geoXwalk – A Gazetteer Server and Service for UK Academia

J.S.Reid

GeoServices Delivery Team,EDINA, Edinburgh University Data Library,

George square, Edinburgh, EH8 9LJ, Scotland.

Telephone: +44 (0) 131 651 1383
FAX: +44 (0) 131 650 3308
Email:james.reid@ed.ac.uk

Abstract

This paper will summarise work undertaken on behalf of the UK academic community to evaluate and develop a gazetteer server and service which will underpin geographic searching within the UK distributed academic information network. It will outline the context and problem domain, report on issues investigated and the findings to date. Lastly, it poses some unresolved questions requiring further research and speculates on possible future directions.

1. Introduction

The Joint Information Systems Committee (JISC) is a strategic advisory committee working on behalf of the funding bodies for further and higher education (FE and HE) in the United Kingdom. The JISC promotes the innovative application and use of information systems and information technology in FE and HE across the UK by providing vision and leadership and funding the network infrastructure, Information and Communications Technology (ICT) and information services, development projects and high quality materials for education in what is referred to as the JISC 'Information Environment' (JISC IE)- Figure 1. The JISC IE provides access to heterogeneous resources for academia, ranging from bibliographic, multimedia and geospatial data and associated materials.

The geoXwalk project was conceived as a development project to build a shared service which would service the JISC IE by providing a mechanism for geographic searching of information resources. This would complement the traditional key term and author type searches that have been supported. 'Geo-enabling' other JISC services would provide a powerful mechanism to support resource discovery and utilisation within the distributed JISC IE.

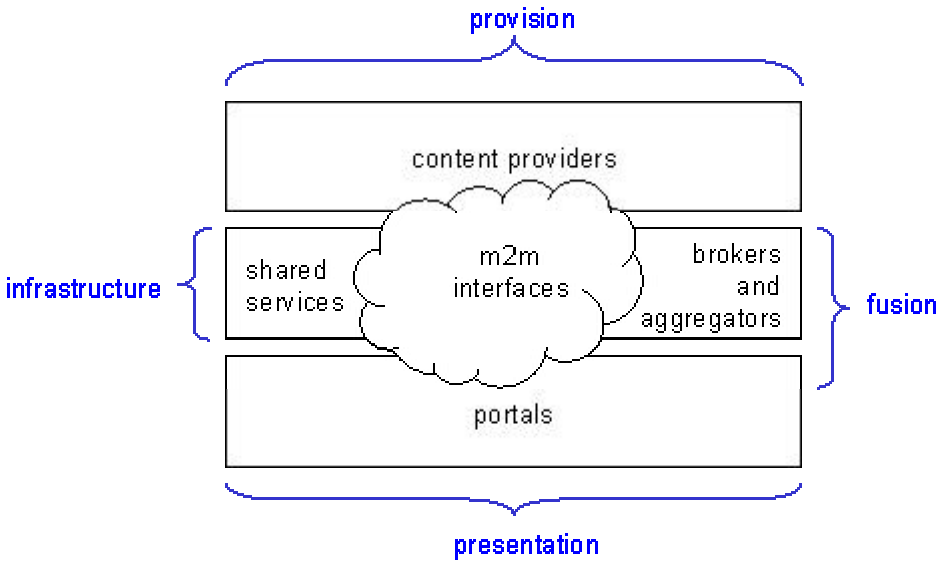

Figure 1. The general architecture of the JISC Information Environment.

Source: Information Environment: Development Strategy 2001-2005.

2.Problem Domain

Geographic searching is a powerful information retrieval tool. Most information resources pertain to specific geographic areas and are either explicitly or implicitly geo-referenced. The UK's National Geospatial Data Framework (NGDF) estimates that as much as eighty per cent of the information collected in the UK today is geo-referenced in some form. Geography is frequently used as a search parameter, and there is an increasing demand from users, data services, archives, libraries, and museums for more powerful geographic searching. However, there are serious obstacles to meeting this demand.

Geographic searching is often restricted because geographic metadata creation is excessively resource intensive. Accordingly, many information resources have no geographic metadata, and where it exists, it usually only extends to geographic names. Search strategies based on geographic name alone are very limited (although they are a critical and often the only access point). An alternative is to geo-reference the resources using a spatial referencing system such as latitude and longitude or, in the UK, the Ordnance Survey National Grids (of Great Britain and Northern Ireland).

The existence of a maze of current and historical geographies has created a situation where there is considerable variation in the spatial units and spatial coding schemes used in geographic metadata. Many geographic names have a number of variant forms, the boundaries in different geographies do not align, and names, units and hierarchies have changed in the past, and will continue to change. In 1990, the UK Data Archive (University of Essex) identified ninety different types of spatial units and spatial coding schemes in use in their collection and more have since been added. By way of illustration, the Resource Discovery Network (RDN) ResourceFinder (http://www.rdn.ac.uk/) does not currently have a specific geographic search function, instead it simply searches using a text based mechanism, thus to find information referring to a particular place the place must be referred to by name in the description field of the resource's accompanying metadata. Using geoXwalk, the placenames in the metadata can be turned into other geographical identifiers enabling searching by e.g. minimum bounding rectangle, postcode, county.

Clearly, no single system of spatial units and coding will suit all purposes as people conceptualise geographic space in different ways and different servers deploy differing geographic naming schemes. Ideally, users should not be forced to have to explicitly convert from one 'world' view to another. However, it is impractical for most service providers to support more than a few geo-referencing schemes, or develop the means to convert from one to another. A comprehensive gazetteer linking a controlled vocabulary of current and historical geographic names to a standard spatial coding scheme (such as latitude and longitude and/or the Ordnance Survey National Grid(s)) would be necessary to perform these translations (or 'cross-walks' – hence the name) and would therefore provide the capacity to be, what might be referred to as 'geographical agnostic' – see Figure 2. This is what the geoXwalk project aims to deliver.

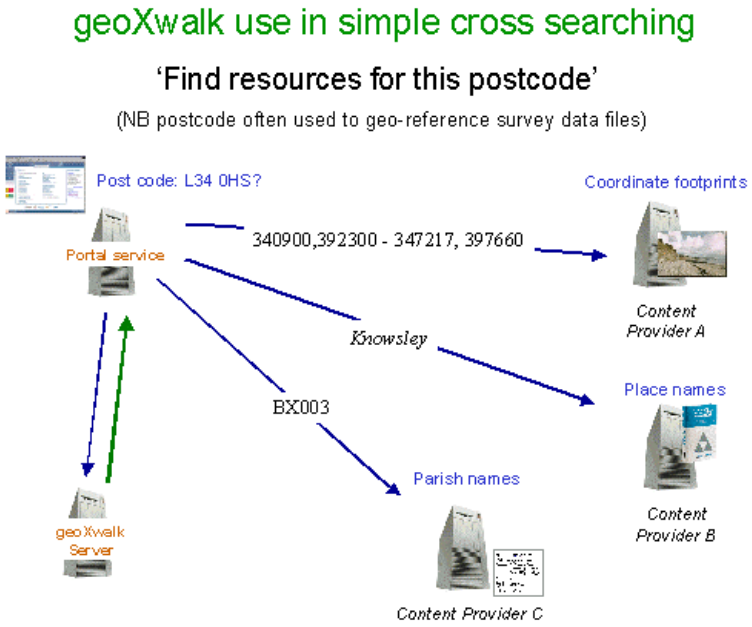

Figure2. Example case of geoXwalk to 'cross-walk' different geographies.

3.Background and Rationale behind geoXwalk

geoXwalk was funded by the JISC as a two Phase development project, the principle aim of which was to assess the feasibility of developing and providing an online, fast, scaleable and extensible British and Irish gazetteer service, which would play a crucial role in supporting geographic searching in the JISC IE. The project was a joint one between EDINA, Data Library, University of Edinburgh, and the History Data Service, Data Archive, University of Essex.

The general aim of the project is to investigate the practicability of a gazetteer service, (a network-addressable middle-ware service), implementing open protocols, specifically the Alexandria Digital Library Gazetteer Protocol (ADL 1999, Janée and Hill 2001) and the Open GIS Consortium's (OGC) Filter Specification (OGC 2001), to support other information services within the JISC IE. Firstly, by supporting geographic searching and secondly, by assisting in the geographic indexing of information resources. As the project has progressed it has become clear that there is also a requirement for it to act as a general reference source about places and features in the UK and Ireland. Phase I was conducted as a scoping study to determine the feasibility and the requirements for such a service while

Phase II which commenced in June 2002 aims to develop an actual working demonstrator geo-spatial gazetteer service suitable for extension to full service within the JISC IE.

4.Technical Implementation

For conceptual purposes the geoXwalk project can be decomposed into a number of interdependent though to some degree independent features:

- a gazetteer database supporting spatial searches;
- middleware components comprising APIs supporting open protocols to issue spatial and/or aspatial search queries;
- a semi-automatic document 'scanner' that can parse non-geographically indexed documents for placenames, relate them to the gazetteer and return appropriate geo-references (coordinates) for confirmed matches – a 'geoparser'.

4.1 The Gazetteer Database

The gazetteer database itself is of course crucial but what is of particular relevance is that geoXwalk extends the concept of a traditional gazetteer. Such gazetteers typically only hold details of a placename/feature along with a single x/y coordinate to represent geographic location . In order to provide answers to the sorts of sample queries that geoXwalk will be expected to resolve (e.g. What parishes fall within the Lake District National Park?; What is at grid ref. NT 258 728?; What parishes fall within the Lake District National Park?) each geographical feature held in the database needs to provide, as a minimum: a) a name for the feature, b) a feature type and c) a geometry.

In the case of the latter, the simplest geometry would be a point location but significantly, geoXwalk accommodates recording of the actual spatial footprint of the feature i.e. polygons as opposed to just simple points. By using a spatially enabled RDBMs, geoXwalk can then determine at runtime the implicit spatial relationships between features (this contrasts with the classic 'thesaurus' approach in which predetermined hierarchies of features are employed to capture the spatial relationships between features). Holding such extended geometries in the database therefore permits more flexible and richer querying than could be achieved by only holding reduced geometries. Furthermore, all polygonal features can be optionally queried as point features if desired.

A feature type thesaurus has been implemented based on a national dialect of the ADL Feature Type Thesaurus (ADL FTT) (http://www.alexandria.ucsb.edu/~lhill/FeatureTypes/) and provides a means of controlling the vocabulary used to describe features when undertaking searches. Phase I of the project had identified the ADL FTT as a suitable candidate given its relative parsimony balanced against richness and adaptability.

The database schema used in geoXwalk is itself based upon the ADL Content Standard (http://www.alexandria.ucsb.edu/gazetteer/gaz_content_standard.html). Examination of the ADL standard showed that it could embody the richness of an expanded gazetteer model. It further allows temporal references to be incorporated into entries and defines how complicated geo-footprints can be handled. Explicit relationships can be defined, which is of particular use when a gazetteer holds significant amounts of historical data for which geometries do not exist (although implicit spatial relationships between features based on

their footprints is arguably the more flexible approach to favour).

The standard also recognises that data sources will not hold all the detail permitted by the standard. The minimal requirement for an entry and the holding of attribution information with each (and in some cases on fields within) sections means that features can be easily introduced into a gazetteer and be updated with additional information from other sources at a later date. Potentially this could be save a significant amount of time when the gazetteer is constructed from a variety of data sources (as in fact geoXwalk is). Also, from a pragmatic point of view, the proposed standard has been implemented in a web based gazetteer service, containing over 10 million entries.

The current implementation of the gazetteer resides in an Ingres database with customised external support routines providing the spatial functionality. Benchmarking of this solution against an implementation in Oracle 9I is due to commence shortly as response times and scalability issues are important design goals and it is essential that the geoXwalk server does not become a bottleneck for other servers using it to answer their geographical queries.

4.2 The Middleware Components

Two protocols are currently supported by geoXwalk – the ADL Gazetteer Protocol (Janée and Hill 2001) and the OGC Filter Encoding Implementation Specification (OGC 2001). In order to process incoming queries a suite of servlets has been written that transforms the XML based queries generated by the client and maps these to database specific SQL statements. Additionally, in a sample demonstrator that has been developed, an OGC Web Map Server (WMS) is deployed to produce map based visualisations of the returned geographical features. The demonstrator is purely for illustrative purposes as the primary focus of the geoXwalk server is to service machine to machine (m2m) queries. A related project, Go-Geo! (www.gogeo.ac.uk), which is a geospatial metadata discovery service for UK academia, utilises the geoXwalk server to perform spatial searching as illustrated in Figure. 2. SOAP implementations have also been developed.

4.3 The geoparser

Much of the data and metadata that exists within the JISC IE, while having some sort of georeference (such as placename, address, postcode, county etc.) is not in a format that allows it to be easily spatially searched. One task of the project was to investigate how existing, non-spatially referenced documents could be spatially indexed. Using the gazetteer as reference, a prototype rule based geoparser has been implemented that can semi-automatically identify placenames within a document and extract a suitable spatial footprint (see Figure 3).

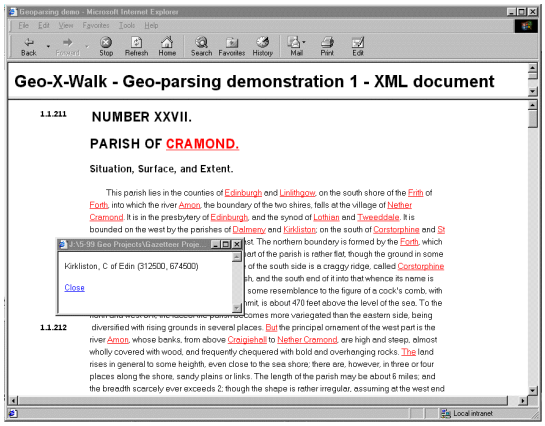

Figure 3 A sample geoparsed document – underlined words have been automatically identified by the geoparser as potential placenames and where possible a georeference (OSGB National Grid coordinates) derived from the geoXwalk gazetteer.

The approach taken has not relied on a 'brute force' approach as this is too resource intensive and in tests proved too unreliable. The rule based approach takes account of the structure of the document and the context within which the word occurs. As a refinement to the basic geoparser, a web-based interface has also been developed to allow interactive editing/confirmation of the results of the geoparser (Figure 4). Resultant output from this stage can then be used to update the document's metadata to include the georeferences thus making the original document spatially searchable.

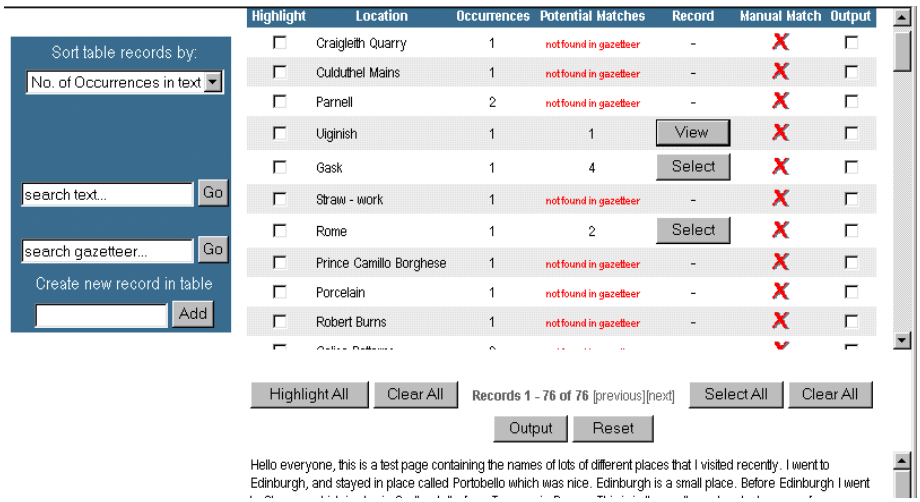

Figure 4 An editing interface to the geoparser to assist in the semi-automatic georeferencing of documents.

5. Outstanding Issues

The development of both the gazetteer and the geoparser have identified a series of questions that require further investigative research. Amongst the more intractable of these are issues associated with map conflation and improvements to the geoparser.

The first of these, referred to elsewhere as 'map conflation' (Yuan 1999) essentially is concerned with the identification and resolution of duplicate entries in the gazetteer. In essence the problem is one of being able to discriminate between existing and new features and to be able to determine when two (or more) potentially 'similar' geographic features are sufficiently 'similar' to be considered as the same feature. As geoXwalk dimensions a geographic feature on the basis of:

- its name;
- its feature type (ADL FTT entry) and
- its geometry (spatial footprint).

there are three distinct 'axes' along which the candidate features must be compared (see Figure 5).

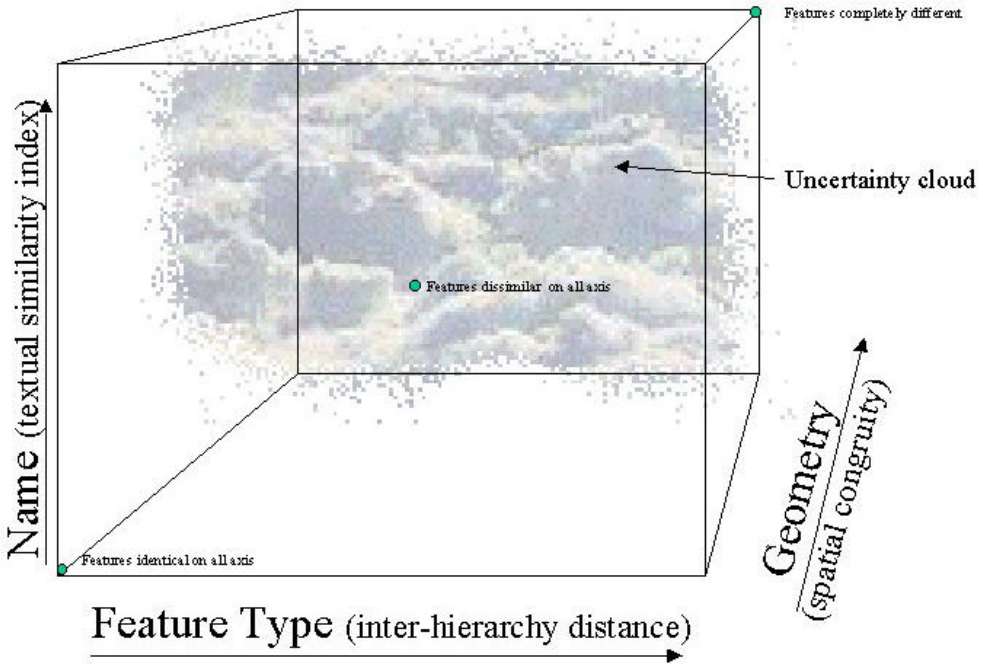

Figure 5 The three 'axes' used to discriminate between new and existing features.

Furthermore, the question arises of how closely on all three axis must the features correspond and whether one axis should be weighted more highly than another. As geoXwalk relies on answering queries by using the implicit spatial relations between features using their geometries, one rule-of-thumb would be that a features geometry is it's principal axis and takes precedence over name and type. However, as more historical data is added to the gazetteer this problem intensifies – take for example the case of London (contemporary) and Londinium (Roman) – while both may be regarded as the same place but with radically different spatial footprints and possibly feature types (town vs. city). The derivation of some metric that would allow confidence limits to be attached to features as they are added to the gazetteer is therefore a pressing research priority.

A lot of refinement work could be conducted on the current basic implementation of the geoparser, specifically optimising it in terms of performance and accuracy in order to facilitate realtime geoparsing as well as minimising the number of false positives identified. Additionally, the capacity to derive a feature type automatically as well as the footprint itself would be extremely useful.

6. The Future

The project ends in June 2003 at which point evaluation by JISC may lead to an extension to a fully resourced shared service within the JISC IE. However, significant interest in the project exists outside the academic community that could forseeably lead to a more 'public' service that is of relevance to a wider audience that just the UK academic sector. Indeed, the model employed is sufficiently adaptable to scale to the development of a European geoXwalk comprising a series of regional geoXwalk servers. Such a facility would form a critical component of any European Spatial Data Infrastructure.

7.Acknowledgements

The geoXwalk project has been funded by the Joint Information Systems Committee (JISC) as part of its 'shared services' strategy.

8. References

Joint Information Systems Committee 2001, *Information Environment: Development Strategy 2001-2005.*

Alexandria Digital Library Project,1999 [online] <http://www.alexandria.ucsb.edu/> , and ADL. *Alexandria Digital Library Gazetteer Development Information*. [online] <http://www.alexandria.ucsb.edu/gazetteer/> 01 October 1999.

Janée G. and Hill, L., 2001, The ADL Gazetteer Protocol, Version 1. Available online: <http://alexandria.sdc.ucsb.edu/~gjanee/gazetteer/specification.html>

Open GIS Consortium Inc., Filter Encoding Implementation Specification Version 1.0.0, 19 September 2001. [online]: <http://www.opengis.org/techno/specs/02-059.pdf>

Yuan S., 1999, Development of Conflation Components, *Geoinformatics and Socioinformatics* '99 Conference Ann Arbor,19-21 June 1999,pp. 1-13.

geoXwalk Phase I and II documentation available at: <www.geoXwalk.ac.uk>