web, queries submitted to search engines also often contain a reference to location - a study by Zhang et al. (2006) found that some 12.7% of four million queries sampled contained a place name. However, such geographic terms are treated by conventional search engines in the same way as other terms and thus important semantic information is potentially discarded.

Egenhofer (2002) in his description of a potential “Semantic Geospatial web” describes some typical issues through the example of the query “lakes in Maine”. If this query was submitted to a conventional search engine, the spatial relationship “in” would generally be treated as a stop word and discarded from the search. As a further example, Egenhofer (2002) suggests that it is possible that data describing such lakes might be available, but described only in documents which name counties in Maine, but not Maine itself. Thus, some form of spatial join must be made between data describing the lakes and counties (and their geometry) and counties and the state of Maine (and its geometry). Egenhofer suggests that a central challenge for a geospatial web is therefore that it “captures, analyses and tailors geospatial information, much beyond the purely lexical and syntactic level.”

Geographic search can therefore be seen as a key element of a semantic web, since geography is perhaps one of the most familiar and common means by which we ascribe context to information. In turn, as described above, since conventional search engines lack the ability to consider semantics, geographic qualifiers describing spatial relationships such

as *inside*, *north of* or *near* are not treated geographically. Equally, geographic terms which may be shared between locations – for example, there are literally hundreds of Newports – are not disambiguated, and thus the user is unlikely to find relevant results for a small hamlet called Newport. Documents relating to somewhere inside a large place, such as London, may contain no mention of that place name, but refer instead to a district (e.g. Westminster) inside London and thus not be located by a search for “cathedrals in London”. Equally geographic names may be different in different languages (e.g. Geneva, Genf and Genève) or be used in a non-geographic context (e.g. Jack London). These and other limitations can only be addressed if geographic information is treated differently, and in particular if spatial relationships can be in some way stored and used in analysis of the results of a query.

## **Related Work**

Recent years have seen an increase in research dedicated to geographical information retrieval. The detection of geographic content is of much interest and as Himmelstein (2005:29) states retrieval of this kind of information has generated much interest in both research and commerce. Research 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, methods for the formulation of spatial queries and interaction with the results of geographic search. Many of these research areas are directly related to providing geographical web search. Wang et al. (2005) classifies the areas of research in this area into three major aspects: identifying

and disambiguating place names, developing effective computational approaches to handle geographic information and exploiting various geographic information sources. We do not set out to review each area here, but rather focus on research projects dedicated to geographical information retrieval.

Ding *et al.* (2000) described an experimental web search facility that used a gazetteer to analyse the presence of geographic references, including their frequency of occurrence, and hence to index the content of online newspapers. Techniques for georeferencing web documents with respect to their originators were described by Buyukokkten *et al.* (1999) who derived addresses associated with web sites, following analysis of IP addresses. The prototype system defines a geographical scope for a US web page based on the geographical distribution of web pages linking to it. The extent of a page is then computed as a circle, the radius defined by the distribution of geo-referenced links to the page and visualised on a map.

McCurley (2001) described an experimental system for web navigation that employed a number of methods for analysing the geographic context of web pages and allowed the user to find web sites related geographically to a currently retrieved web site. An experimental geographic web search facility, employing a map interface for specification of the area to be searched, based on the city of Kyoto has been presented by Lee *et al.* (2003) called KyotoSEARCH.

The GeoSearcher project (Watters & Amoudi, 2002) has experimented with location-based ranking of search engine results by analysis of the URL's of retrieved web pages. The prototype provides an alternative ranking order for results returned by search engines in response to users queries related to physical locations and distances. As an add-on to an existing search engine, search results were re-ranked based on distance between the query and scope of retrieved document determined by grounding the URL of a page to longitude/latitude.

Rauch et al. (2003) describe their approach for incorporating geography into web search which is being developed commercially by MetaCarta (<http://www.metacarta.com>). Their approach computes a geographical focus for each web page which is used to restrict results to a specific region and display results on a map. A disambiguation method involving a confidence-based framework is used to model the probability that a given name refers to a given place. An additional stage during the extraction of locations also attempts to understand grammatical expressions which define some kind of relative positioning such as "15 miles north of Washington".

Markowetz *et al.* (2004) described aspects of an experimental geographic web search engine that is similar in intent to the search engine described in this paper, with its emphasis upon general web search. They describe methods for determining the geographical scope of web pages based on multiple sources of evidence including the WHOIS directory and the text content of the web pages. For each web page a geographic footprint is created consisting of a set of grid cells recording the degree of spatial relevance

of the document for each cell. Their approach does not include interpretation of spatial relationships that relate the subject of interest to a particular place, other than distance, and no evaluation of their methods is described.

Amitay *et al.* (2004) describe a system for geotagging web content called Web-a-Where that is able to associate locations to pages with the aim of applying this to location-based applications. The system extracts and grounds geographic references found in web pages and uses these to compute a focus for the page – a locality the page is assumed to refer to as a whole. The authors distinguish between two types of location: source and target. The former has to do with the origin of the page (e.g. the physical location of the server hosting the page); the latter reflects the coverage of a pages' content. The authors showed how the Web-a-Where system could be integrated into an existing data-mining framework to geotag web pages gathered from a web crawl and thereby enable geographic search. Similarly, Wang *et al.* (2005) investigated the detection of geographic locations from web pages through the assignment of a single *geographic scope*. In particular, they categorised geographic references into three types: provider location, content location and serving location to reflect the observation that various types of location can coexist in a single source. Provider location identifies the physical location of the provider who owns the web resource (e.g. derived from Yellow Page addresses), content location refers to the location that the content of a web resource describes, and serving location defines the geographic scope that a page can reach (e.g. through analysis of usage logs or links pointing to the web resource).

Silva et al. (2006) describe the GREASE (Geographic REasoning in Search Engines) project in which a prototype system called Geotumba (<http://local.tumba.pt>) has been created to demonstrate the capabilities of geographic web search for a subset of documents written in Portuguese. The architecture developed is modular and incorporates components that include: a Portuguese geographic ontology, a module for extracting geographic entities from Portuguese Web pages and assigning them to entities in the ontology, an indexing and ranking module and a component that deals with user interaction. The approach makes use of a graph ranking method to assign a single scope to each web document.

Not all approaches to geographic search include explicit representation of locations through geocoding. For example, Delboni et al. (2005) reported on the development of a relatively simple system which, through analysis of spatial relationships found in unstructured text documents textually expands queries with spatially-related terms. Graupmann and Schenkel (2006) present a further example of a system, GeoSphereSearch, which does not explicitly represent footprints, but which rather treats documents as a bag of words, maintaining document structure in a graph-based representation which includes an encoding of the type of term (e.g. a location and its coordinates). GeoSphereSearch is reported to be capable of dealing with vague queries by interactive redefinition of an appropriate search radius.

In addition to research projects, a large number of commercial geosearch facilities have been developed by some of the largest commercial web search engine providers (for example Google, Yahoo and Microsoft). Himmelstein (2005) suggests that information on the web can be used to provide local service support and a number of local search services

have been introduced to address this emerging need of web users. Local services like Yellow Pages, Yahoo! Local and Google Local have all been working towards meeting the challenge of providing spatial information to users. At present these facilities provide limited geographic search as they appear to be seeded through Yellow Pages or other business directories, thus limiting their content to entities found in commercial listings. Furthermore they do not have geographic intelligence in interpreting spatial relationships and appear to treat the geometry of all locations as points.

In this paper we set out a description of the design, implementation and evaluation of a complete solution to geographic information retrieval developed in the context of the SPIRIT (Spatially-aware Information Retrieval on the Internet) project (Jones *et al.*, 2002). The implemented search engine utilises a footprint-based solution to the problem of document retrieval for unstructured text such as that commonly found in web documents. Central to SPIRIT's aims were the interpretation of spatial relationships in query formulation, search and the ranking of relevant documents enabling us to test the fundamental hypothesis underlying all of the research described here on geographic information retrieval – namely that search techniques which take explicit account of geographic content and spatial relationships will provide more accurate results than pure text search for queries which include geographic content. Although many of the works described above address a part of this problem, arguably none introduce a holistic and evaluated solution.

The paper is set out as follows. Firstly, the collection of requirements on which the architecture is based is presented, before an introduction to the basic architecture implemented is given. Next, pre-processing steps necessary to identify geographic footprints in unstructured text, and methods for indexing these footprints together with document terms are set out. The components required to for runtime search, including the user interface, core search engine and relevance ranking are then described. An evaluation of the system, firstly examining the accuracy of search results for spatial and non-spatial search in terms of standard IR metrics is then presented, before presentation of some results from a study of the systems overall usability. Finally, the extent to which these results have met the needs specified initially in the paper, and areas where further research is required are discussed.

## **2 Gathering requirements for spatially-aware search**

An early stage in the development of any system's architecture is the collection of requirements from both a system perspective, defining the goals that a system must achieve, and a user's perspective focusing on the needs and context of use of a typical user (Nielsen, 1993).

The basic methodology to analyse the requirements for SPIRIT was two-fold. A set of mock-ups together with scenarios for their use were developed by the project team. These were presented to potential user groups, and semi-structured interviews were used to collect information about users' views on potential interactions and functionalities of the system.

An analysis of existing web-based systems that provided some of the required search and mapping functionalities was also carried out. For instance we analysed the way in which users specify place through textual input and interactive maps in applications such as MultiMap ([www.multimap.com](http://www.multimap.com)). Both the analysis of user's needs and the shortcomings identified in the existing applications resulted in a set of functionalities and characteristics that required innovative solutions from the SPIRIT project.

### ***2.1 Analysis of Requirements***

The requirements gathering process produced a wide ranging set of goals, some of which fitted within the programme of work in SPIRIT, and some which could be considered to be outside the scope of the project for reasons of resource or focus. A core list, which formed both the requirements on which the architecture was based and goals against which the implementation of the system must be measured against is given here. These requirements fall into three broad categories – requirements for the formulation of queries, requirements for the nature of results from a spatially-aware system and requirements for the user interface for such a system.

- The system should support geographical query expansion to allow exhaustive retrieval of relevant documents in a specified area (e.g. a query for documents in Scotland should retrieve documents with respect to Edinburgh, Glasgow, Dundee and other Scottish towns and cities).
- Place names should be automatically identified, and interactively disambiguated.

- Users should be able to query for geographical areas such as Central London whose boundaries are imprecise.
- Spatial concepts relating different geographic entities should be represented, e.g. beaches outside Nice.
- The system should support the use of multiple-place names for the same location, and handle multiple instances in search.
- It should be possible for users to specify the area of interest on a map.
- Users should be able to view query results on a map, and these results should be linked to relevant web documents.
- Documents should be ranked differently with respect to the relevant spatial relation.
- Document ranking should combine both spatial and thematic aspects of document relevance.

These requirements can be decomposed into several specific functionalities which a GIR system must support and forms the basis of the system described below. In summary, the first step in the search process is the specification of a query which takes the form of a triplet containing a thematic component, a geographic (place name) component and some form of spatial relationship that links them, e.g. <theme><relationship><location>. The system must interpret and disambiguate the user's specification of a place name, and provide mechanisms to deal with potentially ambiguous place names (e.g. London, UK vs. London, Ontario) or imprecise regions, such as the south of England (Purves *et al.*, 2005). Having interpreted a query, it must be submitted to a search engine which incorporates

techniques to deal with the thematic and geographic aspects of the query and provide ranked results to a user (Kreveld *et al.*, 2005). Users working through scenarios made clear the importance of displaying results on a map, especially in cases where the spatial relevance of a document with respect to a query is unclear to the user. This typically happens when the user does not have detailed local knowledge of the area under query.

Cartographic representation of the query area and the retrieved document's locations should allow users to make some assessment of the *spatial relevance* of the query results, such as adjacency or containment relationships between a document location and a specific place the user is familiar with. Linking of the *geographic footprints* of the retrieved documents with the content of the documents themselves further allows the user to assess the *thematic* and *geographic* relevance of the query. The possibility of the user using the interface to redefine their search, for example through selection of the most relevant documents or (re)specification of the query region, allows some form of relevance feedback between the user and the system.

### **3. Overview of SPIRIT architecture**

**Insert Figure 1 about here**

A working prototype of the SPIRIT system has been developed, which comprises a number of components responsible for key spatial-awareness functionalities (Figure 1) based on the user requirements expressed above.

SPIRIT has a multimodal interface, allowing both textual and graphical query formulation, together with results presentation and a mapping backdrop. A geographical ontology acts as a repository of knowledge about place names, and relationships between them, for the regions covered by the search engine. Data in this repository are used to recognise the presence of place names in web resources and hence to “ground” the web resources to geographic locations, a process referred to as geo-tagging (Clough, 2005). The geographic ontology is also used to disambiguate place names in a user query and to generate a query footprint that reflects the region of space to which the query refers. The query footprint is a geometric interpretation of the place name and the spatial relationship employed in the query.

The web data collection is a set of geo-tagged web resources that have been extracted from a one terabyte collection of approximately 94 million web pages resulting from a crawl of the web (Joho and Sanderson, 2004).

A metadata component attempts to associate the documents from the original web crawl with one or more footprints representing the regions of space to which individual documents relate. The resulting geo-tagged collection consists of about 900,000 documents that refer to parts of the UK, France, Germany and Switzerland. This is used in generation of a spatial index and in spatially-aware relevance ranking. Spatial indexing supplements text indexing by associating documents with one or more cells of a subdivision of geographic space.

The search engine provides the core information retrieval (IR) functionality of the system, accessing the pre-processed indexes to obtain matches to user queries. The retrieved documents are scored initially only on the basis of textual matching. The geographically-specialised relevance ranking component provides a number of configurable methods for ranking of retrieved web resources. This includes integrating different types of spatial relevance scores with a textual relevance score and a distributed ranking method that avoids clustering of retrieved documents that are very similar to each other. In order to facilitate experimental comparisons, the SPIRIT prototype can also be configured to employ a pure text index for comparison with traditional, non-spatial, search engines.

These components were deployed in a distributed architecture using SOAP (Simple Object Access Protocol), with components existing on remote sites and communicating via defined functional interfaces. This allowed the components to be developed and maintained from individual project consortium partner sites. The components are connected through use of a central, broker component, which acts as a session manager, controlling and scheduling the information flow through the system as well as enabling the recording of all steps involved in processing a query for monitoring and evaluation purposes.

### **Insert Box 1 about here**

Box 1 illustrates formally how for a given document collection,  $D$ , the set of documents  $R$  which are relevant to a query  $Q$  can be returned, given a suitable interface allowing a user

access to the system. In the following, the steps involved in pre-processing the document collection to create searchable indexes, search these indexes and rank the documents returned are detailed, together with the elements of the system dealing with query specification and results presentation.

## **4 Functionality of the Spatially-Aware Search Engine Components**

### ***4.1 Pre-Processing the Document Collection***

The architecture of SPIRIT is based on web documents which have been pre-processed and assigned spatial footprints to be stored within a spatial index. The documents themselves are not altered in this process, but rather metadata describing both terms within documents (the traditional bag of words) and spatial footprints are assigned and represented in text and spatial indexes which support the search functionality of SPIRIT.

#### **4.1.1 Assigning Spatial Footprints to web Documents**

Given a set of web pages, we must identify geographical references and assign them to spatial coordinates (Hill *et al.*, 1999). These two tasks are commonly referred to as *geoparsing* and *geocoding* respectively (Larson, 1996; McCurley, 2001). Geoparsing is performed using the GATE (General Architecture for Text Engineering) Information Extraction (IE) system (Cunningham *et al.*, 2002). Candidate terms are identified using a combination of lists of known locations, organisations and people derived from gazetteers, together with rules which capture elements of the surrounding context. Gazetteer lookup is simple, language-independent and often effective (Mikheev *et al.*, 1999), but alone is unable to identify locations not found in the list and distinguish between locations not used

in a geographical context (e.g. Chicago can represent the US city, the name of a pop group, or the internal name for Windows 95 (McCurley, 2001)). To identify locations in a feasible time, we used a simple gazetteer lookup approach but also applied context rules and additional name lists (proper names and commonly occurring terms) to filter out names in the gazetteer lists most likely used in a non-geographical sense.

For gazetteer lookup, the geo-ontology was used (described in §4.2.2) and was populated with two main sources of data (though only one of these was applicable outside of the UK) – the SABE (Seamless Administrative Boundaries of Europe) dataset and the Ordnance Survey 1:50,000 Scale Gazetteer. These two datasets contain, for the UK, a total of around 270,000 locations of which about 10% are ambiguous (i.e. not unique entries). Locations include regions such as villages, towns, cities, counties and places of interest represented spatially as points and polygons. Based on empirical evaluation (Clough, 2005), our approach using gazetteer lookup with additional context rules gave an accuracy of around 72% and 25% false positives for all annotations found.

When geocoding we can either apply multiple locations to a single reference in these ambiguous cases (called referent ambiguity), or define a *default* location associated with location metadata. A default location can be determined using, for example, the most commonly occurring place (Smith and Mann, 2003), by population of the place name (Rauch *et al.*, 2003) or by semi-automatic extraction from the web (Li *et al.*, 2003). We used a default sense approach and global geographical world knowledge to resolve ambiguity based on features from the geographical resources available to us. Again, based

on the evaluation described in Clough (2005), we were able to ground correctly around 89% of all place names. Locations were assigned an appropriate bounding box, representing a spatial extent derived from polygonal data stored in the geo-ontology, since the overheads associated with passing polygons through the system were too high.

Based on the geo-parsing stage, 885,502 web pages were finally included in the collection used in the SPIRIT prototype. This is less than the total number in the original Terabyte collection because many files either contained no locations, or contained locations which could not be grounded. Table 1 provides some analysis of the SPIRIT collection based on the number of spatial footprints (or Unique IDs (UIDs)) extracted from the processed web pages. The number of UID's indicates the total number of unique locations identified in the collection. For example, 25,841 UID's within the UK were identified within the collection. Within the document collection a total of around 1.5 million references to UK place names were found and, given that around 340,000 pages referred to a UK place name the average number of UID's per document was about four. It is important to note that these grounded UID's include falsely grounded place names (e.g. misidentified instances of place names, such as Jack London) and that the distribution of place names is likely to be strongly biased towards a much smaller subset of locations – the most striking example of this bias is London which occurs in 112477 documents in the collection – that is to say that around 1/3 of the documents in the UK collection have a reference to London (as well as potentially other place names). This result is in itself not surprising, since many administrative organisations in the UK are based in or near London.

Insert Table 1 about here

#### 4.1.2 Building Document Indexes

As the intention of the SPIRIT project was to retrieve text documents with respect to geographic context, the need arose to index on the basis of both text and location. Text indexes are generally maintained as an inverted file structure (Salton & McGill, 1983), and here use was made of the GLASS search engine<sup>1</sup>. Spatial access methods (Samet, 1990; Rigaux *et al.*, 2002) can be used for performing point, range and interval queries. Examples of spatial access methods include regular grids, quadtrees, R-trees, etc. and any one of these could be used in conjunction with the inverted file structure to provide a hybrid indexing structure for addressing spatial queries. For SPIRIT, experiments have been confined to a regular grid-based spatial indexing scheme. This index divides the entire footprint coverage of the document collection into a grid of rows and columns. For each cell of the grid, a list of document IDs was constructed, using the document footprints which resulted from the geo-tagging process. The next section briefly discusses the design choices for combining textual and spatial indexing.

The simplest extension to a text-based IR system is to retrieve all documents that match the concept terms of the query and perform a subsequent filter step that returns only those documents that intersect the geographical scope of the place name in the query (the query footprint). This approach was implemented here using a spatial index of document footprints that is used to match the results of the first step with those documents that

---

<sup>1</sup> For more details see <http://dis.shef.ac.uk/mark/glass/>

intersect the query footprint. The scheme is referred to as “T”. In a more integrated approach, referred to as ST, an individual text index, referring to documents whose footprints lie within the respective grid cell, is associated with each of the cells of the spatial index grid. At search time the cells that intersect the query footprint are determined and then only the corresponding text indexes are searched. The final output is a merged and sorted list of the higher ranked documents from each of the intersecting cells. Because the initial step is essentially spatial, ST is regarded as a space-primary indexing method. The third type of scheme that was implemented is text-primary and referred to as TS. Here the inverted list of the text index is extended so that for each term the associated documents (containing the term) are grouped according to the spatial index cells to which they relate, as determined by intersection of their document footprints with the cell. Thus a form of spatial index is built for each list of documents for each term.

In both the ST and TS approaches, there is a penalty of multiple copies of indexes (multiple text for ST and multiple spatial for TS). The overhead increases with the number of footprints that are selected during geo-tagging and subsequently in the creation of the spatial index. This combined effect of multiple cells and footprints per document constrains the index structure to coarse grid resolutions as the amount of total disk space in storing indexes for our IR system with ST and TS becomes very high. In our experiments to compare ST and TS it was found that TS indexing required slightly more space than ST but exhibited better query response times. In contrast the T scheme resulted in longer query times (up to double) but with very little storage overhead (Vaid et al., 2005).

## ***4.2 Main Components for Run-Time Operation***

### **4.2.1 Query Interface**

The query interface is a bridge connecting users and a software system, the aim of which is to interpret and convey the user's requirements to the system. The SPIRIT query interface encompasses two sub-components, namely a text-based query interface and graphical query interface. These two interfaces provide users with the possibility of interacting with the system in different ways. The text-based interface supports text input in a structured form (e.g. a triplet of <theme><relationship><location>) to implement a query. This interface is a simple extension of that provided by most web search engines and designed to make it easy and fast for a new user of the system to formulate a query. The graphical interface, through the use of a map backdrop and basic interactivity, provides a user unfamiliar with an area with a means of specifying a query (for example, a user who is not familiar with local place names can draw a polygon approximating to a region of interest) and also allows an expert user to specify multiple locations for search (Figure 2).

**Insert Figure 2 about here**

Underlying the query interface is the broker component, which is a middle tier connecting the query interface and other components. The broker component produces and interprets SOAP requests for transmission between the different system components described in §3 and logs all interactions allowing developers to trace bugs in different system components, monitor system usage and allow evaluation experiments.

#### **4.2.2 Geographic Ontology (Query Disambiguation and Expansion)**

To support spatial search, the SPIRIT system employs a geographical ontology (or geo-ontology), which provides knowledge of places within the geographic coverage of the search engine. For each place, the geo-ontology maintains all of the names that a place is known by, the place types with which it can be categorised, the geographical footprints which indicate its spatial extent, and its topological relationships (such as *part\_of* and *containing*) with other places. The actual design and structure of this geo-ontology and issues with populating it are described in more detail elsewhere (Abdelmoty *et al.*, 2005, Fu *et al.*, 2005).

As well as representing places for which the spatial extent is well defined, the geographic ontology also contains a small number of so-called imprecise regions, referring to places such as the British Midlands, the Swiss Mittelland or the American Midwest. The boundaries of these regions were derived by first mining the web for documents mentioning such regions, then geoparsing and georeferencing occurrences of other places names in such documents – based on the assumption that occurrences of place names in documents are likely to be spatially autocorrelated. By performing such operations for many documents, and thus identifying a large pool of candidate locations for the imprecise region, boundaries for the regions could be derived using a variety of techniques (Arampatzis *et al.*, 2006; Purves *et al.*, 2005). The regions themselves are represented in the ontology by a spatial extent and place name, as with all other locations.

The geo-ontology is used both at the pre-processing stage and query time to support spatial search. At pre-processing, the geo-ontology helps primarily in geo-parsing of web documents and organisation of spatial indexes. At query time, the ontology helps disambiguate place names in the query, spatially expand the query, and provide data to enable ranking of retrieved documents. How the geo-ontology helps with geo-parsing, spatial indexing and spatial relevance ranking is elaborated on in other parts of this paper. The following focuses on how the geographical ontology performs spatial query disambiguation and expansion to assist spatial search.

Spatial query disambiguation is necessary because many of the place names which may appear in a query can be shared by multiple places (e.g. there are a number of places named *Newport* in the UK). The geo-ontology helps the user to resolve these ambiguous place names by providing them with the broader spatial contexts of the place, by utilising the *containment* (*part\_of*) relationships encoded between places. For example, for a query involving *Newport*, a user will be prompted to select which *Newport* is intended from a menu of hierarchy information obtained from the geo-ontology:

UK, Wales, Newport

UK, England, Essex, Uttlesford, Newport

UK, England, Leicestershire, Melton, Newport

UK, England, Devon, North Devon, Newport

Spatial query expansion is the process where spatial query terms are expanded to generate a geometric footprint that can be employed by further spatial processing within the SPIRIT prototype. The footprint generated is dependent not only on the footprint of the place name mentioned in the query, but also on the spatial relationship used. A challenge in dealing with spatial query expansion is that a spatial relationship such as *near* is vague. Its interpretation can vary with respect to different users' intentions, as well as depending on the types of spatial and non-spatial terms involved in a query. In Fu *et al.*, (2005) a method was proposed for intelligent treatment of spatial query terms whereby the process of query footprint generation is able to take account of the potential for vagueness in the spatial operator. The range of spatial operators currently supported include *in*, *near*, *outside*, *north*, *south*, *east*, *west* and *within distance of*.

#### **4.2.3 Core Search Engine**

Given the generation of spatial and textual indexes, the core search engine retrieves a set of initial documents in response to a query  $Q$  sent from the user interface. In practise, thematic terms are first stemmed by a Porter stemmer (Porter, 1980), and stopwords (e.g., of, the, he, that) are removed. The retrieval of documents then involves matching terms in the query to the document collection and ranking of the matched documents. The initial ranking of matched documents in the system is based on a probabilistic IR model using the BM25 weighting function (Robertson, *et al.* 1998). BM25 is based on three sources of weighting which have been shown empirically to be useful for different retrieval tasks (Sparck Jones and Willet, 1997):

- **Document frequency** - terms occurring in only a few documents are likely to be more useful than terms appearing in many.
- **Term frequency** - the more frequently a term appears in a document the more important it is likely to be for that document.
- **Document length** - a term occurring the same number of times in a short document than a longer one is likely to be more important in the shorter one.

A set of candidate documents  $R$  are retrieved by the intersection of the set of documents which lie within the query footprint and those which contain the thematic terms matched as described above. The particular order of these calculations depends on whether T, ST or TS indexing is used. Having retrieved a candidate set of up to 1,000 documents with the core search engine, these are forwarded to the relevance ranking component for geographic relevance-based re-ranking.

#### **4.2.4 Relevance Ranking**

Web documents identified by the search engine are ranked according to both textual and spatial relevance. From a geo-spatial perspective, each document in the web document collection is represented by a bag of footprints following the grounding of web locations to places, and a query is also represented as a footprint. Depending on the spatial relationship used in the original query, different formulae are used to calculate footprint similarity scores between query and document footprints. When all footprints in a document are assigned a similarity score with respect to the query footprint, a *document spatial similarity score* for the document can be calculated. The relevance ranking component combines the

spatial and textual document scores to generate a single ranking. A number of different options for score combination have been investigated and compared with purely textual scoring (solely BM25).

Thus, in order to produce a relevance ranking of documents with respect to a query, the following steps are taken:

- For every document footprint, a footprint similarity score is produced with respect to the query footprint and connector.
- For every document, a document spatial similarity score is produced based on the footprint similarity scores of all the footprints contained in the document.
- Document spatial similarity scores are usually combined with textual BM25 scores into a document similarity score.
- Documents are ranked in descending order of their document similarity scores.

The formula used to calculate footprint similarity scores is dependent on the spatial relationship in the original query, i.e.

- **inside:** Binary operator defined between a query's bounding-box and a document's footprint (MBR or centroid). Coordinates are checked for containment.

- **near:**  $\text{near}(a,b) = \exp(-L * D(a,b))$ , where  $a$  and  $b$  are the centroids of a query's and a document's footprint,  $D(a,b)$  is their Euclidian distance. Thus, proximity scores decay exponentially from 1 to 0 with increasing distance.  $L$  controls the rate of decay, or, in real-world terms, “how far is far”, and can for example be a function of the query footprint – thus, for example, near things can be further from large objects.
- **north-of, south-of, east-of, west-of:** Assuming that  $a$  and  $b$  are the centroids of a document footprint and a query footprint respectively, and that  $\psi$  is the angle of the vector  $ba$  from the positive x-axis with the origin assumed on point  $b$ . For north-of, if  $\psi \geq 180$  or  $\psi \leq 0$  then  $\text{north-of}(a,b) = 0$ , otherwise  $\text{north-of}(a,b) = 1 - |\psi - 90|/90$ . The other directional operators are calculated in a similar manner. Proximity is also taken into account, so to obtain the final score they are multiplied with  $\text{near}(a,b)$ .

For the SPIRIT prototype the “best-match” approach is followed to determine document spatial relevance, i.e. a document's spatial similarity score is the highest footprint similarity score of the footprints it contains. Initial experiments with a few other approaches that use partial score contributions from all document's footprints did not appear promising; they tend to reduce effectiveness returning dubious results for some queries.

Before multiple textual and spatial scores are combined, they first have to be normalized into the range [0,1]. BM25 scores can be quite unpredictable in their range and they are currently normalized linearly by dividing with the highest document score for the query.

**Insert Figure 3 about here**

Figure 3 depicts 8 documents which each have both a spatial and a BM25 score for which a number of methods have been trialled to combine these two scores into one. The non-distributed method ranks the documents in ascending order of their Euclidian distance from point (1,1) that is assumed to be the most relevant possible document. The distributed method tries to de-cluster documents that have almost the same score components. In Figure 3 ranking is according to a distributed method. Thus, note that although 4 is further from (1,1) it is ranked above 5. Two variations of the distributed method exist, one based on the angle of a yet un-ranked point to the already ranked, and one based on its distance from the already ranked. The exact details of the algorithms are reported elsewhere (Kreveld *et al.*, 2005).

#### **4.2.5 Results Display**

The results returned from the relevance ranking component are a ranked collection of pointers to web documents, a variety of ranking scores and a set of document footprints. The display component of the search engine uses this information to ease the user's task in sorting relevant from irrelevant data. We display search results using a map backdrop generated from standard web mapping components. The results interface only displays

document footprints which have been identified as being relevant to the query in question - that is not every document footprint from every document is displayed. Locations are displayed as points on the map, enhanced through the use of brushing and linking, allowing the user to quickly and simply associate web documents with locations. Footprints which are shared between documents are displayed as stacks, with all documents related to a stack revealed through mouseover.

A further issue in dealing with the results of any search engine, including a spatially-aware one, is to provide users with a means to visualise the potentially very large numbers of documents returned. In SPIRIT, we provide users with a simple, map-based interface on which only the ten documents are displayed at a time (Figure 4) and a set of more complex set of visualisations which allow users to view a variety of representations of larger numbers of document (Figure 5). These include the use of density surfaces, cartograms and a filtering mechanism based on footprint size and the number of shared footprints at a location (Yang *et al.*, 2006).

**Insert Figure 4 about here**

**Insert Figure 5 about here**

## 5 Evaluation

In this paper we concentrate on the holistic evaluation of the SPIRIT prototype - the majority of individual components have also been evaluated separately (e.g. Clough, 2005 and Vaid *et al.*, 2005).

To date, few GIR systems based on querying of unstructured text exist, and thus, to our knowledge, limited evaluation has so far been performed of such systems (although various proposals have been made for the evaluation of specific components of systems (e.g. Martins *et al.*, 2005; Leidner, 2006). The GeoCLEF evaluation campaign (Gey *et al.*, 2005) addresses some of the issues concerning system evaluation. By contrast, in IR a long tradition of evaluation exists. In most cases, an IR system is used to assist with finding answers (in the form of documents) to a user's information needs (Mizzaro, 1997). How well a system meets those needs can be evaluated along a number of dimensions. For example, two predominant evaluation strategies have emerged from IR evaluation: those which are system-focused; and those which are user-centred (Spark Jones and Willett, 1997). In the former, the goal of evaluation is to measure system accuracy, e.g. to compare and rank different IR systems or components of the same IR system. This approach typically uses an established benchmark to simulate retrieval tasks without requiring involvement of end users. The latter strategy - user-centred evaluation – aims to evaluate IR systems with respect to usability (e.g. assessing the suitability of an IR system interface or some feature of interface design through a task-based user study), requiring that the IR

system has an interface through which users can interact and upon which observations can be made.

In this paper we present results from both system and user-centred evaluations. Our initial approach to performing system-centred evaluation was based around standard IR notions and involved the creation of a *test collection* which provides the necessary resources and framework in which to assess the system (Bucher *et al.*, 2005). A typical IR test collection consists of the following: a set of documents representative of a selected domain, a set of typical user information needs based on the document collection (queries) and a list of which documents are relevant to each query. The document collection should fairly represent the search domain and the topics should balance between representing realistic user requests and providing controlled queries (Peters & Braschler, 2001), and relevance assessments should be as complete as possible (which can be difficult in large document collections). However, we encountered significant difficulties with this approach. Crucially, we considered that document relevance should be measured with respect to two dimensions – spatial and thematic relevance. Experiments revealed that thematic relevance (i.e. is this document about castles) was easier to assess than spatial relevance (i.e. is this document about somewhere to the east of Edinburgh). A description of some of our experiments with different relevance schemes and the importance of spatial and thematic dimensions of relevance can be found in Clough *et al.* (2006) and Purves and Clough (2006). These difficulties in assessing spatial relevance meant that when we performed automated tests on a set of queries, many documents appeared whose relevance had not been judged.

Our second approach to system-centred relevance judgement was to assess the relevance of the top ranked documents retrieved by spatially aware search, and documents retrieved from the same document collection using purely textual search with BM25 ranking. For this second case we concatenated the triplet of <theme><spatial relationship><location> into a query which was submitted to the search engine.

The spatial and thematic relevance of the top ten documents for spatial and text search were then assessed independently by two assessors. This approach allowed us to measure normalised precision (the proportion of relevant documents returned where the maximum number of documents returned was ten per search). For spatial search, the experiments described here are for text primary indices which provided a more rapid query response (§4.1.2), and non-distributed ranking (§4.2.4), which appeared in preliminary experiments to give best results. Although assessing the accuracy of the system, particularly with other ranking methods would be of interest, in practice it was not feasible to both judge large numbers of queries and do this for multiple ranking methods.

Judgements were made for a total of 38 queries covering the range of spatial relationships handled by SPIRIT and with locations of differing granularities. The queries and values of normalised precision are shown in Table 2. Thematic and spatial relevance were assessed on a binary scale, where thematic relevance was defined as a document whose had some significant relevance to the theme and to be spatial relevant a document had to have a footprint which was considered to be similar to the query footprint, and importantly, not of a much coarser granularity. Thus, for example, a document about distilleries in Scotland

would not be considered relevant to a query for distilleries in the north of Scotland. Interannotator agreement, in terms of the number of judgements for which the two annotators agreed, was 82% (BM25) and 81% (Spatial) for thematic relevance and 84% (BM25) and 74% (Spatial) for spatial relevance. It is important to note, that particularly in the case of spatial relevance for BM25 search, the high inter-annotator agreement is partially due to the very low numbers of spatially relevant documents.

In order for a document to be considered relevant to the query, it had to be both thematically and spatially relevant. Thus, the most important value in Table 2 is the **combined relevance**. In general, it is clear from Table 2 that combined relevance values are lower for BM25 than spatial search, and furthermore T-tests showed that for both annotators the distributions of combined relevance are significantly different ( $p < 0.001$  in both cases). Overall, for both annotators combined relevance is better for spatial search for 27 out of 38 queries, and BM25 outperforms spatial search for only 7 and 4 queries for annotators 1 and 2 respectively.

**Insert Table 2 about here**

In Table 3, summary statistics for the four different spatial relationship types for BM25 and spatial search are presented. In general, it is clear that spatial search returns more relevant results than BM25 search for all categories of relevance. Furthermore, the difference in relevance is considerably more marked for spatial relationships other than inside. This suggests that a underlying hypothesis - that pure text search cannot deal with the semantics

relating spatial relationships to themes - is correct. The poorer performance of BM25 search in terms of thematic relevance is however, at first glance, surprising. We believe this is a result of the concatenation of the query, which in turn will result in a downweighting of the thematic component of the query in the ranking (as it is no longer the only term in use in the textual ranking).

The key result here is that, spatially-aware search out performed text-only search. However, care is required in considering this result. Mean normalised precision is never more than 0.4 and thus the system's overall precision could be considered as rather low. The primary reason is simple – the number of georeferenced documents is relatively small – the SPIRIT collection for the whole of the UK, for example, consists of only around 340,000 documents. Further, as discussed in Clough (2005) not all documents are correctly georeferenced (around 90% in the experiments reported by Clough).

In the SPIRIT project user-centred evaluation focussed on assessing the more general usability of the system. This evaluation was carried out as part of a larger study of the success of a range of different visualisation methods for query results, such as maps, cartograms and density surfaces (Yang *et al.*, 2006) which will be reported in a forthcoming paper. A range of questions, focussed on the interface, but also providing some information about the system overall were posed to 50 participants in this evaluation. Participants used the SPIRIT system to explore a number of queries which were designed to assess some of the different system capabilities identified in the requirements analysis (e.g. geographical query expansion, place name identification and the use of spatial

relationships). The data gathered includes information on the usability and performance of the system, thus allowing a user-centred assessment of document relevance to be made. Figure 6 shows the answers to a subset of questions from this study, illustrating aspects of the usability of the system in terms of basic usability, query formulation, the use of spatial relationships and the ability of users to associate documents with locations on the map as shown in Figure 4. These results show that most users in this sample were happy with the scope of the spatial relationships presented and had no problems in formulating a query.

**Insert Figure 6 about here**

As well as quantitative data on the usability and effectiveness of SPIRIT, qualitative information about user experiences was gathered during the evaluation process. Perhaps the most commonly repeated remark from users concerned the speed of the system – many users commented on the sluggishness of the system, particularly in comparison to commercial search engines such as Google. However, it was also apparent that users were stimulated to experiment in building relatively complex queries which could not be served by current commercially available search engines, although because of the small sample dataset available in SPIRIT, these queries were not always successful.

## **6. Discussion**

This paper sets out the architecture for a geographic information retrieval system, and reports on the components necessary to pre-process documents for spatial search and

further, to index and retrieve these documents using both spatial and geographic information. The architecture itself was set out in the context of a set of requirements for spatially-aware search and we now consider how well these requirements have been met.

The geographical ontology is a central component of the SPIRIT search engine, and was crucial in meeting several of the requirements set out. Through the user interface, users are able to identify place names and interactively disambiguate such place names where more than one instance of the name exists. Disambiguation is further assisted by the use of graphical query techniques. This is particularly important where users are unfamiliar with a region, since textual disambiguation assumes that users are familiar with administrative names within regions. Since the geographic ontology contains knowledge of the relationships between locations, a disambiguation approach based not on the administrative region to which an area belongs, but rather well-known “nearby” locations may be worthy of investigation. A similar approach has been shown by Naaman *et al.* (2006) to be an effective in assigning textual names to images. Furthermore, methods were developed here to enhance the ontology with imprecise regions that might be employed in natural language.

The results of system-centred and user-centred evaluations show that the implemented architecture successfully expands queries and locates documents which do not mention the place name specified in the query. This functionality depends upon a combination of spatial and textual indexing methods. Initial experiments found that a text-primary indexing scheme gave the best results but at the cost of high storage overheads. Response times were

not at the levels which users' experiences of commercial search engines leads them to expect. The main reason for this is the fact that the SPIRIT prototype uses disk-based indexing methods in contrast to the very much faster main memory methods of commercial systems. It may well be that the combination of spatial and textual indexing that gave the lowest storage overheads (method "T") could be viable in a main memory indexing environment. The primary reason for overheads in data volume of the integrated spatial and textual indexing methods is that many documents have multiple references to locations, with an average of around 4 locations per document in the UK. Moreover, the relatively low precision of the system is partly attributable to many of these locations not being relevant to the document's theme, or having a very coarse granularity which is not relevant to many geographically specific queries. Further work on associating "dominant" locations with web pages (e.g. Amitay *et al.*, 2004; Wang *et al.*, 2005) should reduce the number of terms referenced per web page, thus increasing precision and reducing the size of spatial indexes with a corresponding improvement in performance times.

The system, as implemented, supports a number of spatial relationships through a simple structured interface. These relationships are used to generate a query footprint based on the geometry of the query location and the relationship in question. The paper illustrates a number of queries which make use of such spatial relationships. The user study showed that users were both happy with the list of relationships provided and confident that they understood the purpose of the different elements of the structured query.

The system-centred evaluation concentrated on comparing the normalised precision of spatial search with pure BM25 search. Spatially-aware ranking methods were shown to have better precision than for pure text search, where no query expansion through geometry takes place. Only a single relevance ranking scheme (non-distributed ranking) was tested. However, since the aim of the different ranking methods is to return more useful *collections* of documents – and the usefulness of a collection is not measured by precision, measure other than precision would be required to carry out a useful evaluation of ranking methods. Ongoing research will attempt a qualitative assessment of such properties. As has been shown in IR, such test collections can only be built by large cooperative projects such as the Text REtrieval Conference (Voorhees & Harman, 2000), and the emergence of initiatives which have some parallels, such as Geo-CLEF (Gey *et al.*, 2005) is therefore to be welcomed.

The main reason for the relatively small sub-sample of queries and documents used in the system-centred part of the evaluation lies in the difficulty of building a test collection for GIR. As is well recognised in the information retrieval community, the building of test collections is a time consuming task. Furthermore, our research (Bucher *et al.*, 2005; Clough *et al.*, 2006; Purves and Clough, 2006) suggests that evaluating geographic relevance is more difficult than thematic relevance – in other words it is easier for example to say if a document is about castles than if it is about a castle that is near Leeds without local knowledge of Leeds. Further work is ongoing to collect larger test collections suitable for the evaluation of GIR systems and forms an important component of the research necessary to improve quantitative evaluation of such systems.

Finally, the user requirements stressed the importance of providing a map-based interface. User experiments reinforced this conclusion, with the majority of users happy with the interface functionality and agreeing that the display of results on a map made it easier to judge the relevance of retrieved documents.

## **6. Conclusions and Further Work**

This paper describes a unified approach for introducing spatial-awareness into search engine technology and the architecture used to enable this. A prototype system has demonstrated the effectiveness of a strategy based on indexing and querying spatial footprints found in web documents. This system provides a complete solution to the problem of geographical IR, from the processing of web resources for spatially-aware search, to all the components necessary for managing this type of information at run-time. The processing of web resources is key to the success of this approach and involves a preliminary stage of analysis to identify geographical content. This process of geographical categorisation of documents supports many of the essential aspects of the spatially-aware search engine. In particular, it allows documents to be indexed spatially as well as thematically which in turn enables a full set of geographical query operators, graphical query formulation, the ranking of results according to conceptual as well as spatial match to the original query, and the graphical display of search results.

The analysis and use of geographical content from web resources is currently an area of increased interest and research (from both the academic and commercial sectors). Given that a large proportion of what people do is based around location, and that many web resources contain some kind of geographical context, then the potential benefits of projects such as SPIRIT are obvious. Exploiting geography to enhance the user's experience for web search has enormous potential, but most research has shown that the handling and processing of web content is non-trivial and requires much more attention (especially if the objectives of the semantic web are ever to be realised). The SPIRIT project has demonstrated that retrieval from web resources can be improved by making search spatially-aware. Although several projects have addressed this, SPIRIT is unique in that the whole lifecycle of providing geographic retrieval has been addressed from finding out what users actually want from such systems (and using these to influence and guide the design), to considering user interaction and more importantly, to thinking about how we can evaluate such systems (which is certainly different from how we evaluate traditional IR systems).

In the future, there are a number of areas which we plan to explore to further improve the system. This includes improving the geotagging stages through the use of adaptive information extraction techniques to learn *automatically* context rules for identifying potential locations (e.g. using linguistic clues, punctuation, HTML tags, hyperlinks and other named entities), rather than using rules which are defined by hand. We plan to explore the assignment of a geographic scope to web resources, rather than using individual scopes for all locations found in a web page. We believe this will help improve the

precision of the search results. Further work is also required to evaluate both effectiveness of different relevant spatial ranking methods both in terms of metrics such as precision and their usability in a more holistic sense (for example, are a set of results as returned by distributed ranking more different from one another, and thus more interesting to a user, than results ranked by using non-distributed ranking as evaluated in this paper. Equally, work is ongoing to evaluate the effectiveness of a range of techniques to visualise large document sets resulting from geographic search.

## **Acknowledgements**

This research was supported by the EU-IST Project No. IST-2001-35047 (SPIRIT) and the Swiss BBW (01.0501). We would like to thank all those who took part in requirements and usability studies, and all members of the SPIRIT consortium for a constructive and profitable relationship. The constructive comments of two anonymous referees helped to significantly improve this paper.

## **References**

- Abdelmoty, A.I., Smart, P.D., Jones, C.B., Fu, G. & Finch, D., 2005, A critical evaluation of ontology languages for geographic information retrieval on the Internet. *Journal of Visual Languages & Computing*, **16**(4), pp. 331-358.
- Amitay, E., Har'El, N., Sivan, R., and Soffer, A., 2004, Web-a-where: geotagging web content, In Proceedings of the 27th annual international conference on Research and development in information retrieval (SIGIR04), Sheffield UK, 2004, pp. 273-280.
- Arampatzis, A., van Kreveld, M., Reinbacher, I., Jones, C. B., Vaid, S., Clough, P., Joho,

- H., & Sanderson, M., 2006, Web-based delineation of imprecise regions. *Computers, Environment and Urban Systems*, 436-459.
- Berners-Lee, T., Hedler, J. & Miller, E., 2001, The Semantic Web. *Scientific American*, pp .35-49.
- Bucher, B., Clough, P., Joho, H., Purves, R.S., & Syed, A. K., 2005, Geographic IR Systems: Requirements and Evaluation. In: *Proceedings of the 22nd International Cartographic Conference*, A Coruña, Spain, CD-ROM.
- Buyukokkten, O., Cho, J., Garcia-Molina, H., Gravano, L. & Shivakumar, N., 1999, Exploiting Geographical Location Information of Web Pages. In *Proceedings of Workshop on Web Databases (WebDB'99)*. ACM Press.
- Clough, P., 2005, Extracting Metadata for Spatially-Aware Information Retrieval on the Internet. In *Proceedings of the 2005 Workshop on Geographic Information Retrieval*, Bremen, Germany. pp. 25-30.
- Clough, P.D., Joho, H. and Purves, R.S., 2006, Judging the Spatial Relevance of Documents for GIR, In *Proceedings of the 28th European Conference on IR Research (ECIR'06)*, London, UK, April 2006, Springer-Verlag LNCS Volume 3936, pp. 548-552.
- Cunningham, H., Maynard, D., Bontcheva, K. & Tablan, V., 2002, GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*. Philadelphia, July 2002.
- Delboni, T.M., Borges K.A.V. and Laender A.F., 2005, Geographic Web Search based on Positioning Expressions. In *Proceedings of the 2005 Workshop on Geographic Information*

Retrieval, Bremen, Germany, pp. 61-64.

Ding, J., Gravano, L. & Shivakumar, N., 2000, Computing Geographical Scopes of Web Resources. In: *Proceedings of the 26<sup>th</sup> International Conference on Very Large Databases (VLDB '00)*, pp. 546-556.

Egenhofer, M.J., 2002, Toward the Semantic Geospatial Web. In: *Proceedings of the 10<sup>th</sup> ACM International Symposium In Geographic Information Systems*, ACM Press, pp. 1-4.

Fu, G., Jones, C.B. & Abdelmoty, A.I., 2005, Building a Geographical Ontology for Intelligent Spatial Retrieval on the Web. In *Proceedings of IASTED International Conference on Databases and Applications (DBA2005)*, Innsbruck, Austria.

Gey, F., Larson, R., Sanderson, M., Joho, H., and Clough, P., 2005, GeoCLEF: the CLEF 2005 Cross-Language Geographic Information Retrieval Track. In: *Working Notes for the Workshop on the Cross Language Evaluation Forum 2005*, CD-ROM.

Graupmann, J. and Schenkel, R., 2006, GeoSphereSearch: Context-Aware Geographic Web Search. In *Proceedings of the 2006 Workshop on Geographic Information Retrieval*, Seattle, USA, pp. 64-67.

Hill, L.L., Frew, J., and Zheng, Q., 1999, Geographic Names. The implementation of a gazetteer in a georeferenced digital library. *Digital Library*. 5(1).

Himmelstein, M., 2005, Local Search: The Internet Is the Yellow Pages, *IEEE Computer Society Journal*, 0018-9162/05, pp.26-34.

Joho, H. & Sanderson, M., 2004, The SPIRIT collection: an overview of a large web collection.” *SIGIR Forum*, **38**(2), pp.57-61.

Jones, C.B., Purves, R.S., Ruas, A., Sanderson, M., Sester, M., van Kreveld, M. & Weibel,

- R., 2002, Spatial Information Retrieval and Geographical Ontologies: An Overview of the SPIRIT Project. In: *Proceedings of SIGIR-02, the 25<sup>th</sup> Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM Press, pp.387-388.
- Kreveld, M. van, Reinbacher, I., Arampatzis, A. & Zwol, R. van., 2005, Multi-Dimensional Scattered Ranking Methods for Geographic Information Retrieval. *Geoinformatica*, **9**(1), pp.61-84.
- Larson, R., 1996, "Geographic Information Retrieval and Spatial Browsing." In Smith, L. & Gluck, M. (eds.) *GIS and Libraries: Patrons, Maps and Spatial Information*. University of Illinois.
- Lee, R., Shiina, H., Takakura, H. & Kambayashi, Y., 2003, Map-based Web Indexing and range Query Processing for Geographic Web Search Systems. In *Proceedings of International Conference on Internet Information Retrieval 2003*, Koyang, Korea.
- Leidner, J.L., 2006, An evaluation dataset for the toponym resolution task. *Computers, Environment and Urban Systems*, *30*(4) pp. 400-417.
- Li, H., Srihari, K.R., Niu, C. & Nli, W., 2003, InfoXtract Location Normalization: a Hybrid Approach to Geographic References in Information Extraction. In: Kornai, A. & Sundheim, B. (eds.), *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References*, Alberta, Canada, pp.39-44. ACL.
- Markowetz, A., Brinkhoff, T. & Seeger, B., 2004, Exploiting the Internet As a Geospatial Database. In Post-Workshop Book of *International Workshop on Next Generation Geospatial Information*, Cambridge, MA, 2003.

- Martins, B., Silva, M.J. & Chaves, M.S., 2005, Challenges and resources for evaluating geographical IR. In *Proceedings of the 2005 Workshop on Geographic Information Retrieval*, Bremen, Germany. pp. 65 – 69.
- McCurley, S.K., 2001, Geospatial Mapping and Navigation of the Web. In *Proceedings of the 10<sup>th</sup> International WWW Conference*, Hong Kong, 1-5 May, ACM Press, pp.221-229.
- Mikheev A., Moens M. & 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, pp.1-8.
- Mizzaro, S., 1997, Relevance: The whole history. *Journal of the American Society for Information Science*, **48**(9), pp. 810-832.
- Naaman, M., Song, Y.J., Paepcke, A. & Garcia-Molina, H., 2006, Assigning Textual Names to Sets of Geographic Coordinates. *Computers, Environment, and Urban Systems*. pp. 418-435.
- Nielsen, J., 1993, *Usability Engineering*. Academic Press, Boston.
- Peters, C. & Braschler, M., 2001, Cross-language system evaluation: The CLEF campaigns. *Journal of the American Society for Information Science and Technology*, **52**(12), pp. 1067-1072.
- Porter, M.F., 1980, An Algorithm for Suffix Stripping. *Program*, **14**(3), pp.130-137.
- Purves, R.S. and Clough, P. (2006), Judging spatial relevance and document location for Geographic Information Retrieval, extended abstract, In *Proceedings of 4th International Conference on Geographic Information Science (GIScience 2006)*, Münster, Germany, September 2006, pp. 159-164.

Purves, R.S., Clough, P. and Joho, H., 2005, Identifying imprecise regions for geographic information retrieval using the web. In: Billen, R., Drummond, J., Forrest, D., and João, E. (eds), *Proceedings of the GIS RESEARCH UK 13th Annual Conference*, 313-318, Glasgow, UK.

Rauch, E., Bukatin, M. & Baker, K., 2003, A Confidence-Based Framework for Disambiguating Geographic Terms. In: Kornai, A. & Sundheim, B. (eds.), *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References*, Alberta, Canada, pp.50-54. ACL.

Rigaux, P., Scholl, M. & Voisard, A., 2002, *Spatial Databases With Application to GIS*. Morgan Kaufmann.

Robertson, S.E., Walker, S. & Beaulieu, M., 1998, Okapi at TREC-7: Automatic ad hoc, Filtering, VLC and Interactive Track. In: Voorheer, E.M. & Harman, D.K. (eds.), *NIST Special Publication 500-242: The 7<sup>th</sup> Text Retrieval Conference (TREC-7)*, Gaithersburg, MD, pp.253-264. NIST, Gaithersburg, MD.

Salton, G. & McGill, M., 1983, *Introduction to Modern Information Retrieval*. McGraw-Hill.

Samet, H., 1990, *Applications of Spatial Data Structures*. Addison-Wesley.

Silva, M., Martins, B., Chaves, M., Cardoso, N. and Afonso, A. P., 2006, Adding Geographic Scopes to Web Resources , *Computers, Environment and Urban Systems*, Elsevier Science, Volume 30, pp 378-399.

Smith, D. A. & Mann, G. S., 2003, Bootstrapping Toponym Classifiers.” In: Kornai, A. & Sundheim, B. (eds.), *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of*

*Geographic References*, Alberta, Canada, pp.45-49. ACL.

Sparck Jones, K. & Willett, P. (Eds.), 1997, *Readings in Information Retrieval*. Morgan Kaufmann, San Francisco, USA.

Vaid, S., Jones, C. B., Joho, H., and Sanderson, M., 2005, Spatio-Textual Indexing for Geographical Search on the Web. In: *Proceedings of the 9th International Symposium on Spatial and Temporal Databases*, 218-235, Angra dos Reis, Brazil.

Voorhees, E.M. and Harman, D., 2000, Overview of the sixth Text REtrieval Conference (TREC-6). *Information Processing and Management*, **36**(1), pp.3-35.

Wang, C., Xie, X. Wang, L., Lu Y. & Ma, W., 2005, Detecting geographic locations from web resources. In *Proceedings of the 2005 Workshop on Geographic Information Retrieval*, Bremen, Germany, pp. 17 – 24.

Watters, C. & Amoudi, G., 2002, GeoSearcher: Location-Based Ranking of Search Engine Results. *Journal of the American Society for Information Science and Technology*, **54**(2), pp. 140-151.

Yang, B., Purves, R.S., Syed, A.K. and Weibel, R., 2006, Web-based visualisation tools for spatial information retrieval. In: Preistnall, G. and Aplin, P. (eds), *Proceedings of the GIS RESEARCH UK 14th Annual Conference*, 263-267, Nottingham, UK.

Zhang, V.W. Rey, B. Stipp. E. and Jones, R., 2006, Geomodification in Query Rewriting. In *Proceedings of the 2006 Workshop on Geographic Information Retrieval*, Seattle, USA, pp. 23-27.

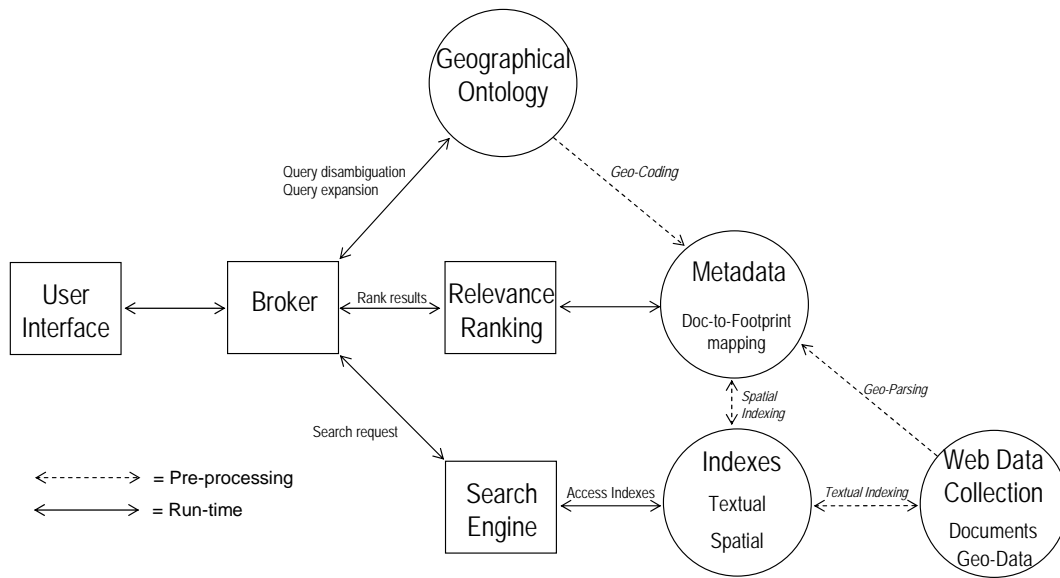



Figure 1: SPIRIT architecture showing run-time and pre-processing components and linkages



[Help](#) [User tests questionnaire](#) [About SPIRIT](#)

### Structured Query

This is a prototype spatially-aware web search engine accessing an experimental collection of web pages. It provides access to only about 900,000 geo-tagged web pages relating to parts of the UK, Germany, France and Switzerland. Coverage is best around a small number of cities especially Cardiff, Edinburgh and Zurich.

Search for

castles

Near

▼

newport

United Kingdom ▼

Town or City name

Region name

Country name

search

2005 Spirit



[Help](#) [About SPIRIT](#)

### Graphical Query

Select Country for Search

United Kingdom ▼

Choose theme to search

cottages

In

▼

Help on searching

1) Choose a country

Select Country for Search

United Kingdom ▼

2) Choose a theme, e.g. castles...

Choose theme to search

castles

3) Choose a spatial relationship from the list, e.g. in, near, north of...

In

▼

4a) To search the whole area shown on the map, click search.

search

4b) To search a smaller area of the map, click on the polygon button and then select an area to query.

Map View

Document Density

Footprint Selection

Cartogram



Figure 2: SPIRIT textual query interface showing a structured query with triplet of <theme><spatial relationship><location> and graphical query interface where the region is graphically specified on a background map

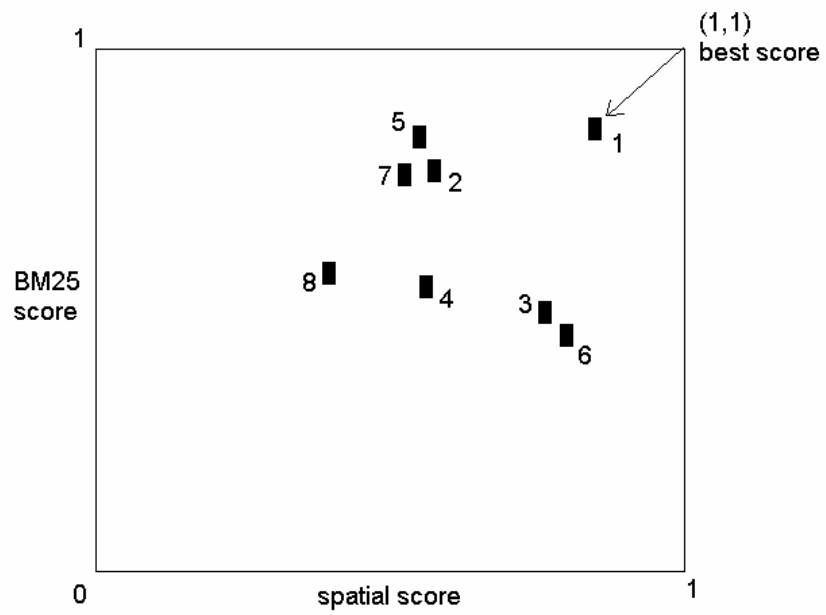
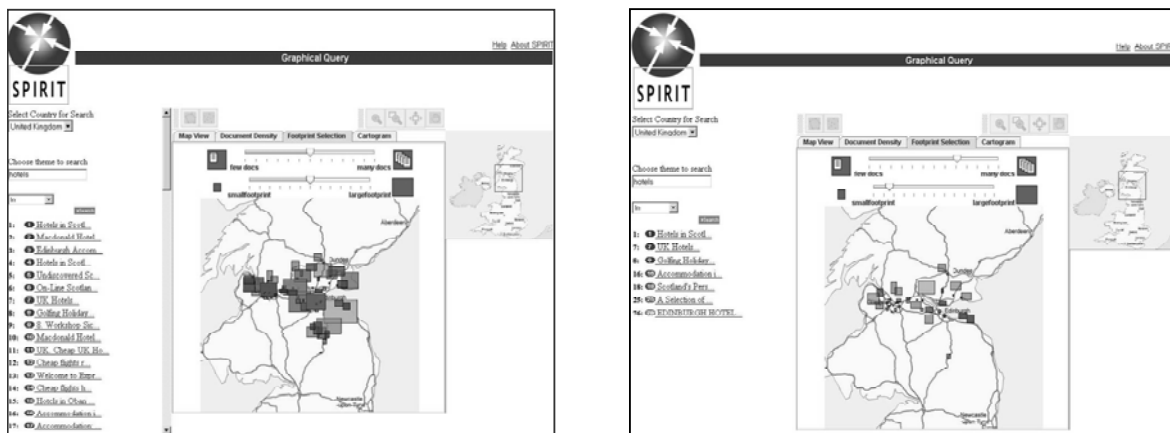


Figure 3: Plot of BM25 (textual) against spatial document scores and associated ranking for a distributed method



Figure 4: SPIRIT results display for a structured query – elements on the map and in the documents list are linked and locations with multiple documents are displayed as columns – with a list of documents at the location displayed on mouseover



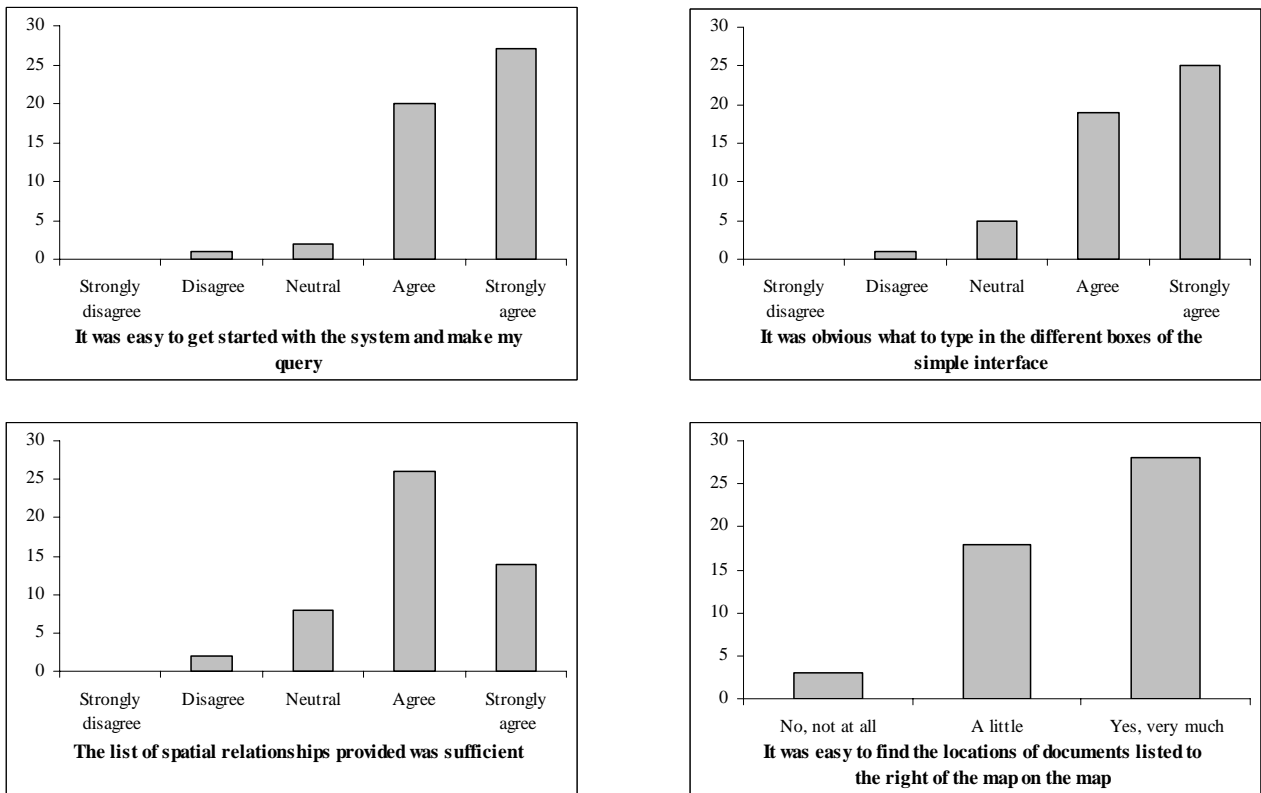


Figure 6: Selected results from a usability of SPIRIT focussing on the ability of users to understand the options presented and relate the results displayed on the map to the list of documents displayed – these results are for simple query interface (Figure 4)

For a document collection  $D$  there exist  $n$  documents  $d_i$  such that

$$d_i \in D$$

and each document also contains terms  $t^{d_i}$ , such that

$$t^{d_i} = \{t_1, \dots, t_m\}$$

where  $m$  is the number of terms in document  $i$ .

Each document  $d_i$  also has footprints  $f^{d_i}$  where

$$f^{d_i} = \{f_1, \dots, f_j\}$$

i.e. where  $j$  is the total number of footprints in a document, which may be 0.

Now, given a query  $Q$ , we define our query as  $Q = \{T, F^Q\}$  where  $T$  is the thematic element of the query

$$t_k \in T$$

i.e. where  $t_k$  are individual thematic query terms and  $F^Q$  describes the spatial extent of the query footprint derived from the spatial relationship and location, then the document set  $R$  returned will be defined as follows :

$$R = \{d_l \in D, \{t_k \in d_l\} \wedge (F^Q \cap f_k), f_k \in f^{d_l}\}$$

Box 1: Set theoretic description of footprint-based spatial search

	# Documents	Unique UIDs	Unique UID occurrences	Avg. UIDs per document
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 1: Summary of footprints (UIDs) identified by geoparsing for SPIRIT collection

Query	Annotator 1, BM25			Annotator 2, BM25			Annotator 1, Spatial			Annotator 2, Spatial		
	Th (P <sub>n</sub> )	Sp (P <sub>n</sub> )	Com. (P <sub>n</sub> )	Th (P <sub>n</sub> )	Sp (P <sub>n</sub> )	Com. (P <sub>n</sub> )	Th (P <sub>n</sub> )	Sp (P <sub>n</sub> )	Com. (P <sub>n</sub> )	Th (P <sub>n</sub> )	Sp (P <sub>n</sub> )	Com. (P <sub>n</sub> )
<beaches><in><east lothian>	0.40	0.40	<b>0.20</b>	0.40	0.30	<b>0.20</b>	0.67	0.00	<b>0.00</b>	0.67	0.00	<b>0.00</b>
<camping><in><highland>	0.80	0.20	<b>0.10</b>	0.50	0.10	<b>0.10</b>	0.50	0.30	<b>0.20</b>	0.20	0.10	<b>0.10</b>
<mountaineering><in><scotland>	0.60	0.60	<b>0.50</b>	0.40	0.40	<b>0.30</b>	1.00	0.90	<b>0.90</b>	1.00	0.90	<b>0.90</b>
<oil industry><in><aberdeen>	0.70	0.40	<b>0.30</b>	0.60	0.40	<b>0.30</b>	0.80	0.40	<b>0.20</b>	0.60	0.30	<b>0.30</b>
<pubs><in><edinburgh>	0.80	0.80	<b>0.70</b>	0.30	0.60	<b>0.30</b>	0.50	0.70	<b>0.50</b>	0.50	0.40	<b>0.40</b>
<walking><in><fife>	0.00	0.25	<b>0.00</b>	0.00	0.00	<b>0.00</b>	0.67	0.17	<b>0.00</b>	0.67	0.17	<b>0.17</b>
<art festivals><in><edinburgh>	0.80	0.50	<b>0.50</b>	0.90	0.40	<b>0.40</b>	0.60	0.70	<b>0.40</b>	0.50	0.50	<b>0.40</b>
<beaches><in><highland>	0.22	0.11	<b>0.00</b>	0.20	0.20	<b>0.20</b>	0.50	1.00	<b>0.50</b>	0.50	1.00	<b>0.50</b>
<museums><in><switzerland>	0.33	0.78	<b>0.33</b>	0.30	0.50	<b>0.22</b>	0.89	0.78	<b>0.78</b>	0.80	0.70	<b>0.70</b>
<museums><in><zurich>	0.30	0.20	<b>0.10</b>	0.40	0.30	<b>0.20</b>	1.00	0.44	<b>0.44</b>	0.90	0.40	<b>0.40</b>
<red kites><near><munlochy>	0.00	0.00	<b>0.00</b>	0.00	0.00	<b>0.00</b>	0.57	0.43	<b>0.43</b>	0.43	0.43	<b>0.29</b>
<canals><near><glasgow>	0.10	0.10	<b>0.10</b>	0.10	0.10	<b>0.00</b>	0.70	0.40	<b>0.20</b>	0.50	0.30	<b>0.30</b>
<walking><near><beaulieu>	0.30	0.50	<b>0.10</b>	0.10	0.20	<b>0.00</b>	0.70	0.60	<b>0.50</b>	0.50	0.40	<b>0.20</b>
<climbing><near><aviemore>	0.50	0.50	<b>0.50</b>	0.40	0.40	<b>0.30</b>	0.50	0.70	<b>0.40</b>	0.50	0.40	<b>0.30</b>
<skiing><near><glencoe>	0.40	0.40	<b>0.40</b>	0.30	0.40	<b>0.20</b>	0.70	0.70	<b>0.40</b>	0.70	0.70	<b>0.40</b>
<beaches><near><portree>	0.10	0.10	<b>0.10</b>	0.00	0.00	<b>0.00</b>	0.57	0.43	<b>0.43</b>	0.43	0.43	<b>0.43</b>
<skiing><near><bern>	0.00	0.10	<b>0.00</b>	0.00	0.00	<b>0.00</b>	0.50	0.60	<b>0.30</b>	0.50	0.30	<b>0.30</b>
<mountains><near><zurich>	0.20	0.10	<b>0.00</b>	0.10	0.10	<b>0.10</b>	0.70	0.50	<b>0.30</b>	0.40	0.30	<b>0.30</b>
<walking><near><luzern>	0.00	0.00	<b>0.00</b>	0.00	0.00	<b>0.00</b>	0.70	0.60	<b>0.40</b>	0.60	0.50	<b>0.50</b>
<camping><near><zurich>	0.20	0.20	<b>0.20</b>	0.00	0.00	<b>0.00</b>	0.33	0.22	<b>0.00</b>	0.20	0.10	<b>0.00</b>
<castles><east><edinburgh>	0.20	0.20	<b>0.00</b>	0.20	0.10	<b>0.20</b>	0.70	0.50	<b>0.50</b>	0.60	0.40	<b>0.40</b>
<camping><west><fort william>	0.10	0.10	<b>0.10</b>	0.10	0.00	<b>0.00</b>	ND	ND	<b>ND</b>	ND	ND	<b>ND</b>
<islands><north><inverness>	0.10	0.00	<b>0.00</b>	0.30	0.30	<b>0.30</b>	0.57	0.14	<b>0.14</b>	0.29	0.29	<b>0.29</b>
<camping><north><inverness>	0.11	0.22	<b>0.00</b>	0.10	0.00	<b>0.00</b>	0.00	0.00	<b>0.00</b>	0.00	0.00	<b>0.00</b>
<cottages><north><ullapool>	0.40	0.50	<b>0.20</b>	0.70	0.20	<b>0.10</b>	0.67	0.67	<b>0.67</b>	ND	ND	<b>ND</b>
<walking><north><dunfermline>	0.11	0.22	<b>0.11</b>	0.20	0.20	<b>0.20</b>	0.67	0.67	<b>0.67</b>	ND	ND	<b>ND</b>
<hotels><north><ullapool>	0.50	0.20	<b>0.00</b>	0.60	0.40	<b>0.40</b>	1.00	0.33	<b>0.33</b>	1.00	0.33	<b>0.33</b>
<beaches><east><dingwall>	0.20	0.20	<b>0.20</b>	0.00	0.10	<b>0.00</b>	0.56	0.44	<b>0.44</b>	0.44	0.33	<b>0.33</b>
<climbing><south><aviemore>	0.50	0.10	<b>0.10</b>	0.30	0.10	<b>0.00</b>	0.67	0.22	<b>0.22</b>	0.67	0.33	<b>0.33</b>
<skiing><east><zurich>	0.50	0.40	<b>0.30</b>	0.70	0.50	<b>0.50</b>	0.30	0.10	<b>0.10</b>	0.20	0.10	<b>0.10</b>
<cottages><within 15km of><portree>	ND	ND	<b>ND</b>	ND	ND	<b>ND</b>	0.80	0.30	<b>0.20</b>	0.80	0.30	<b>0.30</b>
<skiing><within 20km of><fort william>	0.29	0.00	<b>0.00</b>	0.29	0.00	<b>0.00</b>	0.70	0.90	<b>0.60</b>	0.60	0.60	<b>0.50</b>
<skiing><within 100km of><glasgow>	0.00	0.00	<b>0.00</b>	0.00	0.00	<b>0.00</b>	0.70	0.00	<b>0.00</b>	0.50	0.00	<b>0.00</b>
<museums><within 50km of><musselburgh>	ND	ND	<b>ND</b>	ND	ND	<b>ND</b>	0.90	0.20	<b>0.20</b>	0.90	0.10	<b>0.10</b>
<hotels><within 20km of><stirling>	0.00	0.00	<b>0.00</b>	0.00	0.00	<b>0.00</b>	0.80	0.30	<b>0.30</b>	0.80	0.30	<b>0.30</b>
<sailing><within 40km of><grangemouth>	ND	ND	<b>ND</b>	ND	ND	<b>ND</b>	0.40	0.30	<b>0.20</b>	0.40	0.30	<b>0.30</b>
<music><within 50km of><horgen>	ND	ND	<b>ND</b>	ND	ND	<b>ND</b>	1.00	0.56	<b>0.56</b>	0.90	0.50	<b>0.50</b>
<walking><within 50km of><zurich>	0.00	0.00	<b>0.00</b>	0.00	0.00	<b>0.00</b>	0.56	0.78	<b>0.44</b>	0.44	0.89	<b>0.44</b>

Table 2: Thematic, spatial and combined normalised precision for documents retrieved for 38 queries from the SPIRIT collection using textual search (BM25 ranking) and spatial search with T-index and non-distributed ranking (ND means no documents returned/judged)

	Annotator 1, BM25			Annotator 2, BM25			Annotator 1, Spatial			Annotator 2, Spatial		
Spatial relationship	Th (P <sub>n</sub> )	Sp (P <sub>n</sub> )	Com. (P <sub>n</sub> )	Th (P <sub>n</sub> )	Sp (P <sub>n</sub> )	Com. (P <sub>n</sub> )	Th (P <sub>n</sub> )	Sp (P <sub>n</sub> )	Com. (P <sub>n</sub> )	Th (P <sub>n</sub> )	Sp (P <sub>n</sub> )	Com. (P <sub>n</sub> )
Inside	0.50	0.42	<b>0.27</b>	0.40	0.32	<b>0.22</b>	0.71	0.54	<b>0.39</b>	0.63	0.45	<b>0.39</b>
Near	0.18	0.20	<b>0.14</b>	0.10	0.12	<b>0.06</b>	0.60	0.52	<b>0.34</b>	0.48	0.39	<b>0.30</b>
Directional	0.27	0.21	<b>0.10</b>	0.32	0.19	<b>0.17</b>	0.57	0.34	<b>0.34</b>	0.46	0.26	<b>0.26</b>
Within distance of	0.07	0.00	<b>0.00</b>	0.07	0.00	<b>0.00</b>	0.73	0.42	<b>0.31</b>	0.67	0.37	<b>0.31</b>

Table 3: Mean thematic, spatial and combined normalised precision for spatial relationship types using textual search (BM25 ranking) and spatial search with T-index and non-distributed ranking