

Go-Geo! - Geo-data Portal

Geo-Data Portal – Transition to Service

Final report to the JISC IE Programme

JISC Portals Programme

Authors	David Medyckyj-Scott (Project Manager), Hilary Beedham (UKDA Projects Manager), Julie Missen (Project Assistant), Philip Abrahamson (Software Engineer), Eddie Boyle (Software Engineer), Tony Mathys (Information Officer)
Date	8 September 2004
Version	1a
Status	Final
Contact:	Dr David Medyckyj-Scott, EDINA National Data Centre University of Edinburgh, Main Library Building, George Square, Edinburgh EH8 9LJ Tel: 0131 651 1308 Fax: 0131 650 3308 d.medyckyj-scott@ed.ac.uk

Table of Contents

<i>Table of Contents</i>	1
<i>Acknowledgements</i>	2
<i>Executive Summary</i>	3
<i>Background</i>	5
The Metadata Issue	6
Relationship with the AGI Glgateway	6
<i>Aims and Objectives</i>	7
<i>Methodology</i>	7
Metadata standards	8
The portal application (as of the end of Phase 2)	8
Roll out of trial service	11
User and stakeholder engagement and evaluation	11
A study into the requirements for data distribution	11
<i>Implementation</i>	12
Trial Service Roll Out and Operation	14
Evaluation of the Go-Geo! Trial Service	14
The Data Distribution Study	15
<i>Outputs and Results</i>	16
The portal today	16
Use of the portal	21
User feedback	22
Status of the development work	24
From discovery to accessing data	30
<i>Issues arising</i>	34
Cross-walking between metadata standards	35
Which geospatial metadata standard?	35
Widening the dissemination of the metadata	36
<i>Implications and future plans</i>	37
The next 12 months	37
Further development opportunities	37
Exit strategy	41
<i>Recommendations</i>	42
<i>Conclusions</i>	43
<i>References</i>	44
<i>Appendix 1 The People Search tool</i>	45
<i>Appendix 2 Final Budget</i>	47

Acknowledgements

This project was part of the JISC Portals programme and was funded by JISC. The project wishes to acknowledge the support of the staff of the EDINA National Data Centre and UK Data Archive, the other JISC and Research Council funded service providers and the members of the JISC Geospatial Working Group.

Executive Summary

Go-Geo! (www.gogeo.ac.uk) is an online resource discovery tool which allows for the identification and retrieval of records describing the content, quality, condition and other characteristics of geospatial data that exist with UK tertiary education and beyond. (Geospatial data are any data with geographic references, for example, Ordnance Survey national grid references, postal addresses, place names or the names of administrative areas.) The portal supports geospatial searching by interactive map, grid co-ordinates and place name, as well as the more traditional topic or keyword forms of searching.

Go-Geo! has been a cooperative effort between EDINA National Data Centre, University of Edinburgh, and the UK Data Archive, University of Essex.

Using an ANSI standard, Z39.50-1995, the portal undertakes simultaneous searching across many resources including the national Ggateway service and its network of catalogue services. A key feature is the ability for users to find other related resources, such as books, photographs, projects, maps, for their geographic area of interest. These resources are discovered by cross searching the JISC Information Environment and other online information services. The focus of the portal is therefore on 'where' a resource about and less on the 'what' it is about, which is the focus of other JISC portals.

Another important aspect of the portal is the resources section, which provides comprehensive information about geographic information resources. This includes links to information about training courses, learning materials, organisations, books and journals and software which may be of interest to the user.

To date there have been 3 phases in the development of Go-Geo!, all funded by the JISC. This phase, phase 3, had the twin aims of developing the portal to a point where it is suitable for roll out as a full service and trialling it with the user community. The trial service was launched in November 2003 and there has been steady use since then. The results from an ongoing programme of evaluation found that users liked the look and feel of the portal, especially the resource listings, they felt that the search facility was easy to use, they would use the portal again and that it could be an extremely useful facility.

As well as the rolling out of a trial service, a number of specific technical developments were undertaken to develop and enhance the trial service. This included the linking in of additional metadata catalogues including those provided by the Archaeology Data Service and the UK Data Archive, the implementation of a people search facility and a metadata creation tool, and the development of a portlet which can be plugged into other portals and permits the Go-Geo! network of catalogues to be cross searchable from those portals.

The evaluation work, coupled with other feedback, clearly demonstrated that academic users have need of a mechanism for discovering resources in the context of their geographical coverage. The goal of the portal has been to fulfil this need by creating a universally applicable system under which resources, and in particular geospatial data, collected and catalogued under other headings, (for example, archaeological and historical data), can be discovered in terms of their geographical coverage. This goal has been met. However, it is also clear that users want access to data as well. While data residing in specialist services are available for academic use from those services, access to data held by individuals is more problematic. Part of the work undertaken in this project was an investigation into cost-effective ways of distributing geospatial data held by individuals, research teams and departments.

The study found that a number of issues needed to be addressed by any solution including user and depositor support, quality control, copyright and licensing, security, preservation, ease of use and cost. Three solutions were proposed: a distributed one where repositories are set up at individual organisations, one based on a central archive and finally a mixture of these two. No firm conclusion was drawn as to which was the best and further work is required.

As well as identifying further development opportunities, a number of recommendations arise from the project of which the keys ones are as follows.

1. That the existing trial service should become a full service within the JISC IE.
2. Two important needs for the management of geospatially referenced data have been identified. First, a system to assist in institutional geospatial data management. Second, an enhanced national service for geospatial resources with an emphasis on access to data. The portal has the potential to serve both these needs and we seek further funding to meet these goals.
3. Metadata is critical to the success of the service and, in turn, to publishing the existence of data to encourage re-use. Additional funding is required to undertake the task of metadata co-ordination and QA.
4. The JISC should formally state that new HE/FE application profile, based on the recently published ISO 19915, is to be used by those documenting geospatial data in UK HE and FE institutions.
5. Major barriers to data sharing need to be lowered if not removed entirely. A further study investigating how improved data sharing can be facilitated is required, possibly within the JISC's new repositories programme.
6. A broader strategy for the management, preservation and storage, and dissemination of geospatial datasets within the UK tertiary education needs to be developed by key stakeholders. This is something that needs to be taken up by the relevant parties with some urgency.

Background

The Geo-data Browser Project Phase 1 (August 2000 - June 2001) was funded by the JISC to investigate the demand for and the issues surrounding the establishment of a Z39.50 compliant resource discovery tool for geospatial data for UK academia. The project was a joint one between EDINA, Data Library, University of Edinburgh, and the History Data Service, UK Data Archive, University of Essex. MIMAS and the Archaeology Data Service were involved in an advisory capacity.

Between May 2002 and June 2003, JISC funded a second project (Phase 2) with the goal of developing what by then had become perceived as a geospatial data portal. The aims of the phase 2 project were: development of a demonstrator suitable for extension to full service; promoting the possibilities of a fully functioning service for integration in the JISC Information Environment (IE), and to act as a proof of concept.

The earlier scoping and demonstrator projects focused on the idea of a resource discovery tool for identifying and retrieving metadata describing geospatial data available within the UK tertiary education community along with other related resources of use to the user. (Geospatial data is a term used to cover any data with a geospatial reference - for example, an Ordnance Survey National Grid reference, postal address, place name or administrative area. At least 80% of all data has some geographic reference that makes it implicitly geospatial data.) The portal is thus an example of what JISC has termed a media-specific portal.

The problem the portal is attempting to help solve is the lack of awareness of what geospatial data already exist within the UK tertiary education sector. How does a student or researcher find out what geospatial data already exist within UK academia; having located them, how do they determine its quality and fitness for purpose; then how do they gain access and, finally, how do they use them? The solution is to provide comprehensive, standardised metadata, available through a web-searchable service.

Increasing amounts of geospatial data are being created within the UK tertiary education and research community. To gain maximum benefit from these data, to ensure that their teaching and research potential is fully exploited, and to minimise the redundancy of data collection, the existence of geospatial data should be recorded and then promoted to others. It is recognised that much of these data, subject to there being no licensing issues, could also have value outside academia within the UK Geographic Information (GI) community more generally. These data range from topographical or thematic digital maps, digital boundary data to statistical, environmental and historical information suitable for the purpose of spatial analysis. The demand for access to geospatial data is growing; one reason being the availability of Ordnance Survey digital map data through the EDINA Digimap service. However, the project team now have ample evidence that there are significant barriers to the sharing of data.

The core function of the Go-Geo! portal is the identification and retrieval of metadata records describing the content, quality, condition and other characteristics of geospatial data. However, the geo-data portal extends the data discovery function by the addition of other functions that add value to the data identified and retrieved. Primarily this means providing users with ways to find other related resources of use to them in teaching and research.

Two sorts of related resources have been identified. First, those that help the full exploitation and use of the geospatial data discovered by the user, such as appropriate application software and information on data formats. Second, resources that are used alongside these data and assist the user in the process of understanding a problem and its investigation e.g. related case studies and projects, articles, lists of journals, reports, personal contacts, mailing-lists.

These resources are found either by searching the JISC IE and other online information services or through quality assured links listed within the Portal. The Portal therefore goes beyond the JISC definition of media-specific portals by the inclusion of related resources not just specific media i.e. geospatial datasets. A key and novel aspect of the search capability available within the portal is that searching can be constrained to only look for resources for specific geographic locations. The portal is therefore also an access point to geographically related resources within the JISC IE. The focus is on

the 'where' is a resource about and less on the 'what' is it about, which is the focus of other JISC portals.

In January 2003, the JISC requested an exit strategy for the project¹. EDINA and the Data Archive proposed that the user feedback from the workshop strongly testified that the demonstrator should become a service within the JISC Information Environment and a date of January 2004 was proposed by which this should occur. The strategy document noted however that a full service required some further development work. The project team therefore requested additional funding in order to roll out the service and undertake further development work. In March 2003, the JISC agreed to make available funding for a new project (phase 3) to undertake some of the work proposed. However, this excluded collection of metadata and the creation of tools to assist in geospatial data metadata collection. The JISC was also unable to support the proposal for the launch of the service in January 2004. Instead it has proposed that the JISC Geospatial Working Group review the case for long term funding² and that a case then be made to the JISC JCCS in May 2004 to launch Go-Geo! as a service.

The Metadata Issue

The exit strategy produced in January 2003 pointed out that the issue of metadata provision was critical to the success of the service. UK academia has a poor record when it comes to data management. There are a growing number of undocumented datasets and not just in the area of geospatial data. The amount of metadata describing geospatial data existing within UK academic institutions is very low. There was not (and still is not) a ready source of metadata with which to populate the portal. There is a need to encourage geospatial data creators/providers to provide quality metadata. The exit strategy therefore requested money to fund a member of staff who would have responsibility for co-ordinating metadata collection and quality assurance of metadata records. However, creation of content, in this case metadata, is not an activity normally funded by JCIE.

The issue of metadata creation while a challenging one cannot be ignored. Users will come to the service with an expectation of discovering the existence of geospatial datasets that meet their needs. The project team recognised that a critical mass of metadata was required in order to encourage use of the service. To obtain this critical mass it made it a priority early on in phase 1 to link up with the National Geographic Data Framework (NGDF), now known as Glgateway. However, without the first step of metadata creation by the academic community, the vision for the Go-Geo! portal will be diminished.

In May 2003 JISC informed EDINA and the UK Data Archive that it was willing to fund a sister project to this one specifically to deal with the metadata issue. The aim of the Geospatial Metadata Initiative Project was to introduce and promote metadata and metadata creation amongst members of the geospatial academic community through metadata workshops and meetings. It was also hoped that an output would be content for the Go-Geo! portal i.e. metadata records. While the former has been very successful, it has not yet led to the provision of records to the Portal. The reasons for this outcome have been summarised in the final report submitted to the JISC on the Geospatial Metadata Initiative Project. It should be read alongside this report as many of its findings relate directly to the Go-Geo! portal. Some of the key findings are included below.

Relationship with the AGI Glgateway

The project has an ongoing and close relationship with the AGI Glgateway team (formerly NGDF), offering advice in a number of areas. Glgateway is a government-funded initiative. The Go-Geo! project team is recognised by the AGI and others as being a point of expertise not found elsewhere in the UK. (Indeed as a result of the project team's expertise, EDINA was successful in winning the contract to host the Glgateway service on behalf of the AGI and EDINA is now responsible for on-going technical development of the service. The UKDA is supporting the AGI in the transition of the Glgateway metadata guidelines³ to a compliant ISO 19115 standards version.)

¹ This was a challenge as the project still had 6 months to run and the main user engagement activity had not yet commenced.

² This actually happened at a meeting in London in June 2004.

³ Formerly the NGDF Discovery Metadata Guidelines

The AGI are very keen for the establishment of an academic ‘node’ through which the wider GI community in the UK can become aware of what data has been created and/or are curated by researchers in UK academia. Integrating an academic geospatial metadata catalogue into Glgateway framework means that geospatial data produced within UK academia could be visible more widely to, for example, civil servants, local government officers, environmental researchers, etc. It may be possible for such individuals to get access to these datasets although licence conditions may restrict some datasets to academic use.

Having allowed Go-Geo! to cross-search the Glgateway catalogues for over two years, thus providing a critical mass of geospatial metadata for Go-Geo!, AGI have expressed a desire for a reciprocal arrangement that would allow Glgateway to cross-search the Go-Geo! catalogues. They would like this to take place in Autumn 2004 but it would require the consent of Go-Geo! metadata contributors.

Aims and Objectives

The **aims** of the Phase 3 project were:

- To develop the geospatial portal to a point where it is suitable for roll out as a full service
- To trial the service such that issues relating to performance and usability might be measured
- Promote the possibilities of a fully functioning service to potential users
- Promote within the HE and FE communities the benefits of the ability to process material at a geographical level

The key **objectives** were:

- to continue to develop and enhance the service in preparation for service launch
 - 1) link in two existing Z-targets provided by two geospatial data service providers within the JISC IE to form the basis of the Geo-data Network⁴. The targets selected were the Archaeology Data Service catalogue and, time-scales permitting, the UK Data Archive catalogue.
 - 2) addition of user profiling.
 - 3) improvements in the spatial searching of directory services, remote databases and resource catalogues
 - 4) development of a facility (the ‘Geo-Locator’) similar to RDN-I that would allow the Geo-data Network to be searchable via the Gateway from other Portals.
 - 5) identification of new developments and enhancements to the portal.
- trial service roll out to institutions supported by an on-going programme of evaluation of 6 institutions;
- undertake a study to investigate a cost-effective way of distributing geospatial data held by individuals, research teams and departments;
- continue to engage with users to assess their needs and to promote the possibilities of a service through demonstrations at conferences and general awareness raising etc, and
- continue to develop relationships with relevant initiatives in geospatial searching, specifically the AGI Glgateway and the RDN Geography and Environment Hub.

The key **deliverable** of the project was to be a functioning, scalable portal service ready for launch within the JISC Information Environment.

A number of changes to the objectives have occurred over the duration of the project. Some of these resulted from external factors that arose during the project period; others arose as a consequence of internal project issues. The changes and the reasons for them are documented below.

Methodology

Note: This section should be read in conjunction with similar sections in previous Go-Geo! project final reports. The methodology adopted was an extension of that employed previously, since it had proved to be successful.

⁴ The portal is already able to cross-search the HDS data catalogue, an existing, remote, structured directory service which contains geospatial data.

Metadata standards

The importance of creating standardised metadata to describe geospatial datasets has been recognised for many years (e.g. Medyckyj-Scott et al, 1989). Metadata standards help to organise geospatial data so that they are accessible and understandable to people other than those who produced them. Typically, many people other than the producer could use the geospatial data. Proper metadata provides those unfamiliar with the data with a better understanding, and enable them to use it properly. Over the years many geospatial data specific metadata standards have been proposed and implemented. A review of these was undertaken as part of the phase 1 project (O'Hanlon, 2001).

In the UK, the NGDF produced a standard for describing geospatial resources, the NGDF Discovery Metadata Guidelines. This was a subset of a US standard, the FGDC Content Standard for Digital Geospatial Metadata. An analysis of NGDF Discovery Metadata Guidelines standard in Phase 1 found that it was a suitable standard to be used by UK academia to describe geospatial datasets. However, some elements required modifications and new pieces of information need to be added so that the standard met the needs of HE and FE. These included audience, why these data were collected, archival responsibility and data quality. An application profile of the NGDF Metadata Guidelines was created which was entitled 'FE/HE Application Profile of the NGDF Metadata Guidelines' to distinguish it from the NGDF Guidelines proper (Medyckyj-Scott et al, 2001).

One of the major outputs of phase 2 was the production of a final version of FE/HE Application Profile of the now renamed Glgateway Discovery Metadata Guidelines standard along with guidelines for the creation of metadata. This standard is used by the Go-Geo! portal. In anticipation of having to migrate from the Glgateway Discovery Metadata Guidelines to a new ISO Standard (19115 Metadata standard for geographic data) should Go-Geo! become a full service, a cross-comparison was made between the HE/FE Profile and ISO elements to see how well they matched. As part of this comparative evaluation, the HE/FE profile was also compared to the FGDC Content Standard. The results showed that the elements of both standards could be mapped with few problems. As will be seen below, the issue of metadata standards and mappings has become increasingly relevant as the project has gone on.

The portal application (as of the end of Phase 2)

Go-Geo! is built upon the Z39.50 protocol which supports searching of remote databases. The Z39.50 Protocol is used for searching and record retrieval using the GEO Profile and Bath Profiles, which define the Attribute Set and acceptable Record Syntaxes. Metadata can be output in the following formats: FE/HE Application Profile of the Glgateway Metadata Guidelines, NGDF Discovery Metadata (to provide to the NGDF Gateway) and Dublin Core although within the service records are presented as FE/HE Application Profile of the Glgateway Metadata Guidelines. Go-Geo!'s Z-39.50 client uses a PHP implementation of the YAZ toolkit.

During phase 2, a demonstrator was built which was able to cross-search the following:

- a database local to the portal containing sample metadata records provided by ADS, EDINA, and MIMAS;
- the AHDS Data Service catalogue;
- a variety of existing resource catalogues containing geo-related resources including COPAC (maps, government publications, books and articles), and
- the ERSC Regard project database.

Due to problems with the Z-server hosted by GE:source and issues regarding the way they geographically referenced their records, it was not possible to show cross-search the GE:source resource catalogue. This remains the case.

Interoperability with the locally developed "geoXwalk"⁵ web service allows cross-searching of geographic resources held in databases that have geographic references, e.g. place names, but not explicit co-ordinates e.g. COPAC and REGARD. To enhance performance and permit sorting and relevance ordering of records, a 'metadata search engine' was built. This comprises a set of indexes,

⁵ geoXwalk is an online, fast, scalable and extensible British gazetteer middleware server with the goal of supporting geographic searching in the JISC IE. It is being developed jointly by EDINA and the UK Data Archive. Details about the use of geoXwalk in Go-Geo! can be found in the final report of the Go-Geo! phase 2 project.

created by harvesting records from catalogues holding geospatial metadata operating as Z-targets, and application code which supports searching using the Z39.50 protocol. This accepts query inputs from the Go-Geo! portal client, undertakes a search of the indexes which describe the metadata held at remote Z-targets and returns a list of results (titles) and the Z-target on which the actual record can be found. The Z-targets can be individually queried to retrieve the metadata records themselves.

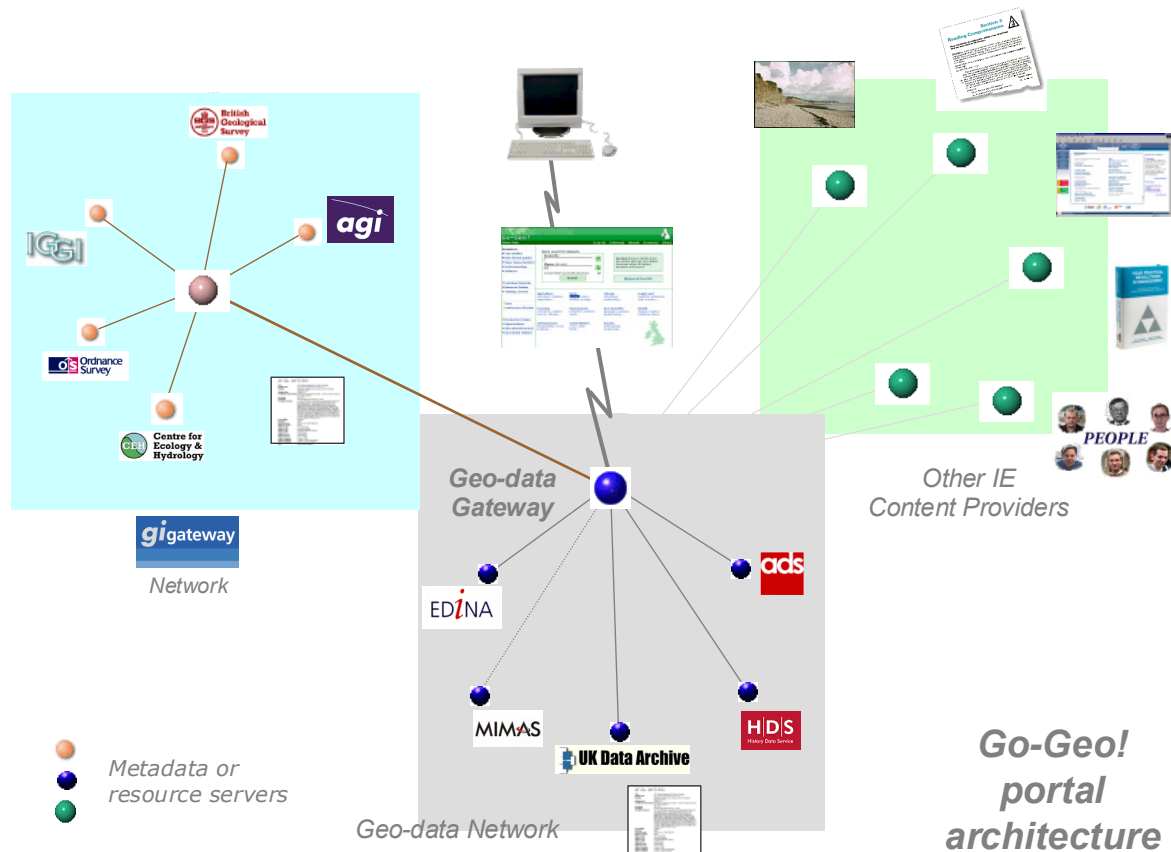


Figure 1 The Go-Geo! portal architecture

By the end of Phase 2, Go-Geo! supported both simple and advanced searches. The simple search offers a 'google-like' free-text subject and place-name search and provides a list of relevance ordered results that match the geographic area of the inputted place-name. Each resulting record is presented in a succinct tabbed format along with context maps showing the geographic extent of the data described by each record. E-mail and print functions are available by which single and user selected record sets can be downloaded.

Advanced search functionality extends the simple search by allowing users to search for geospatial data and resources by subject/topic, date, resource type and geographic location. A tabbed format is used to assist users in constructing a query. The geographic area of interest can be defined by place name, administrative area, grid co-ordinates or by using an interactive map.

The demonstrator also provided links to information on major geographic data providers, GIS organisations, a glossary of terms, news, events and learning resources. Supporting material was provided to help users in their use of the service.

Developments proposed in phase 3

A number of enhancements to the portal were planned both in preparation for trial service launch and during the trial period.

1) One of the project aims was to link in other existing catalogues which hold geospatial metadata and which were provided by major service providers within the JISC IE. This was part of the plan, described in the phase 1 final report, to establish a Geo-data Network⁶. The targets selected were the Archaeological Data Service catalogue and the Data Archive catalogue. Time permitting we also proposed to investigate adding in other services such as the CHCC Census Resource Discovery System and the MIMAS catalogue. The work would involve configuring Go-Geo! so that it could search the respective Z39-50 targets and request records. Because the metadata are held in different formats to that used by Go-Geo!, the records would need to be mapped into the Glgateway HE/FE metadata format for display to users. Furthermore, because these Z-targets were set up for different purposes than the one for which it was wished to use them, part of the work would involve customising the respective Z39.50 searches at the portal end to deal with the unique characteristics of each Z-target. Funding was available to pay for staff time at the Archaeological Data Service and the Data Archive if it was found that changes were needed in the Z-targets themselves.

2) Addition of user profiling.

Phase 2 of the project had raised the issue of user profiling and customisation. Users would gain certain benefits by, for example, extending the functionality of the portal to allow users to save queries and search results. It was envisaged that there might be other uses to which user profiling could be put once it was looked at it more closely.

3) Improvements in the spatial searching of remote databases and resource catalogues.

Limitations in Z39.50 means that it is not possible to say whether resources being searched for overlap, cover, exactly match or fall within the user's area of interest. This means search results often include erroneous items. Ways of supporting polygon searching instead of searching using minimum-bounding rectangles (MBRs) are also required to improve the accuracy of results. In the phase 2 project a mechanism was implemented for sorting the result of searches by spatial relevance. However, the project team wanted to see if this could be improved further in the phase 3 project. It was proposed to investigate the problems, identify solutions and where practical and resources allowed, implement them.

4) Development of a facility (the 'Geo-Locator') similar to RDN-I⁷ that would allow the Geo-data Network to be searchable via the Gateway from other Portals.

As early as phase 1, there was a requirement for the Geo-data Gateway to be searchable from other portals (subject, media and institutional) and third party web sites such as those operated by data centres. Users accustomed to accessing resources from a preferred site could do so while at the same time query and retrieve geospatial metadata from the Geo-data Gateway. It was proposed that this would work in a similar way to the RDN-I and be called the 'Geo-Locator'.

In this project it was proposed to actually implement such a facility. However, at the request of the JISC, the team were asked to look at the use of portlet technology as a way to achieve this rather than the mechanism used by the RDN. The idea was that a portlet would be published by the Go-Geo! portal. Other portals would be able to find and bind to it. The Go-Geo! server would also implement a simple web service to which the portal invoking the portlet would be able to send a search request using SOAP. Go-Geo! portal would return an XML document containing the result of the search.

⁶ The architecture set out in phase 1 of the Geo-data portal project envisaged a Geo-data Gateway being established to underpin the portal. This would comprise compliant metadata service providers, data providers with metadata but no metadata services and data providers with data who require their metadata to be hosted by service providers. Users would search from a single point (a 'hub') and a search would be directed towards a set of distributed services (the Geo-data Network) hosting metadata via the Gateway. The Geo-data Gateway would appear as a 'virtual' catalogue in the Glgateway service, thus providing a central access point to metadata about geospatial data within UK academia.

⁷ RDN-i allowed site administrators to embed the RDN query engine within a standard HTML page within their Web sites.

In the phase 3 project it was also planned to look at possible new developments and enhancements to the portal. This included RSS (Rich Site Summary) for news feeds, a web-based discussion tool, web page searching filtered by geography, pushing of records to users when new, relevant records are added (alerts). These were to be reviewed in the context of user needs. Those identified as of potential use to users would form part of an on-going programme of development of the service after launch.

Roll out of trial service

A central aim of the project was to roll out a trial portal service to institutions in UK HE and FE supported by an on-going programme of evaluation (see below). As well as giving the project critical feedback on the demand for such a service, the project team wished to test the service with respect to performance (response times and scalability) prior to its launch as a full service. The trial service would be operated as if it was a real service although the resources available to the project would limit promotion and user support. Once 'operational' it was recognised that staff effort would be required to undertake bug fixing and general maintenance of the service.

From the start it was understood that work might be required on the portal before it could be released as a trial service. As will be reported below, this turned out to be more than planned for at the start of the project. It also was not anticipated that interoperability and metadata standards problems would arise that would need addressing alongside the technical development.

User and stakeholder engagement and evaluation

Continued user and stakeholder engagement was important. The project had built up support from a large group of academics and researchers in previous phases. In the phase 3 project, the team wanted to

- facilitate further communication between members of the project, stakeholders and the wider community;
- promote the Go-Geo! service and the possibilities of a service and to do this through demonstrations at conferences and general awareness raising etc;
- develop further relationships with relevant initiatives in geospatial searching specifically the AGI, the RDN Geography and Environment Hub and the Resource Guide for Geography and the Environment, but also the RDN network in general, other JISC services and JISC IE development programme projects, and
- ensure that project learning was disseminated to the HE/FE and GI communities.

A key objective of the project was to carry out a user trial to seek feedback on the trial service and assess future needs of the user community once a service was launched. It was planned to monitor usage of the service through analysis of log files and the searches conducted.

A study into the requirements for data distribution

Phase 2 identified that a major barrier to the sharing of geospatial data was the lack of a data distribution mechanism that could be used by geospatial data producers and custodians. Feedback from the demonstrator project found that provision of metadata would be limited by concerns data providers/custodians had about how they would distribute their data. Individuals/departments are not organised in such a way as to be able to manage distribution of data to people who approach them. Phase 2 found that a cost effective and easy way for those holding geospatial data to share their data with others within UK tertiary education was required. In this phase it was therefore planned to undertake a study to investigate how this could be achieved. Two approaches were to be investigated.

- a) Through a peer to peer (p2p) application. Data holders/custodians would set up a p2p server on institutional machines and store data in them. (Probably at an institutional or department level.) Metadata would be published announcing the existence of servers and data. Metadata could also be published to the Geo-data Gateway. Users would use a p2p client to search for data or the Go-Geo! Portal and then can locate a copy of the data and download it to their machine.
- b) Self-archiving service. This involves establishing one or more self-archiving services which data producers/holders can use to publish data for use by others. It would provide mechanisms for

users to provide data, metadata and accompanying documentation (pdf, word files etc.). Metadata would also be published, possibly using OAI, and therefore harvested and stored in the Go-Geo! cache.

These two approaches would be compared with existing centralised data archiving services. The Department of Informatics at City University, who have an interest in p2p technology and geographic information, were to act as consultants in the study. The work itself would be led by the Data Archive. The report entitled “Go-Geo - Data Distribution Study” which accompanies this report, provides more details on the aims of the data distribution study and the methodology.

Implementation

The project partners were the EDINA National Data Centre and the UK Data Archive. EDINA acted as the lead partner for the purposes of project administration and finance. Overall, responsibility for the project rested with senior staff of EDINA and the UK DA. The project was co-ordinated by the Project Director, Mr Peter Burnhill, and Project Manager (0.1 FTE), Dr David Medyckyj-Scott, based at EDINA. The project consortium was originally divided into two teams: a Project Co-ordination team and a Development and Evaluation team. However, after the launch of the trial service, a third team formed with responsibility for the operation of the trial service and content development.

A mixture of telephone and video conferencing, supplemented by more conventional face-to-face meetings were successfully utilised during the project.

The work itself was split between 5 work packages (see Figure 2).

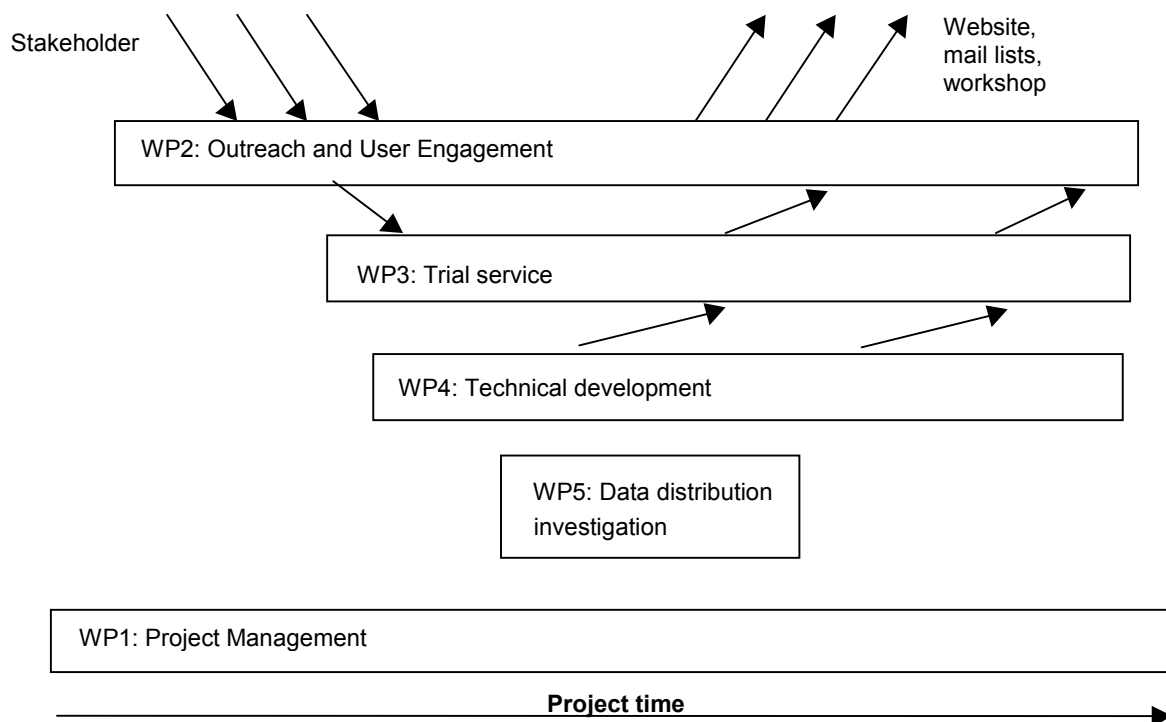


Figure 2 Overall Project Structure

Following the JISC project guidelines, a Detailed Project Plan was developed which set out the objectives of each work package, the outputs, tasks and quality assurance.

The funding received by the JISC meant that the following resources were available to the project.

At EDINA

- 10% of a Project Manager
- 60% of a Software Engineer

70% of a Software Engineer

At the UK Data Archive

60% of an Information Officer

There was also a small amount of funding to pay for the involvement of the Department of Informatics, City University and contributions to the Archaeology Data Service and UKDA for staff time to modify their Z- targets to allow cross-searching. The staff costs of Peter Burnhill (EDINA) were not charged to the project nor was EDINA Helpdesk staff time (the Helpdesk acted as a first point of help for users of the trial service). The majority of Hilary Beedham's (UKDA) time was not charged to the project. After the launch of the trial service, a member of the EDINA staff also assisted in the identification and QA of useful resources for inclusion in the portal. Again, this time was not charged to the project.

Stakeholders were initially contacted in August 2003 to announce the start of a new project phase. Additional announcements were made at the launch of the trial service and subsequently to inform users about the second evaluation period and the prize draws which users were invited to enter after completing the user evaluation questionnaire.

News of the trial Go-Geo! service and invitations to complete the data distribution questionnaire was also disseminated to contacts for the sister project, the Metadata Promotion and Creation Initiative. Contacts for the metadata initiative were identified from university department information, from a variety of disciplines and included both potential users and creators of geospatial data. This combined effort between the two projects allowed for a greater number of contacts to be made, without overburdening stakeholders with the same information.

Contact with users was made via the project JISCmail list and the stakeholder list as well as personal emails, in particular for the purpose of inviting stakeholders to complete the portal evaluation and data distribution questionnaire. Invitations to complete the surveys were also disseminated via the following mailing lists:

- ESDS site reps and website (<http://www.esds.ac.uk>);
- IBG Quantitative Methods Research Group (<http://www.ncl.ac.uk/geps/research/geography/sarg/qmrg.htm>);
- Geographical Information Science Research Group (GIScRG) (<http://www.giscience.info/>);
- GIS-UK (<http://www.jiscmail.ac.uk/lists/GIS-UK.html>).

Dissemination of the project was achieved through email, mailing lists, web sites, workshops and conferences.

Details of both the project and the trial service were advertised on the UKDA, ESDS and EDINA web sites as well as the Go-Geo! project web pages. Announcements were also distributed via the UKBORDERS and ESDS mailing lists, the jisc-development, portals, lis-jibs-users and interoperability JISCmail lists, as well as a number of internal JISC lists, including other portal projects. An article about the project was included in the October issue of WYVERN, the University of Essex newsletter and in the UK Data Archive newsletter.

Details of the project were also disseminated to metadata workshops contacts and workshop participants. Attendance at conferences, such as GeoSolutions 2003, AGI Scotland, BURISA 2004, GIS Research UK (GISRUK) and IASSIST 2004, allowed for the distribution of promotional material and for face-to-face discussion with potential users and data creators. Presentations were also made at various conferences (e.g. the ESRI Geo-Portals Workshop, London; the ESDI State of the Art Workshop on Portals, Warsaw) and workshops and papers published.

All documents produced or commissioned by members of the project were designated an Internal Technical Document or a Project Document. All Internal Technical Documents, Project Documents and minutes of meetings are stored in a password-protected area on the UKDA website. This area of the website is only accessible to project members. Certain internal documents, presentations and papers are available from the project web site.

Trial Service Roll Out and Operation

Prior to roll out of the trial service further development of the portal was undertaken in response to feedback received during phase 2. This included the following.

- The portal home page was redesigned and navigation of the service improved.
- GEO attribute set compatibility in the Go-Geo! search engine was extended to handle all the Go-Geo! web interface requirements and future integration with GIGateway as an academic virtual catalogue. GEO attributes now supported include date, place name, title, abstract, keyword, free-field, and bounding-box queries. There is also support for boolean query logic of search fields. All development was undertaken in Perl and interacts with the Net::Z3950::SimpleServer library.
- An improved email/print facility was implemented.
- Users were offered more search results by the removal of an artificial optimisation that limited results to 100 records and the way records were displayed was improved to more accurately display Go-Geo! metadata elements.
- The portal content was updated, including the resources pages, the help facility and metadata creation pages.
- Several bugs and issues relating to interfacing with the geoXwalk gazetteer server were fixed, and request response times improved.
- Detailed statistics and usage logging mechanisms were written.
- Service administration mechanisms were put in place including boot scripts and cron jobs to deal with server housekeeping for the portal.

The trial service was launched on 19 November 2003. This was two weeks later than planned due to the need to undertake additional development work.

Accessibility testing of the Go-Geo! portal prior to launch as a trial service resulted in the identification of a number of problems. In part this reflected a difference in interpretation of the accessibility guidelines relating to the use of JavaScript between the web designer of the site, the programmer and the EDINA web master. As a consequence, it was decided to revisit the design of the site. It soon became clear that it would not be a case of just redesigning the site, rather it required a separate accessibility-enabled version of the portal to be built. This was done although the advanced search facility was excluded because of the difficulty of supporting user interaction with the advanced search web pages without the use of JavaScript. This parallel site was released in late-February 2004. This was an additional work item not identified during the project-planning phase. Undertaking this work had a significant knock on effect on other work items. There still remains the problem that only part of the search aspect of the site meets accessibility guidelines.

Alongside the launch of the portal, it was necessary to put in place end user support. First line support was provided via the EDINA helpdesk. Use and technical queries were dealt with by staff manning the helpdesk and, where necessary, were passed onto the developers. Content or project queries were passed to Information Officer at the UKDA.

It was acknowledged very early on that it was important that the site was designed such that users would want to make frequent return visits. Doing this would help establish the site as a key resource. To encourage users to return to the site again and again, content needed to be useful, up to date and added to over time. This was achieved through ongoing updates to resource pages in the portal particularly the news and events pages (see below). The project was fortunate in that it was able to call upon the geo-data services and user support teams at EDINA for input.

Evaluation of the Go-Geo! Trial Service

As in previous phases, a range of quality assurance and formal and informal formative evaluation instruments were employed to ensure that the project was meeting the requirements of the stakeholder community. An evaluation plan for the project was developed and project staff prepared an evaluation framework to carry out testing and evaluation of the service.

The first evaluation period began straight after the launch of the Go-Geo! trial service in November 2003. This first evaluation period ended 27 February 2004. A second evaluation ran from 1 March to 30 June 2004 (see below).

User evaluation was carried out via an online survey and through hands on sessions. A short survey was posted on the Go-Geo! portal site, with a link included from the project web pages. Go-Geo! portal users were invited to try out the portal and to complete an online evaluation questionnaire, reporting on the usability and usefulness of the portal. The survey concentrated on asking if the users liked the portal and if they would want to use it again. All academic respondents were entered into a prize draw which was held at the end of each evaluation period.

Academic institutions were also invited to participate in a more detailed evaluation of the portal. Originally the plan had to been to involve six institutions. However, by undertaking evaluation in conjunction with the workshops held jointly with the 'Geospatial Metadata Promotion and Creation Initiative' project, a much larger scale evaluation was possible.

During the workshops participants were invited to try out the portal. The 'hands on' session allowed users to identify navigational and procedural issues and to establish the effectiveness of Go-Geo! features. The participants were then asked to complete a detailed usability questionnaire in order to provide feedback on the functionality, usability and design of the portal. The questionnaire was aimed at identifying issues in the interface design and layout as well as functionality, usefulness and usability. The participants were given 45 minutes to try out the portal and to complete the questionnaire. All participants were invited to be entered into the prize draw for completing the questionnaire.

The institutions targeted to host the evaluation sessions were identified from the Go-Geo! stakeholder list, members of the project JISCmail list, metadata initiative contact list and from users of the demonstrator portal service. A contact was identified at each institution who distributed posters and emails advertising the workshops. In some cases, if no key contact could be found at an institution, potential participants were emailed personally. In all, evaluation sessions were held at eight universities and in some cases more than one university was represented at a workshop.

As the user evaluation sessions were being held as part of the metadata initiative workshops it allowed for a larger number of participants to attend and to ensure that participants are not being overburdened by attending two separate sessions. This combined effort also ensured that a wide audience was reached, including both users and metadata creators, from a variety of disciplines.

Details about the evaluation methodology can be found in the 'Go-Geo! Phase 3 Evaluation Report' which accompanies this final report.

The Data Distribution Study

There were two parts to the data distribution study: a requirements survey and a review of peer-2-peer, self-archiving and traditional (centralised) archiving.

A change in the timing was necessary for this work item due to an overrun of one of the work items of the geoXwalk phase 3 project which the Information Officer at UK Data Archive was also working on. There were further delays due to the member of staff from the Department of Informatics at City being unavailable and then not responding to emails and phone calls. In the end the relationship with the individual had to be terminated. Unfortunately, the project was unable to identify a suitable replacement i.e. with knowledge of the use of p2p for research data sharing. The Programme Manager of the JISC Portal Programme provided details for some people and these individuals were approached for advice during the study. As a consequence of the delay and problems, with the agreement of the Programme Manager, the study was scaled back to just a requirements study and a user survey.

A requirements study was undertaken to investigate how services and organisations currently store and distribute data and what they might consider they would use in the future. The requirements survey was carried out electronically due to time constraints for completing this work package. Ideally, if time had allowed, face-to-face interviews or focus group discussion sessions would have been conducted, as these could have produced a greater response rate.

The survey was disseminated via the web, workshops, mailing lists and by emailing UK GIS lecturers and other key academics. The investigation focused on access, licensing, funding and policies.

Outputs and Results

The portal today

Figure 3 shows the home page of the Portal as it looks at the end of the phase 3 project.

Figure 3 The Home Page of the Portal

Resource Panel

The home page is divided up into a number of panels. The left-hand side panel provides access to a quality-controlled set of links to useful resources organised as channels and sub channels. The channels are structured as follows.

- **Case studies (of use of geospatial data)**
 - * Archaeology
 - * Environmental and Geographical Science
 - * Government
 - * Health
 - * Land Use/Planning
 - * Transport
- **Data format guides**
- **(List of) Major Data providers**
 - * Academic Organisations
 - * Commercial Organisations
 - * Public Organisations

- **Geoterminology**
- **Software/Online services**
 - * Online Services (UK only)
 - * Data Translators
 - * Commercial Software
 - * Free Software
- **Research Materials**
 - * Bibliographic Resources - links to bibliographic resources for GIS-related research materials.
 - * Books and Printed Guides - sources of GIS-related printed books and guides.
 - * Journals - educational geographical information journals.
 - * Magazines and Newsletters
- **Learning Materials**
 - * Bibliographic Resources - links to bibliographic resources for GIS-related learning materials.
 - * Books and Printed Guides - sources of GIS-related printed books and guides.
 - * Online Guides and Tutorials - educational GIS information and online tutorials.
- **Subject Resources**
 - * Subject Gateways - further information and links to subject gateways.
 - * Resource Guides - link to the Resource Guides on the JISC web site.
- **Training courses**
 - * Higher and Further Education Courses - information on undergraduate and postgraduate GIS courses offered by UK Higher and Further Education institutions.
 - * Private Courses - GIS courses run by private organisations.
 - * Short Courses - links to short GIS courses run by academic organisations.
- **News**
 - * Selected News Items
 - * Links to GIS-Related News Resources
- **Conferences/Events**
 - * Selected Conferences and Events
 - * Links to sites which list Conferences and Events
- **(List of GI related) Discussion groups**
 - * Environmental Science and Geography
 - * Census, Population and Health
 - * GIS and Mapping Science
 - * Library and Information Science
 - * Software and Technology
- **(List of GI related) Organisations**
 - * Commercial Organisations
 - * Government and Public Organisations
- **(List of GI related) Education/Research groups** (GIS-related research projects within higher and further education)

On a weekly basis, the channels on Training Courses, News and Conferences/Events are updated. Sources include English language email list servers relevant to geospatial data, online news sites and printed magazines. The channels on Major Data Providers, Organisations, Discussion Groups, Education/Research, Subject Resources and Software are reviewed and updated regularly. User suggestions, derived from feedback questionnaires and emails to user support, were analysed for content development and changes in structure. As a consequence, the channels on Learning Materials and Research Materials were reorganised.

A list of journal titles which publish English language papers describing GIS techniques and the use of geospatial data, with links to the relevant parts of the publishers web sites, were put up under the Research Materials channel. A review was undertaken of free online geospatial services in the UK and a list created of these with short descriptions about each and links. In the last few months the project has approached major academic publishers and requested details on the geospatial and GIS books they publish. At the time of writing, these are being authored into a series of web pages with entries organised by topic. Each entry will contain the title of the book, the names of the author(s), date of publication, a short abstract, a picture of the cover of the book and a link to more information on the publishers web site. The publishers now plan to send information to the team on books they are about to publish.

Web forms are accessible throughout the site, which allow users to tell us about resources they would like to see mentioned in the Portal, events they are holding and news items.

With around 500 links to maintain, a link checker has been implemented which once a week checks all links are still valid. If a broken link is found, an email is automatically sent to one of the project staff notifying them that a link needs fixing.

Figure 4 shows an example of a channel page.

The screenshot shows the 'Go-Geo!' website interface. At the top, there's a green header with the site name and navigation links: [Home Page], [Back], [Accessible Site], [About Project], [Help], [FAQ], and [Athens Login]. Below the header is a sidebar with a 'Resources' section containing links for Case studies, Data format guides, Major data providers, Geoterminology, Software/Online Services, Research Materials, Learning Materials, Subject Resources, Training courses, News, and Conferences/Events. The main content area is titled 'Geoterminology' and contains the following information:

- Geoterminology**: This page provides sources of information for geoterminology. Includes links for A-Z and K-Z.
- AGI GIS Dictionary**: This online dictionary of GIS terms was created by the Association for Geographic Information and the University Of Edinburgh, Department of Geography. The dictionary includes definitions for 980 terms compiled from a variety of sources which either relate directly to GIS or which GIS users may come across. The dictionary is also supplemented by 52 diagrams. URL: <http://www.geo.ed.ac.uk/agidict/welcome.html>
- ESRI Glossary of GIS Terms**: ESRI have produced a free online GIS glossary and a dictionary of GIS terminology which may be purchased from their online store. URL: http://gisstore.esri.com/acb/showdetl.cfm?&DID=6&Product_ID=3341&CATID=12
- Geocommunity**: Geocommunity provide a free online GIS glossary, which can be accessed from their web site. URL: <http://www.gisdatadepot.com/helpdesk/glossary.html>
- GeoExplorer**: The GeoExplorer dictionary provides coverage of the most common terms used in physical, human and environmental geography. Some of the definitions are further enhanced by a diagram or photograph. URL: <http://www.geoexplorer.co.uk/sections/dictionary/dictionary.htm>
- Back to Top**: A link to return to the top of the page.
- NH GRANIT**: The New Hampshire Geographically Referenced Analysis and Information Transfer System (NH GRANIT) web site, provides access to a GIS-related glossary.

Figure 4 Sample channel page

Search Interfaces

The middle panel of the home page supports three tasks: a simple 'google-like' search for metadata describing geospatial datasets; a link to an advance search area where the user has more control over the parameters of a search and the type of resource to search for (Figure 5), and a link to an area which provides information and guide notes for the creation and supply of metadata records. The latter area will be the location of a simple metadata creation and editing tool to support the online creation of metadata records (see below).

Figure 5 How users control the type of resource to search for under Advanced Search option

The final panel on the right-hand side is the area where news about the service is announced and links to recent news items.

The geospatial metadata catalogues currently available for cross-searching are

Within the Geo-data Network

Go-Geo! catalogue (includes metadata about datasets held by EDINA and MIMAS and individual researchers)

AHDS History Data Service catalogue

AHDS Archaeological Data Service catalogue

Within GIGateway

Ggateway catalogue

British Geological Survey catalogue
 Centre for Environment and Hydrology (CEH) catalogue
 Ordnance Survey catalogue
 QinetiQ catalogue
 Local Authorities catalogue
 Central Government (IGGI) catalogue
 British Atmospheric Data Centre catalogue
 NERC Earth Observation Data Centre catalogue

In addition there are a number of other resources searchable via the advanced search. These are

Maps:	COPAC
Images:	British Geological Survey JIDI Photographic Collection
References:	
Books, journal articles etc.	COPAC (filtered by reference type)
Projects:	ESRC REGARD database

During this phase, the capability to cross-search the British Geological Survey JIDI photographic collection was added to the advanced search. The records describing the images contain explicit geographic references, which allows spatial searches to be performed. The records users view include a postcard size version of the BGS image. A link is provided to the BGS web site to place an order. Figure 6 is a screenshot of a BGS image record as displayed by Go-Geo!.

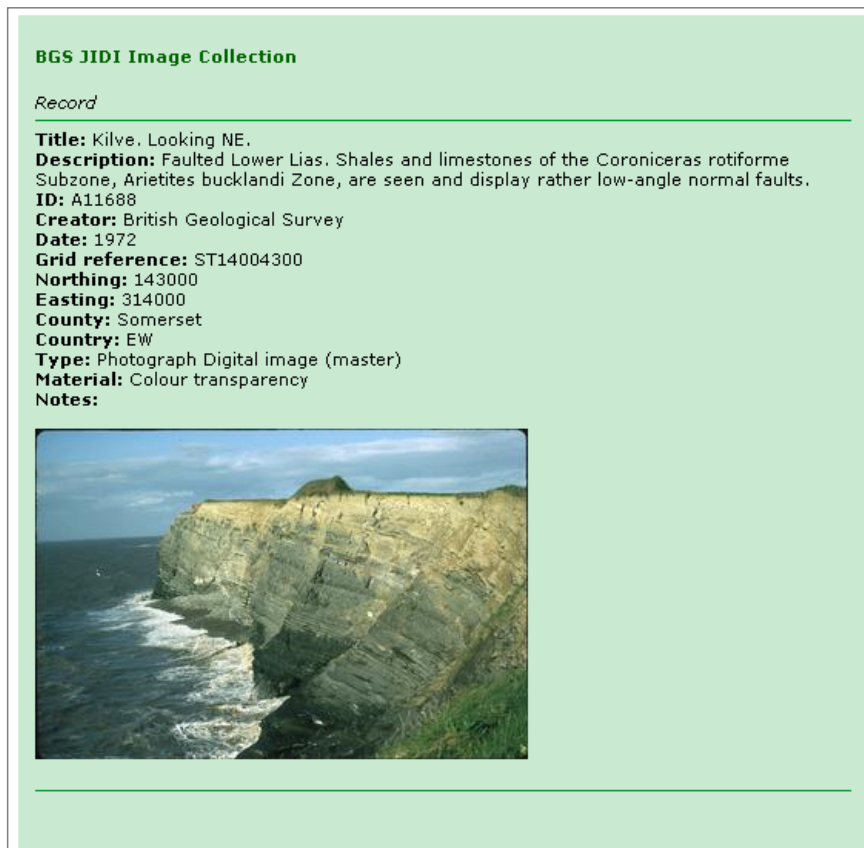


Figure 6 Screenshot of a record from the BGS JIDI Photographic Collection

Despite a variety of unanticipated problems arising e.g. metadata record format, at the very end of the project records from the AHDS Archaeological Data Service catalogue became searchable from Go-Geo!. Unfortunately, due to a slippage in the timetable for redevelopment and roll out of the new UKDA catalogue, the team were not able to link in the UKDA catalogue before the end of the project.

All being well however, the UKDA catalogue should be searchable from the Go-Geo! portal by the end of September.

The inclusion of Global Change Master Directory, operated by NASA, was also investigated as a way of providing metadata on global and regional geospatial datasets whose geographic coverage includes the UK. This raised again the problem of how to deal with catalogues whose metadata is in a format other than the FE/HE Application Profile of the Glgateway Metadata Guidelines. This is discussed more fully in the issue section of this report.

Two Glgateway catalogues were added during the phase 3 project, these were the British Atmospheric Data Centre and NERC Earth Observation Data Centre.

Use of the portal

Use of the service⁸ has been monitored since its launch.

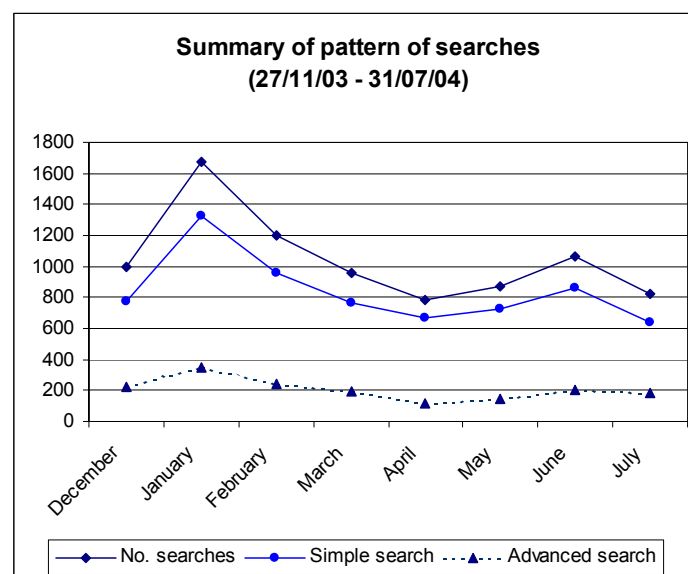


Figure 7 Summary of pattern of searches

The simple and advanced search metadata facilities are the most used aspect of the service. January was the month with the highest number of searches conducted (1,676) after which the number declined with another small peak in June. Use was lowest during those months in which university holidays fell.

More simple searches were conducted than advanced searches, possibly reflecting a 'google' mentality on the part of the user. Searches for resources other than geospatial data are low e.g. COPAC has only been searched 214 times, the ESRC Regard project database 27 times and the BGS Images catalogue 170 times. There are a number of reasons for this. First, these resources are only searchable from the advanced search. Unless a user selects the advanced search option, they may not be aware that these resources can be searched. Second, the ability to search for these sorts of resources is not a common function in other geospatial data discovery services. Users may need time to get used to the idea that Go-Geo! can be used for this purpose. It is clear however that users do want access to related resources since reference to the web pages where resources are listed is quite high. The top five most used channels were

- * Case studies
- * Software particularly Open Source and free software
- * Major data providers
- * News items

⁸ The web logs were modified at the end of November, so the first few weeks of use are not available for reporting.

* Data formats

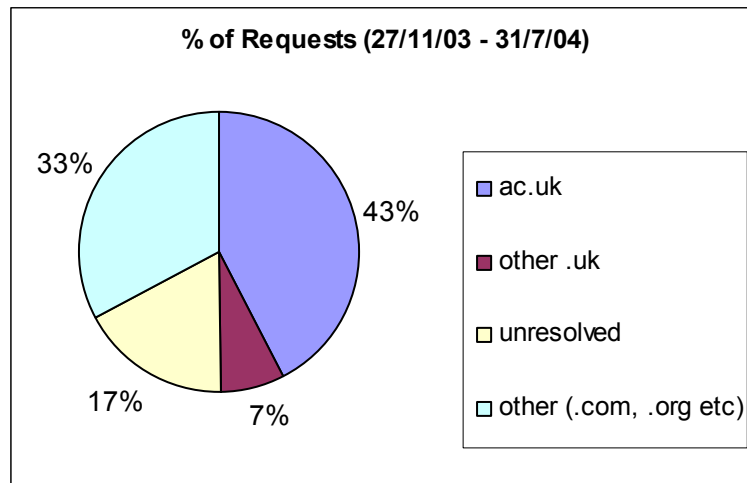


Figure 8 Breakdown of requests by domain

Figure 8 shows that the majority of users of the Portal come from the .ac domain i.e. universities and further education institutions (43%). During the first 3 months the figure was 54% indicating that individuals outwith UK academia have discovered the portal⁹. However, given that universities and further education institutions are the target group for the service, the figure of 43% is reassuring. Users come from a wide number of institutions. The top ten institutions using the portal across the period were Edinburgh, Durham, Newcastle, Cambridge, Liverpool, Manchester, York, Imperial College, Strathclyde and Cambridge. Interestingly, the British Geological Survey was one of the biggest users over the 8½ months that the Portal was available. Within the 'Other' category are users from 24 countries outwith the UK.

User feedback

[Full details on the findings from the evaluation work can be found in the report accompanying this final report entitled 'Go-Geo! Phase 3 Evaluation Report'.]

By the end of this phase of development, 79 responses were received from the online questionnaire and 57 responses from the evaluation sessions. In general, the participants liked the look and feel of the portal, especially the resource channels. Participants found that it was clear how to conduct searches, how to display record details and how to navigate through the portal. Overall, the participants agreed that the portal was easy to use, that they would use it again and that it could be an extremely useful facility.

⁹ Note that many of the unresolved addresses could also be universities.

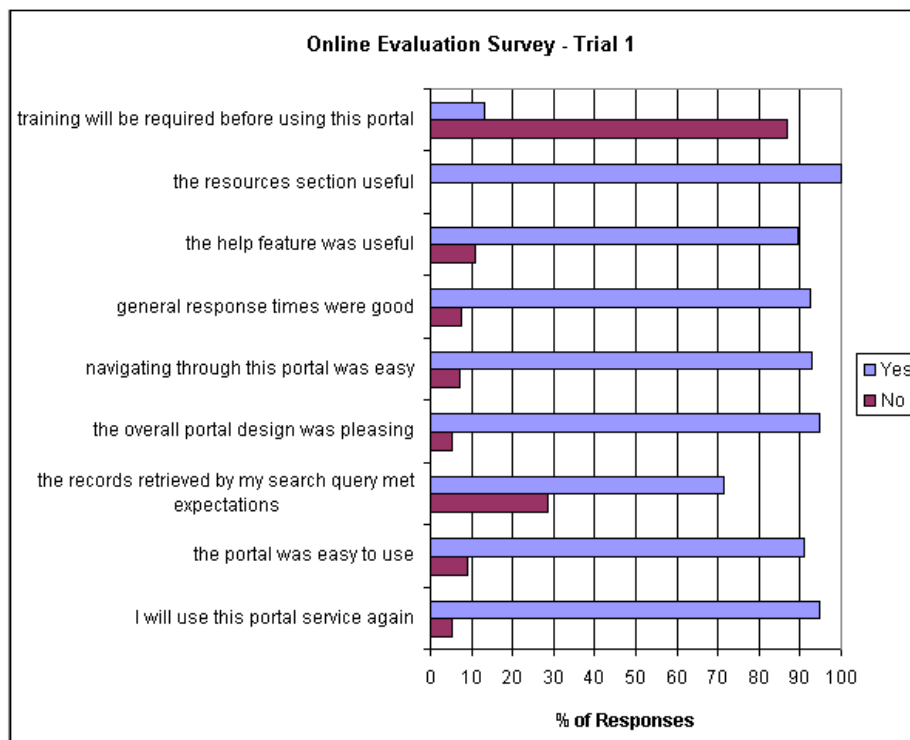


Figure 9 Results from the Online Evaluation Survey – Period 1

The main areas of concern for users were the inconsistency of match between the results and the search query and the lack of completeness of metadata records.

Although there is information on a large numbers of geospatial datasets, the information provided about the datasets is often incomplete. Users found that there was missing information about contact details, access information etc. This is a direct result of the poor quality of some of the metadata. However, because the metadata is found by cross-searching catalogues managed by third parties, the project team has no control over the quality of metadata held in these catalogues. Whilst that found by crossing the HE operated catalogues is comparatively good, some of the metadata in the catalogues under the control of, for example, government agencies is very poor. We understand the AGI, who are responsible for Ggateway, have recently employed a member of staff whose job it is to encourage organisations running catalogues to QA their metadata. It has been suggested that the metadata results discovered through the portal be ordered and ranked based on a measure of 'completeness' of the record. That is, records which are more robust in terms of content appear closer to the top of the results listing than records which are barely minimal. This would have to be an option the user selects to override the default of ranking by topic and spatial relevance.

The quality issue also affects the quality of the searching and thus the results presented to the user. In some cases no results are found, but in others users found that either irrelevant information was retrieved e.g. North American names, or the results were too broad e.g. covered the whole of Scotland when only data for Glasgow was being sought. One user reported that when he searched COPAC for York or Yorkshire with the 'geographically expand' button he received no results, but when York was typed into the simple search it resulted in lots of records being found. We are trying to address these issues through working with geoXwalk project. The geoXwalk gazetteer underpins searching within Go-Geo!. Using geoXwalk allows services that otherwise could not be searched geographically to be spatially enabled. However, this can only be work so far. For example, a number of the geospatial metadata records have poorly specified geographic extents. A search for datasets for Dorset will return datasets for Cumbria because the records describing the Cumbrian datasets have their extent defined as the whole of the UK. In another case a resource may include Boston as a place name term but with nothing about which specific Boston i.e. Boston, Mass. or Boston, Lincolnshire. This makes it difficult to filter out erroneous records. Ideally, the records held within the catalogues need to be explicitly and accurately geo-referenced.

There is also frequently a mismatch between what the service the portal currently provides and what users expect it to provide. Many users were expecting more than simply the reference details i.e. metadata. They were disappointed by how few datasets were online and that there were no links directly to the data providers from the 'Who' page. It was felt that the portal perhaps encouraged the idea that it would allow direct access to the data, not just the home pages of organisations.

The most frequent suggestions for improvement included: the ability to refine the search query to search within a result set; the addition of a postcode search, and the inclusion of an online tutorial was suggested as a means to enrich the help facility. Extending the portal to include areas outside Great Britain was also suggested by a number of users who felt that this would enrich further an already very useful resource.

Another major concern was the lack of a guarantee of continued access to the portal in the longer term (this re-iterated a key finding from the phase 2 project).

Status of the development work

Improving the spatial searching of the catalogues

The geospatial metadata indexes cache has been re-designed so that geospatial footprint of the spatial data described by the metadata is stored as a spatial geometry type rather than real types. This allows Go-Geo! to use the geospatial extensions of the MySQL database. (These features were not stable at the time of Go-Geo!'s original implementation but have since been finalised in the database code.) This has allowed additional capabilities for selecting search results based on map queries to be offered, other than the default geographic functionality of Z39.50's GEO profile. For example, it is now possible to restrict results to those geospatial datasets that fall within a map boundary rather than merely results within or encompassing a boundary. This has improved the quality of results measurably by excluding many poorly specified metadata records which incorrectly state a UK wide geographic boundary.

Despite this improvement, users has said that they expect to see a small sample of results, whereas Go-Geo! tries to return all possible matching results but sorted by order of relevance. This accounts for some users complaining that the results on the second or nth page were irrelevant. Also, where there are no highly relevant results in the databases being searched, for example, if the user has entered an obscure town as the geographic query area, then those results that appear will all have a low relevance, irrespective of the technology used to sort them.

The following steps could be deployed to try to improve this situation:

1. Provide the option of further geographic search functions that eliminate geographically irrelevant results, rather than just sorting them by relevance. This should satisfy people expecting only a small subset of results. However, it is unclear how this could be done without increasing the complexity of query specification in the interface.
2. Introduce a proximity element to the geographic relevance. This would help minimise the inappropriate bounding box data quality problem.
3. Provide a numeric relevance score of results to indicate the relevance sorting that is taking place, and help explain to the user why their results may appear to be irrelevant.

Development of the 'Geo-Locator' Portlet

A "Portlet" is the new term for a plug-in component to a "Portal". It consists of a Web service run by one organisation that functions remotely on the Portal Web site of a different organisation. A classic example is a Weather report box which can be operated by Weather forecasters but runs on an existing Web site Portal such as yahoo.com.

The common misconception gained from examples such as this is that portlets are only useful for services requiring small screen real-estate. Delving into the relevant specification reveals that it is equally applicable to search other information finding services, such as Go-Geo!, which demand the

whole screen or a larger amount of screen estate to display results. In these cases the portlet service can start as a small box or link on a portal that contains other services, but claim a "maximum" or "solo" window-state following the users' selection or search of the service.

"Web Services" initiatives using SOAP, WSDL, and UDDI have approached the same subject of providing remote services to Portals, but they have typically only sent XML data streams to the Portals, which must then repackage the XML for presentation as HTML on their Portal Web sites. This step often requires unique knowledge of the data, and considerable time spent working with the XML from each such service.

The WSRP Portlet initiative (http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=wsrp) proposes a format that allows the remote Web service to send HTML directly to the Portal so that it can be included directly into the source HTML of a Portal Web page. The underlying idea is to make the use of Web services more convenient for the Web Portals by taking out the time consuming process of handling XML themselves.

While a nice idea, the implementation turns out to be surprisingly difficult. Complications encountered include the need to rewrite URLs so that both portal and portlet interact properly with the user, and the need to separate interactions between users and the portal into two phases, a state-change phase and an HTML rendering phase. The standard envisages that software developers will take care of these complications, and that at least Portals will be blissfully unaware of them, even though portlet developers will require some understanding of the standard themselves.

Go-Geo!'s portlet code provides for the following states: the search page, results page, record page, and ambiguous place names page. An example interaction is as follows: starting with the "search page" state - which will exist on a Web page among other portlets and consist of a search form with an input box - if the user enters a search in the Geo-Locator portlet, the portal recognises the input and makes a SOAP XML call ("performBlockingInteraction") to the Geo-Locator portlet to change the state of the portlet following the form input. The portlet will check if the place name entered is recognised and unique, and change the state to the results page. This state is sent back to the portal via a SOAP XML call, but no HTML is yet returned. The portal will then make another SOAP XML call to "getMarkup" from the portlet in a second step to retrieve the HTML to be displayed. At this point the portlet will query the Go-Geo! search engine and send back a display of results for the first 10 records. It will also request a maximised window display on which to display the search results which, if accepted, will result in all other portlets being minimised/ignored.

Third party portals that decide to make use of portlets will need a software infrastructure that supports the standard. U-Portal (<http://www.uportal.org/>) is an example of one such infrastructure. Adding portlets to an existing ad-hoc Portal infrastructure would be difficult unless considerable time were devoted to it. This has created a problem for the Go-Geo! team. The original intention of the Geo-Locator, when it was supposed to be RDN-Include like i.e. based on CGI, was that other JISC service providers, such as MIMAS, could embed the search facility in the web pages of their services; indeed with one catalogue provider this was part of the agreement for us to be able to cross search their catalogue. None of these providers are operating web sites based upon a formal Portal software infrastructure such as U-Portal. As things stand, none of them can use the Geo-Locator portlet facility.

Services that wish to offer a portlet interface to an existing service, such as the Go-Geo! search service, would be wise to study the specification ahead of their service implementation because it makes specific demands on how the service must be developed. The HTML of the service requires the right Cascading Style Sheet (CSS) classes, and the service must be designed to operate in two steps: one step to change the operational "state" of the service, and the next step to output the HTML corresponding to that state. It is something of a misnomer to assume a portlet interface can be retroactively fitted to an existing service, and in fact it is almost certain that new HTML and underlying code will need to be written.

The WSRP Portlets format holds much promise, but there is a nagging sense that perhaps it demands too much of the parties involved. The specification requires a good grasp of SOAP and XML/RPC in addition to HTML and CSS. The standard is no easy read, sadly leaving examples to a side document (<http://www.oasis-open.org/committees/download.php/1237/wsrp-primer-draft-0.3.pdf>) that itself features subtle bugs and differences with the latest version of the standard. To its credit the standard

is platform independent. Developers of various packages may yet take the complications away from Portal and portlet developers and thereby help it achieve its goals.

User profiling and customisation

The first decision to be taken when adding user profiling to Go-Geo! was what authentication and authorisation mechanism to use. Athens was the obvious choice since it meant that users could use existing usernames and passwords and would not have to be supplied with alternative ones. Furthermore, one of the longer-term goals is to provide a geographic entry point to other JISC services including subscription services. However, users would only be allowed to search those services to which they or their institution subscribed. To do this requires a mechanism to ascertain what those services are. Athens would allow this.

Athens is therefore being used as the method of allowing Go-Geo! users access to proposed customisation facilities and to identify a particular user to the Go-Geo! portal. To login, a user clicks on a link on the Go-Geo! home page. (Note that the link on the current live Go-Geo! website is not active.) This takes them to a remote login page on the Athens authentication server, if the user has not already logged in via Athens SSO. The user then supplies their Athens username and password on this page. This means that the sensitive account information is never transmitted over the Internet.

If the user is authenticated and authorised to use Go-Geo! he is redirected back to the Go-Geo! website. The Go-Geo! web server creates a browser session with an associated randomly generated session ID and also uses the Athens username as a user ID. Any variables associated with the session are stored on the Go-Geo! server. The session has an expiry limit of 1 hour of non-use. Both IDs are also set as temporary cookies in the user's browser, destroyed when the user closes the browser down.

Because the amount of work required once the portal was 'live' as a trial service was underestimated, as far as the implementation is concerned, the user profiling/customisation has not gone beyond the authentication and authorisation mechanism. However, now this is in place, further developments are possible. These include the following.

- 1) Users can be given the facility to provide personal details about themselves e.g. their name, institution and email address. This information can be used to auto-complete relevant fields in the portal e.g. their email address when they want to email metadata records to themselves. If the user created metadata records using the Metadata Creator tool, the details could be used to auto-complete contact details.
- 2) Users could set preferences for how searches are conducted and results displayed. For example, they could set a default map view defining their geographic area of interest rather than use the map of Great Britain, which is the service default. They could also set a default value for the number of records to be displayed in a result list. It might also be possible for them to state what elements of a record they want to appear in a search results list.
- 3) A user could set automatic notifications. Users could request that certain queries are stored and rerun at regular intervals defined by the user e.g. weekly, monthly etc. The results of the searches could be emailed to the user. They might also request to be notified if new entries are added to particular resource channels e.g. new books or journals.
- 4) A primary goal for the future is provide users with the means to geographically cross-search other catalogues and resources (possibly even full text services) within the JISC Information Environment subject to
 - a) either the user or their institution being subscribers to that service,
 - b) the service itself be enabled for searching remotely, and
 - c) permission being granted by the various rights holders to allow the service to be accessed and searched remotely.
- 5) It should be possible to enable users to organise searches in a way that best suits them and store the resulting structure for use on subsequent occasions. This is the idea of an Enquiry Organiser, first proposed in the Phase 1 project¹⁰ (Pradhan and Medyckyj-Scott, 2001). An enquiry organiser

¹⁰ An Enquiry Organiser would be used to catalogue, link and related distributed resources containing spatial and temporal data such as multi-resolution vector data, images, video clips, reports, documents, messages and memos containing location-based information, statistical data and so on. In essence, the Enquiry Organiser would be a geo-enabled folder that organises and relates the spatial and temporal properties of resources that contain various forms of geographic information. These

would assist users to categorise search results, link items for relevance and store the queries and linkages into folders for future retrieval and reuse or for transfer to other users.

Incorporating RSS in Go-Geo!

RSS stands for Rich Site Summary, Really Simple Syndication or RDF Site Summary. Use of the term simple is something of a misnomer. There are it seems, nine RSS standards and a rival known as Atom. An RSS feed is a digest of site content in XML format. The level of detail varies but most sites provide a headline, a URL link and sometimes summary information in their feed. Most major news sites provide an RSS feed so that a user can keep an eye out for specific news items without searching through the rest of the site.

An RSS reader takes feeds from various sites and provides a common interface with which to browse, search and organise the content. If something catches the reader's eye, they can click through to the full articles. Users do not need to worry about the various standards. Most RSS readers or aggregators take care of the specifications. They can also poll feeds for updates at regular intervals and automatically download any changes. There are a number of RSS readers available with varying degrees of functionality. Many are free, some are shareware and some are commercial offerings, for example, ROSA which is an Open Source, customisable RSS Aggregator and Filter from NPG (http://www.jisc.ac.uk/index.cfm?name=project_rosa) which is being used by the JISC.

Information about RSS feeds can be found by searching sites dedicated to cataloguing feeds e.g. Syndic8 (www.syndic8.com). Another searchable feed catalogue is Newsisfree (www.newsisfree.com). It is a little different to Syndic8 in that it offers both free and paid-for services. These and a number of other feed catalogues were searched to see whether there were any feeds that would be of potential interest to users of the Go-Geo! portal. Unfortunately, very few RSS feeds were identified. (One example was in geomatics <http://www.2rss.com/index.php?rss=2383>.) Some of those described in the catalogues were no longer working or were seeking syndication. Furthermore, those that were specifically news feeds were US orientated and contained little of relevance to geospatial data users in the UK. Having said all this, RSS looks to be a useful tool and, should relevant feeds be discovered, the technology should be added into the Go-Geo! portal.

Searching for people

One of the things a researcher wants to know when starting a new piece of research is who is working in the same or a closely related area. Often this requires them to look at sources like Abstract and Indexing services or Citation indexes. However, a researcher may be working in a particularly area but not yet published any academic papers or they may have ceased working in an area some time ago. For a researcher working in a specific geographic location it can be useful to know whether any other academics are working on the same topic or similar ones in the same area or nearby. For example, is there a researcher in another university also investigating soil erosion on the North Norfolk coast or is there anyone studying gannet populations in the Firth of Forth.

One potential source of such information is staff web pages. These often summarise a member of staff's current research interests. They are also often more up-to-date than other types of sources because academics use such pages to describe new projects they are involved in.

Searching staff pages with search engines such as Google or Yahoo is tricky. It is difficult if not impossible to restrict searches to just staff pages. The geographic search capability of such tools is also very poor because place names are just treated like any other text. A researcher entering "Cromer" and "soil erosion" into a search engine may get no results and yet there might be a staff web page which contains the terms "soil erosion" and "Sheringham" (Sheringham being 3 miles west of Cromer).

The project team felt that being able to search for people would be a very useful facility to include in Go-Geo!. However, we were unable to identify any existing tool that would meet the requirements, specifically with regard to supporting geographic searching. A decision was therefore taken to attempt to build a tool. (This work was funded outwith of the JISC Portal funding. However, it is hoped that the facility that was built can be incorporated into Go-Geo! after further testing.)

folders could be browsed, navigated through some geographic context, organised according to spatial or temporal properties, and exchanged with other users.

In brief, the people search facility works by spidering and identifying staff pages within UK academic sites (a "spider" refers to automated software that trawls the World Wide Web collecting information for web search engines). The text of the page is then scanned (parsed) for both postcodes and place names within the UK. If these are found the URL of the page is stored along with plain text content, address, keywords and title and the latitude and longitude values for any place names found.

In the current demonstrator, a user conducts a search via a simple form where they can enter a keyword and/or a location (Figure 11). The user is shown a list of the 'records' found and can choose which 'record' they want to view. The searching has been designed such that not only are pages that mention a specific place name returned but also those pages that are for nearby locations. When the user clicks on a 'record', an HTTP request is sent to the respective University server and the retrieved page displayed to the user. The user can also view a map showing the relevant locations mentioned by each member of staff with respect to the user's query (Figure 11). (The mechanism used by the people search tool is described in more detail in Appendix 1).

SEARCHED FOR: in

Results

1. [dr. valerie haynes](#) **Cairngorm, United Kingdom**
the department of environmental science @ stirling university, scotland
<http://www.es.stir.ac.uk/people/haynes.htm>
2. [dr. ian grieve](#) **Perthshire, United Kingdom**
the department of environmental science @ stirling university, scotland
<http://www.stir.ac.uk/envsci/grieve.htm>
3. [dr. david gilvear](#) **Scotland, United Kingdom**
the department of environmental science @ stirling university, scotland
<http://www.es.stir.ac.uk/people/gilvear.htm>

Figure 10 Search and results screen for people search tool

Although the tool demonstrates that it is possible to perform people searching, at the moment it should be considered a proof of concept. A number of improvements are required before it could be incorporated into Go-Geo! (or even offered as a standalone service). The improvements required are documented in Appendix 1.

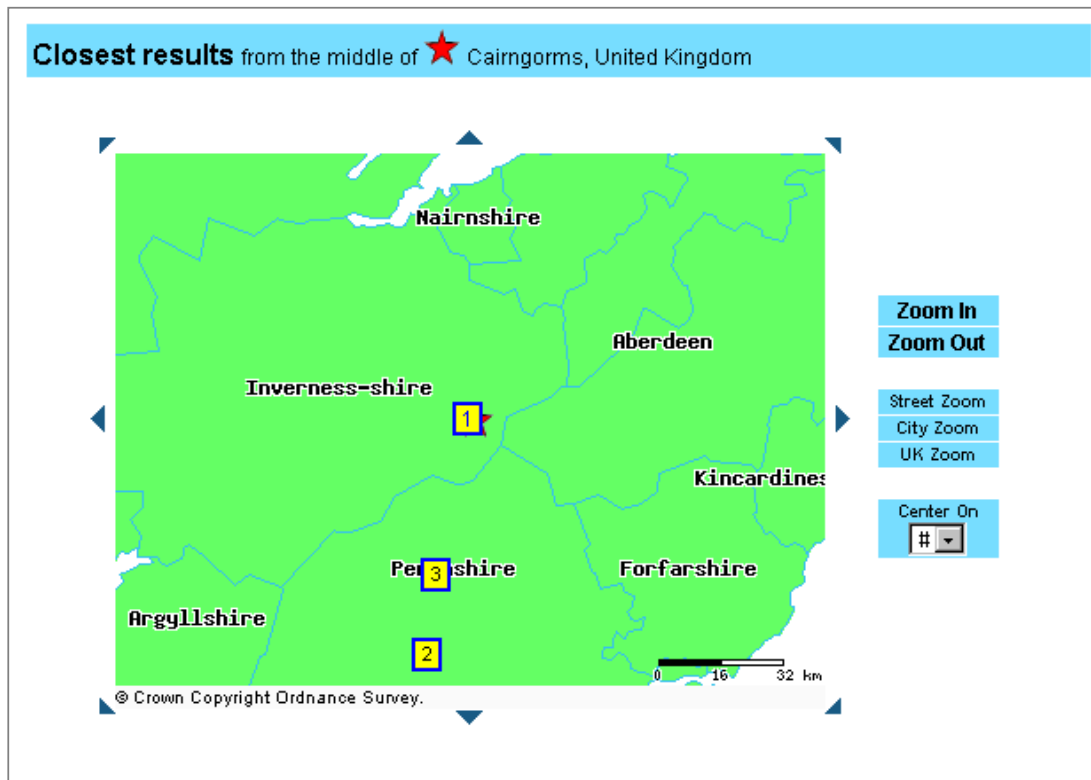


Figure 11 The map view page in the people search tool.
(Numbers refer to locations mentioned in a staff page.)

Metadata creation tool

As early as phase 1 of the project, the development of a metadata creation tool was identified as vital resource for users to use for creating, updating and supplying metadata records for inclusion in the Go-Geo! portal. (Records would be stored in a catalogue hosted by EDINA local to the portal.) An attempt to address this need occurred during the summer of 2003, when negotiations took place with AGI Glgateway team for access to the metadata creation tool they were commissioning. The idea was to install and modify the Glgateway metadata collection browser tool for use by HE/FE. While this was agreed with the Glgateway team, they failed to make this a condition when they contracted the work to build a metadata collection browser tool to a third party software house. In September 2003, this became more apparent when the problems of IPR and sublicensing had not been resolved by the lawyers of the various parties involved.

The decision was therefore taken to develop our own on-line tool. Work begun on this in January 2004 and took about 3 months. Figure 9 shows a sample page. This work was funded outwith of the JISC funding. A beta version of the tool was made available as part of the training courses being run by "Geospatial Metadata promotion and Creation Initiative" project and received very positive feedback. Currently, this tool is not available to the HE/FE community. Its release is awaiting a decision by the UK GI community on the creation of a subset of ISO 19115 at the discovery level which will be the minimum UK standard. This is discussed further below.

Figure 12 Sample screen from Metadata Creator tool

From discovery to accessing data...

There were two parts to the data distribution study: a requirements survey and a review of peer-2-peer, self-archiving and traditional (centralised) archiving. (The report entitled "Go-Geo - Data Distribution Study" which accompanies this report, provides more details on the findings of the data distribution study.)

Fewer responses to the requirements survey were gained than was initially hoped (ten responses). This was thought to be due to the time of year (exam pressure, marking etc) as well as survey fatigue amongst respondents, particularly as many of the contacts had also been invited to evaluate the Go-Geo! portal earlier in the year. From the responses received it was found that the majority of the respondents would like to see the provision of:

- a centralised archive or repository;
- online access to data;
- copyright agreements and data integrity provided for;
- a service where user and depositor support is provided;
- preservation and archiving facilities provided by the archive;
- a system which is both easy and fast to use;
- supporting material made available;
- a service which is free to deposit and preferably free to use.

It should be noted that the above list reflected feedback received at the metadata workshops and conferences. The workshops alone drew 139 attendees, thus significantly augmenting the responses collected from the requirements survey.

The most important factors in choosing a data distribution system were found to be ease of use, cost, preservation and the provision of user and depositor support.

An investigation was also carried out which looked at self-archiving, traditional (centralised) archiving and peer-2-peer. The literature review concentrated on issues such as preservation, ease of use, cost, security, support and access control. Examples of software and case studies were included to enrich the findings.

The review found that there were two main models under which a repository could operate - a centralised approach or some form of distributed model.

The main characteristics of the centralised archive are:

- storage and distribution of data from a single location;
- centralised access control over the supply and re-use of data;
- checking, cleaning and processing of data according to standard criteria;
- centralised support service, describing the contents of the data, the principles and practices governing the collection of data and other relevant properties of data;
- cataloguing technical and substantive properties of data for information and retrieval and offering user support following the supply of data.

The main characteristics of the typical distributed model include:

- data holdings distributed over various sites;
- data disseminated to users from each of the different sites, according to where the data is held;
- the various suppliers of data ideally networked in such a way that common standards and administrative procedure can be maintained, including agreements on the supply and use of data;
- a single point of entry into the network for users, together with some form of integrated cataloguing and ordering service.

There are advantages to both the centralised and distributed approaches.

There are advantages to be had from a single institution liaising with all depositors and users. Some economies of scale are also likely with a centralised service. Centralised, possibly national, services may also be better able to promote national standards for the documentation of research data, new developments and innovation including cross-national linkage, to promote research and teaching which values the use of secondary data and to police usage of the data supplied.

These advantages might also, however, be reproduced to varying degrees through a hub and spoke model with the hub generating much of the benefit provided by a centralised model. Technological developments make this increasingly possible. The distributed model can allow for the development and support of pools of knowledge and expertise and for higher volume usage in clusters related to particular datasets.

Under a hub and spoke model, a centralised facility might be responsible for the most heavily used datasets. Other datasets could then be held and provided by a network of distributed centres. This would allow for specialisation within particular satellite centres. Terms and conditions relating to deposit, access and standards could then be co-ordinated centrally.

	Self Archiving	Peer-2-Peer	Centralised Archiving
Advantages:	<ul style="list-style-type: none"> ▪ rapid, wide and free dissemination of data across the web ▪ documents are stored electronically on a publicly accessible web site ▪ documents can be deposited in either centralised archive or distributed system 	<ul style="list-style-type: none"> ▪ all parties are equal ▪ data stored locally therefore no central repository/server needed ▪ less formal, adding a document is seen as less like publishing than self archiving or centralised archiving, data creators are more willing to share data with their own communities 	<ul style="list-style-type: none"> ▪ both documents and metadata held and managed by archive therefore preservation and copyright not an issue as negotiated and managed centrally ▪ quality control of metadata and data ▪ user support available

	<ul style="list-style-type: none"> ▪ easy access to software ▪ can be published by author at any stage ▪ subject or organisation based therefore specific to given field ▪ cheaper than centralised archiving as quality checking done by peer review ▪ extendible ▪ most self archives use Dublin Core or OAI and are therefore interoperable ▪ some self archives are duplicates of printed material, therefore preservation not an issue 	<ul style="list-style-type: none"> ▪ simpler than client/server architectures ▪ more robust than centralised servers as it eliminates reliance on centralised servers that are potential critical points of failure ▪ can search both the P2P network and centralised repositories ▪ control over a closed network where individuals can set passwords ▪ scalable 	<ul style="list-style-type: none"> ▪ resource discovery supported and provided centrally ▪ reputation of organisation promotes trust and confidence in data creators ▪ required by grant award providers ▪ IPR remains with data creator ▪ existence published by archive ▪ administrative duties carried out by archive ▪ use of data can be monitored ▪ access can be managed centrally e.g. using Athens ▪ can support a large number of data users ▪ can create own search facilities
<p>Disadvantages:</p>	<ul style="list-style-type: none"> ▪ wide dispersal can lead to problems finding the data ▪ copyright - author retains pre-print but not post-print copyright. The author can not self-archive if paid royalties by publisher or if publisher holds exclusive copyright. It can be difficult to enforce copyright ▪ ownership can be disputed if data creator moves organisations ▪ rights over metadata are not clear ▪ preservation concerns over interoperability of online medium and maintenance of the archive ▪ lack of access control ▪ no quality control of metadata ▪ no or little reputation means lack of confidence in method ▪ concern over long term prospects of archive ▪ lack of licensing ▪ cost of data creators time and resources ▪ relies on peer review for quality control ▪ user support reliant on goodwill of publisher or data creator ▪ requires cultural change to ensure researchers submit content to repositories 	<ul style="list-style-type: none"> ▪ no overall control over copyright and licences ▪ ownership can be disputed if data creator moves organisations ▪ no quality control ▪ technical infrastructure ▪ does not offer good performance under heavy loads ▪ use of hard drive as nodes could lead to local slow performance ▪ must have sufficient number of nodes to work successfully ▪ unreliability of peers which may be switched off ▪ no central support system ▪ difficult to enforce rules, such as copyright ▪ network favours client not server usage, therefore creating bandwidth problems ▪ relies on metadata for data retrieval ▪ difficult to manage as no central control ▪ user support reliant on goodwill of publisher or data creator ▪ cost of data creators time and resources ▪ concern over long term prospects of archive 	<ul style="list-style-type: none"> ▪ both documents and metadata held by archive, need specialised staff skills, therefore human and financial restraints ▪ dependency on single network may lead to bandwidth issues especially with large volumes of data, although archive could just hold metadata

Software:	<ul style="list-style-type: none"> ▪ EPrints ▪ Dspace ▪ Kepler ▪ CERN 	<ul style="list-style-type: none"> ▪ JXTA ▪ Gnutella ▪ FreeNet ▪ LimeWire ▪ LionShare 	<ul style="list-style-type: none"> ▪ Archive specific
------------------	---	--	--

Table 1. Comparison of the Three Data Distribution Systems Reviewed

Conclusions of the data distribution study

Whichever method is adopted, there are several issues which need to be addressed and these are discussed below.

User and Depositor Support

Centralised archives would provide the best user and depositor support as resources are focused in one location. These types of archive are often specialised in an academic discipline and can therefore provide the most efficient service. There are examples of traditional archives that work together as one organisation but still provide specialist support at the individual level. One example of this is ESDS (Economic and Social Data Service), where specialist skills remain at the UK Data Archive, ISER, MIMAS and the Cathie Marsh Centre for Census and Survey Research.

Self-archiving and P2P networks do not provide user and depositor support. Instead they rely on in-house help at the data holders organisation, such as a library support system based at a university.

Metadata Standards

The use of the Go-Geo! HE/FE metadata application profile should be considered as a template for the distribution or storage of metadata records for Go-Geo! If a simple standard is used, such as Dublin Core, which only has a few elements, then not all of the mandatory elements will be displayed on the Go-Geo! portal and information about the datasets will be lost. At present self-archiving and centralised archiving mainly rely on Dublin Core as it promotes interoperability and preservation. However, the use of Dublin Core is too simplistic for geospatial metadata records. In contrast, P2P can be more flexible, although difficulties lay in the lack of control and enforcement of standards and protocols.

Ease of Use

All of the data distribution systems reviewed have their own advantages and disadvantages when it comes to ease of use. The centralised archive has the necessary support to assist data creators in the submission of metadata. Self-archiving and P2P archiving are more flexible as the data creator can deposit when they wish to and update the datasets when they like. This point is particularly relevant to P2P which, in theory, should be available all of the time.

Quality Control

Distributed networks, such as self-archives and P2P networks, do not have any central control over data. This means that there is a lack of quality control protocols in place.

Copyright and Licensing

For distributed archives, copyright and licensing must lie with the data creators host institution as this is where the data are stored. At a more traditional archive, licences are drawn up and signed by depositors but also users, who must sign an agreement as to the intended use of the data. This point is very important when the usage of data is restricted to one community, such as academia.

Licences are also a means of mediating IP where the data producer has used material for which that do not have copyright. This issue probably affects the majority of geospatial datasets in the UK as materials such as Ordnance Survey data are incorporated into the geospatial data. Without centralised control over licences, issues over copyright will occur.

Cost

The cost of self-archiving and P2P lies mainly with the data creator, as it is they who will spend the time setting up the system, in the case of P2P and submitting data, as in the case of self-archiving. The cost with centralised archiving lies with the archive, who must pay for overheads, staff expertise and time, resources etc. The cost to the user in most cases will be nothing unless the data is used commercially.

Security

A distributed system provides greater security against denial of service attacks and network crashes as the system does not rely on one data source or network. A centralised archive could provide greater data security and integrity as access is controlled and usage of the data is monitored.

Preservation

The traditional way to preserve material would be to deposit data in a traditional archive. Although this is still the most effective, reassuring way of preserving data, there are other issues to consider. If the data creator is based at a university they may be able to rely on the university library for data preservation, depending on in-house policies. Another option would be to use another service such as demonstrated in the Thesis Alive project. The preservation of the metadata records themselves is an issue which needs further consideration.

Possible solutions

One solution could be to combine the advantages of using centralised archiving with a distributed system. The creation of a distributed archive, could be subject based and would involve using several archives such as the ADS and UKDA to hold the datasets and metadata which they would normally hold, such as archaeology and social science respectively. In this way, the advantages of using a well established, reputable archive which offers quality control and user support can be combined with the advantage of not holding data in one place, therefore not depending on one single network. This ensures that whilst only good quality metadata is published the system will not be slowed down. It also means that data creators can store their data in the place which they would normally do so and therefore there is less likely to be duplication of data. Creating subject repositories of this kind would also allow for the creation of subject-based user support systems, where the repository would be able to provide specialist user support. One major problem with this solution would be that the various archives already use their own metadata standards and are unlikely to want to hold two different standards or convert their standard into a geospatial one. Some confusion may also arise as data creators may not always know which subject repository their dataset is most suited to. Finally, since certain subject archives already exist, why don't more researchers see them as the obvious mechanism by which to share their data?

An alternative solution would be to set up a completely distributed system where repositories are established at the individual organisations who have created the data. (Institutions are already establishing repositories for e-prints and learning objects; these could be extended to include geospatial datasets.) This type of system would rely on the institutions providing quality control and user support, which could be achieved through current archive or library procedures. However, it may also lead to duplication of data and inconsistencies in style. For example, general metadata standards may be used which would probably not be compliant with the Go-Geo! application profile. This could be mitigated against if the JISC and the research councils mandated that the Go-Geo! application profile must be used for documenting geospatial data in institutional repositories.

Yet another solution would be to hold the datasets and associated metadata at an established archive, such as UKDA. This is likely to be a welcomed scenario as provisions will be made for user and depositor support, preservation, licensing and data security. Again, the issue of differences between metadata standards create difficulties with this option. More investigation is needed into this issue though as further changes may be made to standards, such as the DDI, in the future which could make it compliant enough with the Go-Geo! application profile to warrant this as a feasible option.

Issues arising

As in any project, a number of issues arose during the project. Some of these have been discussed above. In this section we discuss those issues not dealt with so far.

Cross-walking between metadata standards

When the original NGDF service was established (what is now GIGateway) there was no already established catalogues. Therefore a standard could be imposed for the metadata that was created. This was not the situation when Go-Geo! was being established. A number of catalogues already existed containing records describing geospatial data but documented using the metadata standard already adopted by that community.

Ideally when these catalogues are searched by Go-Geo! the retrieved metadata should be presented in a standard way, namely using the FE/HE Application Profile of the NGDF Metadata Guidelines. Unfortunately, because the catalogues use a variety of different standards this is difficult.

For example, the project team investigated adding the Global Change Master Directory (GCMD) to the set of catalogues searched by Go-Geo!. This catalogue would give users access to global datasets whose geographic coverage includes the UK. The records retrieved from the GCMD Z39.50 server use the DIF metadata standard in XML/HTML format. The GCMD Z server uses MD8-Isite, which is their own modified version of Isite. They map their DIF records to the GEO Z profile to allow them to be searched. They can claim interoperability because their catalogue can be searched using the GEO profile. However, if one retrieves the full record, it is in the DIF metadata format and not the US FDGC record format, which, recall, is similar to FE/HE Application Profile of the NGDF Metadata Guidelines used by Go-Geo!.

The same is true of the Archaeological Data Service (ADS) collection level metadata and the UKDA catalogue service. The former uses Dublin Core while the latter uses DDI. MIMAS use a version of Dublin Core but have grouped their geospatial data into high level collections e.g. all the datasets in LandMap are documented in a single record.

This means, as far as Go-Geo! is concerned, one cannot just simply plug the GCMD catalogue or the ADS catalogue or the UKDA catalogue into the existing search routines. In the phase 2 project, to cross-search the History Data Service catalogue records, a converter had to be written to map the DDI fields into FGDC and then to the FE/HE Application Profile. This is being used for the UKDA DDI records. During this phase the project has had to write a specific mapping from Dublin Core to the FE/HE Application Profile of the NGDF Metadata Guidelines. This involved extra work, complexity and generally the approach is not really scalable. It also results in incomplete records since there is not a one to one or many to one mapping between each standard.

A key issue surrounds the need for DC and DDI to adopt elements that define a dataset's place/location based on co-ordinates. A place name does not always constitute a unique place because the same place name can be used for multiple locations (e.g. Norfolk is found in England and the US). This in turn can extend place name keyword searches to include all datasets associated with every occurrence of the place name. Including co-ordinate elements in DC and DDI will provide an alternative that will allow for co-ordinate values to define places and this in turn will address a key interoperability issue.

Recently, the DDI committee adopted two geospatial elements; this process also included consultation with several people on the project team. These allow for the metadata creator to enter a) co-ordinate values of a bounding rectangle, or b) co-ordinate values of a bounding polygon. This is an important step towards improving cross-search capabilities between the social science and geospatial communities. Dublin Core has also addressed this and recommends the use of co-ordinates as part of its 'Coverage' element under version 1.1. However, it seems unlikely that catalogue owners will retrospectively add the geospatial elements into their records.

Which geospatial metadata standard?

The adherence to standards is an integral part of the JISC Information Environment. The development of standards is an ongoing process, but it is recognised that standards bodies exist to place markers along this process to facilitate the take-up of standards and encourage their use. The development of the ISO 19915 standard for geospatial metadata is one such standard.

During phase 2 it was recognised that ISO 19115 was going to become the geospatial metadata standard for the international community. Based on this assessment, the mandatory ISO elements

were added to the HE/FE Metadata Application Profile to assure compliance. However, there are eight HE/FE profile elements that do not have equivalent ISO elements. Seven of these elements were derived from the NGDF Discovery Metadata Guidelines. The NGDF guidelines represent an application profile that was created to serve the UK geospatial community. Though the NGDF guidelines are based on the FGDC's standard, a standard widely adopted previously across the international community, a number of non-FGDC elements were added to NGDF to address specific UK needs. The ISO 19915 standard, in many ways, mirrors the FDGC's standard; therefore accounting for the missing counterpart elements.

This issue is being resolved by UK initiative known as the UK GEMINI Initiative. Started in 2003 and including several people from the project team, its goal is to make the NGDF standard compliant with the ISO 19115 metadata standard. This UK profile of the ISO 19115 metadata standard will be adopted by government¹¹, public and commercial bodies to describe geospatial data sets and resources they hold. In time it is planned for it to become a BSI national standard. Based on initial feedback, it appears that the problem with element representation will be addressed with the adoption of the ISO elements.

With the resolution of the mismatch between the UK standard and ISO 19115, the HE/FE Metadata Application Profile can be modified. Existing records which use the HE/FE Metadata Application Profile can be mapped to a new HE/FE Metadata Application Profile. Converting the existing metadata records to the ISO standard will allow the portal to be searched by services outside of the UK and thereby increase its use.

A member of the project team has produced a new version of the HE/FE profile based upon ISO19115 with some extensions to cater for fields required by the HE/FE community (these are in the current version of the profile). A critical question is when the records should be converted. Ideally it should be done before too many records are created using the old standard. However, there is a problem in that XML schema for ISO 19115 (ISO 19319) does not yet exist in a final form. Furthermore the draft version doesn't yet support extensions; although the next version will.

An added difficulty is that the database application used to host the metadata, Isite, does not yet support ISO 19139. Isite uses something called a 'doctype' to specify the structure of records that it is indexing. The 'ISO19115cd2' doctype in the Isite version being used by Go-Geo! to store metadata records (v1.0) is substantially different from the ISO 19139 XML schema that the Go-Geo! metadata editor is using to build records. This means that records created with the editor are not fully compatible with the current version of Isite. The same is true if the records already in Isite are converted. A final version of ISO 19139 is due to be out near Christmas 2004. Hopefully, the developers of Isite will then quickly update their application.

There are two further complications. First, it may be some time before Glgateway converts its records to ISO 19115. This means that for the short term the record display pages of Go-Geo! will need to be modified to support the display of records as either NGDF metadata standard records (from Glgateway) or HE/FE profile of ISO 19115 (from the academic geo-network). Second, new mappings from DDI and DC to the new version of the HE/FE profile will also need to be produced. However, we are concerned that an exact mapping between the standards currently used in some third party catalogues and the HE/FE profile of ISO 19115 may not be possible, particularly as far as compliance with mandatory elements of ISO19115 is concerned. This means that records from these catalogues could not be exposed through Glgateway or other discovery service that mandate the use of ISO19115.

Widening the dissemination of the metadata

Exposure of the geospatial metadata beyond the academic community may present an access-related problem for services. Some data licences restrict access only to academics and even where this may not be the case, services tend not to have the resources needed to supply or support non-academic users.

¹¹ It will map to the UK Government's e-gif GMS standard.

Implications and future plans

The next 12 months

In April 2004, the JISC Portal Programme Manager informed the project team that there was no commitment from the JISC Services Team to fund portals as services as yet. This was because they were awaiting the outcomes from a business and technical sustainability study to advise them on courses of action. As a consequence JISC stated it would not therefore be possible to move Go-Geo! from its current phase into service at the end of the project. In the meantime, the project team were informed that the JISC Development Group were to initiate a review of Go-Geo! and the sister project geoXwalk to assess what options there were for the future.

Given the importance of the Portal, and the unfairness imposed to both existing users and potential new users if Portal operations were to cease, the project team informed the JISC on 26 May 2004 that EDINA had agreed to undertake to run Go-Geo! as a service for one year starting 1 August 2004. At the end of that period EDINA will review the effort required on its part against the level of usage by the community. At that time, if further funding is not forthcoming from the JISC, EDINA and the UKDA will wish to open discussions with the JISC about the long term sustainability of the service. It was noted that it was possible that EDINA might decide to continue to run the Go-Geo! after this date, seeking external funding to do so.

As a result of this proposal, JISC agreed that EDINA could run Go-Geo! as a service for one year commencing from August 2004. It was felt by the JISC that this would be a valuable opportunity to further investigate the worth of Go-Geo! to the wider community, both within academia and beyond. Unfortunately, since JISC was not contributing to this arrangement directly, JISC stated that it would not be possible to promote the service as a JISC service, however any ongoing development funding should be badged as JISC-funded.

The fact that the service will not be badged as a JISC service is unfortunate particularly as EDINA is taking on the operation of the service for a year to avoid bad PR with the academic community while the JISC Services Team decides whether or not it will fund portals as services. It means, for example, that the service will not benefit from being part of the general promotion of JISC services undertaken by the JISC. It also means that EDINA will probably have to cease using Athens as the authentication and authorisation mechanism. (EduServ require payment for use of the Athens service if the service is not a JISC one but since there is no income from Go-Geo! this would mean that EDINA would have to find the money from somewhere else. This EDINA is unwilling to do.)

Further development opportunities

In return for taking on the service aspect of Go-Geo!, EDINA and the UKDA wish that the JISC IE programme help by funding work that targets promoting and supporting metadata creation and Go-Geo! Portal development. (This does not include development of links within channels or new channels.) At the time of writing, discussions are to take place with the JISC to prioritise areas of development and to agree a development path for the coming year.

In parallel with the ongoing development of Go-Geo! alternative uses for Go-Geo! should be explored, especially with respect to the e-research and JCSR interests in portals. Greater links with the research councils should also be investigated to explore how Go-Geo! might play a role for them.

Metadata Creation

A key area, and one that remains a great concern, is metadata creation. The academic community has expressed a greater interest in metadata than was anticipated at the outset of the Metadata Promotion and Creation Initiative project, showing that the mindset is changing.

It is clear that it requires real resource to encourage and co-ordinate those who are producing or have already produced geospatial data to produce metadata. It also needs to be recognised that there is a difference between individuals and small research teams and service providers. The metadata issues for each are different. For example, feedback from academics indicates that they see a lack of mid and long-term commitment to continued support of work in this area. Without these commitments, the

Go-Geo! portal's future is at risk. Furthermore, if these concerns are not addressed, then the Go-Geo! portal is likely to remain static and serve little purpose.

The focus now needs to be on the creation of metadata records to populate the catalogue hosted by EDINA and searched by Go-Geo!. The project has created an on-line tool for doing this but data producers need to be encouraged to use it, the records supplied need to be quality assured and control terms applied and licensing issues need to be dealt with. Metadata providers also need to be encouraged to keep their records up to date. The Geospatial Metadata Initiative Project found that there was interest in the use of the metadata creation tool and Go-Geo! to assist in the internal management of geospatial data held by departments and research teams. A first step would be to provide an internal data management service for each university offering a complete resource package that includes the HFE profile, guidelines, metadata creation tool and portal. (Those creating records would mark whether a record is private to an institution or public i.e. anyone can see it. Members of a particular institution could log in to see their own institution's metadata records.) If such a service was offered, it may be enough of a benefit to encourage metadata records to be created and 'published' for others in other institutions to view.

As these measures are put into practice, the range of concerns that many individuals conveyed at metadata workshops, conferences and survey responses can be addressed. This can set the stage for HE and FE institutions involved in the internal data management service to open and publicise their metadata records to the entire HE and FE community and fulfil the scope of the Go-Geo! Portal as envisaged at the start of this project

Continued support for metadata creation and quality assurances would demonstrate a commitment to supplying content for the portal and reassure users that the metadata records are both accurate and complete. We are therefore requesting that the JISC IE programme consider funding an individual to undertake the task of metadata co-ordination. We would also be seeking T&S for the individual. The individual would be based at EDINA.

Development of Go-Geo!

There are a considerable number of enhancements that can be introduced to the Go-Geo! portal to provide users with extended functionality and performance. Some must be addressed over the short term with further considerations given to those developments that should or could be done with additional funding. The priorities stand as follows:

Must

- Expose the Go-Geo! gateway, and thus the catalogues comprising the geo-data network, as a virtual catalogue within the Ggateway network.
- Investigate, design and implement a data sharing/dissemination mechanism for use by researchers and research groups. As we have stated above, this is very important as difficulties in the mechanics for sharing of data are seen as a barrier to the publication of its existence and therefore its (re)use.
- Modify the portal so that private areas can be created within the portal that can be used for 'local' institutional management of metadata by research teams/departments.
- Migrate the existing metadata records and service from the UK standard created for the Go-Geo! project to the ISO standards (19915 and 19139).

Should

- Develop facilities to support data mining and data visualisation through data that are accessible through OpenGIS Web Map Services, OpenGIS Web Feature Services or OpenGIS Web Coverage Services.
- Work towards facilities to support direct data linkage (access).
- Develop the user profiling and customisation part of the service.

Could

- Create a purely map driven search interface where data can be found based solely on a user specified geographic extent.
- Develop the portlet aspects of the service further and encourage uptake of the Portlet by other service providers whose users have an interest in geospatial data.

- Develop the people search facility to service strength and quality for inclusion in Go-Geo!
- Extend the geographic coverage of the service content to beyond Great Britain. This has to be done in a cost-effective way through linking the service up to other such services being developed around the world.
- Work to identify and include other JISC resources.
- Implement the OpenGIS Catalog Interface Implementation Specification (v.2) on top of the Go-Geo! Gateway thus making it OpenGIS/ISO compliant. This would benefit the academic community by enabling linking up with developments across Europe.
- Investigate extending the portal to a GRID portal for use by geographic information science.

A number of more minor developments have been identified. These include:

- * a facility that, for each result item, provides an explanation of how it was found. This is particularly needed because of the innovative use of geoXwalk to expand the spatial part of a query;
- * introduce a way for users to report errors in the metadata;
- * investigate the use of a terminology service so that searches of resources and catalogues which use different controlled vocabularies and subject classifications are possible;
- * continue to improve the spatial searching capabilities of Go-Geo!;
- * add support for postcode searching;
- * looking again at the spatial searching;
- * inclusion of an ambiguous place name clarification page;
- * an alert mechanism for notification of new items to the portal;
- * provide a mechanism by which metadata records are accessible to the user by “two clicks to content” organised by ISO 19115 Topic Categories, and
- * a web-based discussion tool.

The JISC have suggested development of a set of requirements for non-geospatial datasets to allow their metadata to become discoverable through the Go-Geo! portal. This acknowledges the fact that different datasets need individual attention to make them fully discoverable through the portal. It is important that the data holders have as much information as they can to allow the possible “preparation of datasets for surfacing through Go-Geo!”¹². The creation of a set of requirements would allow this to take place and form the basis for future discussions between Go-Geo! and data holders, whether individuals or service providers (each will have to undertake different activities and require different levels of resource and may be changes to workflows and processes), on the inclusion of additional resources.

Data distribution and sharing – the need for a strategy

The lack of a data distribution mechanism is a significant barrier to the sharing of data. However, like metadata creation, work in this area needs to be part of a broader strategy for the management, preservation and storage, and dissemination of geospatial datasets within the UK tertiary education. This strategy needs to be developed by key stakeholders including the JISC, the research councils, national data centres and archives, data creators and data users.

JISC need to consider the creation of a national repository or repositories for geospatial data. This is very important and is a major project in its own right. It is also a long-term undertaking. However, if we wait for its completion before encouraging the creation of metadata, there is the risk that the geospatial metadata content of Go-Geo! will be restricted to service nodes and we will not have really gone anyway to expose the data we believe is out there in institutions. There needs to be a short, medium and longer term strategy therefore which can be promoted to users and which shows that the key issues for them are recognised and that together we plan to work towards cost effective solutions.

Other considerations: Geo-data Portals – the wider view

On the 23 July 2004 the Commission of the European Communities published a proposal for a directive establishing an infrastructure for spatial information in the community (COM(2004) 516 final). The proposal sets out to make interoperable geospatial information readily available in support of both national and community policy and to enable public access to this information. The proposed Directive creates a legal framework for the establishment and operation of an Infrastructure for Spatial

¹² This may necessitate that that JISC mandate that geospatial searching is enabled where relevant.

Information in Europe. The main beneficiaries are public authorities, legislators and citizens. However other groups such as universities and researchers are expected to benefit. Public bodies in member states are expected to comply with the Directive. This includes Universities.

The Directive contains some 34 articles. Many are relevant to the present discussion but two specific ones stand out.

“Member states shall ensure that metadata are created for spatial datasets and services, and that those metadata are kept up to date.” Article 8(1)

“Member states shall establish and operate a network of the following services for spatial datasets and service for which metadata have been created....

- a) discovery services
- b) view services
- c) download services
- d) transformation services
- e) “invoke spatial data services” services, enabling data services to be invoked.” Article 18(1)

The Directive notes that services for discovering and viewing spatial datasets should be **free of charge**. It also states that Member states must provide access to the services referred to in Article 18(1) through their own geo-portals.

What is being proposed under this directive is the establishment of a European wide Spatial Data Infrastructure. Converging technological trends and increasing maturity in society’s use of geographic information has seen proposals for the establishment of Spatial Data Infrastructures (SDIs) at various levels, e.g. European, National and Regional. A Spatial Data Infrastructure is defined as

“the relevant base collection of technologies, policies and institutional arrangements that facilitate the availability of and access to spatial data”,
(Nebert 2001)

The benefits of establishing an SDI within HE and FE include

- Reduced duplication of geospatial data collection
- Improved awareness of what data exist
- Improved service delivery, in terms of efficiency, content and type of services
- Different information sources can be made available to the user in a consistent and 'joined-up' (i.e. seamless) way
- Improved interoperability of systems working to common standards
- Better integration of GI and GIS into research and teaching
- Alignment with other complementary initiatives to avoid duplication of effort and expenditure

The sorts of services found in an SDI include discovery services, viewing and visualisation services, services for downloading data and services for transforming data. There are also infrastructure services such as gazetteer, geoparsing, geocoding and registry services. These services are being built using published open interoperability specifications primarily those being developed by the OpenGIS consortium although standards from other areas are used where they already exist. Geo-data portals are considered as critical components in any SDI.

JISC is promoting a common and shared vision - the ‘Information Environment’ or JISC IE and geospatial data are an important part of this vision (e.g. one can apply the discover, locate, access, use, and publish functional model to geospatial data and geographic information). Many of the components of an SDI already ‘exist’ in the IE (for example, organisational remits, policies, data, technologies, standards, delivery mechanisms (access)) and, in fact, within a UK context tertiary education is well on the way towards establishing its own SDI (Figure 13). This is something that is increasingly recognised by organisations outside of UK tertiary education even if it isn’t recognised as so by those within the community. A more detailed description of this can be found in the recent paper “Spatial Data Infrastructures and Digital Libraries: Paths to Convergence” (Reid et al, 2004). The key point is that the GoGeo! portal is a core component of a Spatial Data Infrastructure (SDI) for the UK

tertiary academic community but at the same time has an important role to play both in a UK context (though exposure of its metadata by Glgateway) and within Europe. It is one of a number of initiatives but has a number of unique characteristics not least the focus on the academic community, and mechanisms for searching for geo-related resources as well as metadata about geospatial data, and thus its close relationship with the digital library world.

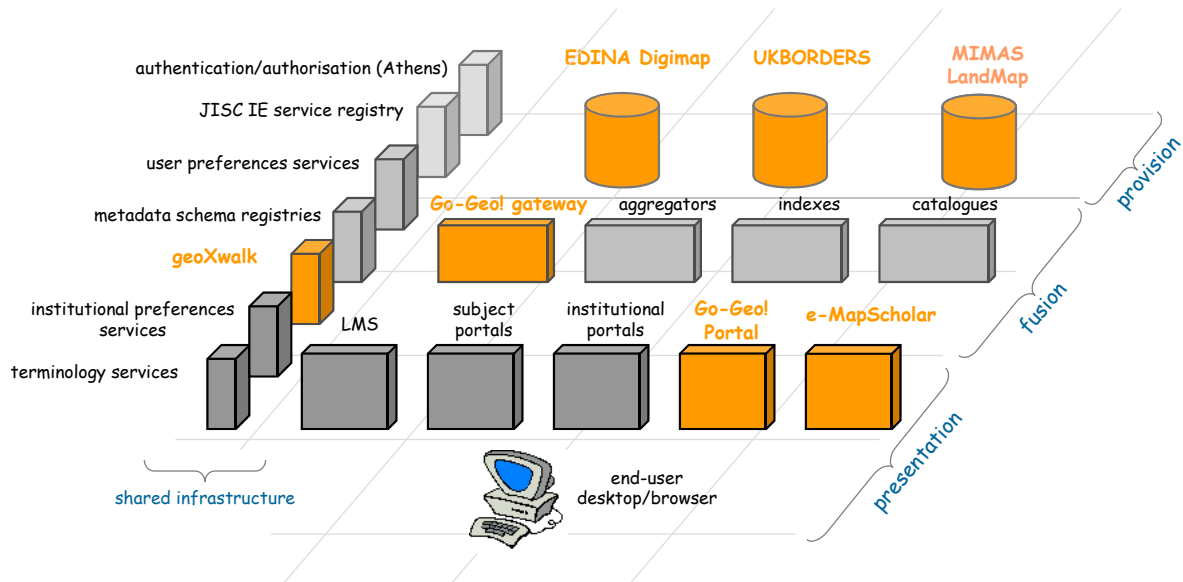


Figure 13 An SDI in the JISC IE

Exit strategy

The goal of our exit strategy is, as has been stated on previous occasions, for the portal to become a service within the JISC Information Environment.

As stated above, we see the portal as forming a core component of a community-wide Spatial Data Infrastructure (SDI) for the UK tertiary academic community. In this regard it could also form the building block for the data integration and presentation portal envisaged by the JISC Geospatial Working Group.

“A portal accessing many services, data types etc via one login and user interface, which might be a simple portal, i.e. a common finding aid but different underlying services, or a more complex portal (in time) providing full data fusion in the JISC IE sense, backed by data interoperability using OGC standards, post-code etc as glue, supported wherever possible by common licensing terms, e.g. minimising special registration requirements (which tend to frustrate the above).”

JISC Geospatial Working Group 2001.

The long-term goal of the portal would be to provide researchers, students and teaching staff with access to tools to discovery geospatial data, and then browse, visualise, extract, integrate and analyse geospatial data. To bring this about requires greater geospatial content within the JISC IE (we note that licence for access to additional geospatial datasets is currently being negotiated by the JISC) and for these data to be accessible through the adoption of open standards and protocols, in particular the OpenGIS Consortiums (OGC) interoperability framework.

It is also becoming clear that the portal could take on other roles. First, as a centrally managed repository, which institutions use to internally manage metadata about geospatial data they hold. Second, our initial discussions regarding providing access to datasets and data sharing indicate a

desire by some individuals for a place where datasets can be deposited (archived and preserved), with appropriate metadata, and then accessed by others. This would eliminate the time and effort required by data creators to deal with orders for data or to dedicate and establish internal resources to these requests. The portal could thus become the front door to a data distribution facility.

In an earlier confidential document to the JISC we described alternative business models through which Go-Geo! might become either fully or partially self funding as opposed to grant funding by JISC. These were

- a subscription service
- a pay as you go service – the Utility model
- through web advertising

We also set out the positive and negative points of each approach and noted that none of them were entirely satisfactory. It is worth noting that other GI portals that exist in Europe, North America and Australia provide the discovery capability for free and, although access to data and mapping services through the portal is restricted, this is normally for copyright or licence reasons rather than because a charge has been placed on the user.

The question of how Go-Geo! should be funded should it become a service is bound up with decisions which need to be taken by the JISC with regard to its overall service strategy, in particular funding of its portals, and with the strategic direction of the JISC Portal Strategy. We await the outcome of the JISC's Portals sustainability study. Not only will the recommendations of that project be significant for the future of the 'service' but also for the higher and further education community's ongoing involvement in national and international activities in the area of Spatial Data Infrastructures.

Recommendations

1. The existing trial service should become a full service within the JISC IE. (This recommendation has the full support of the JISC Geospatial working Group.)
2. The enormous potential of the portal should be recognised. The evaluation work has identified two important needs for the management of geospatially referenced data:
 - a system to assist in institutional geospatial data management;
 - an enhanced national service for geospatial resources with an emphasis on access to data.The portal has the potential to serve both these needs and we seek further funding to meet these goals.
3. Metadata is critical to the success of the service and, in turn, to publishing the existence of data to encourage re-use. There is no ready source. We are therefore requesting that the JISC IE programme consider funding an individual to undertake the task of metadata co-ordination and QA.
4. The Go-Geo! Geo-data network should be linked into the national Ggateway service with metadata exposed through a virtual catalogue.
5. There is a desire to be able to search and gain access to geospatial data whose geographic area of interest is outside the UK. Cost-effective ways of doing this need to be investigated.
6. A number of potential areas for developing Go-Geo! have been identified. Some of these could be separate projects in their own right. JISC should consider prioritising the developments proposed, taking account of the views of users, and provide funding.
7. With the publication of ISO 19115, Go-Geo! should move to use this standard through adopting the new HE/FE application profile produced by this project. JISC should formally state that this application profile is to be used by those documenting geospatial data in UK HE and FE institutions.

8. The project has produced further evidence that there are major barriers to data sharing that need to be lowered if not removed entirely. A further study investigating how improved data sharing can be facilitated is required, possibly within the JISC's new repositories programme.
9. A broader strategy for the management, preservation and storage, and dissemination of geospatial datasets within the UK tertiary education needs to be developed by key stakeholders including the JISC, the research councils, national data centres and archives, data creators and data users.
10. In parallel with the ongoing development of Go-Geo! alternative uses for Go-Geo! should be explored, especially with respect to the e-research and JCSR interests in portals. Greater links with the research councils should also be investigated to explore how Go-Geo! might play a role for them.

Conclusions

The evaluation work for this phase of the project, coupled with feedback from the metadata initiative, has clearly demonstrated that academic users have need of a mechanism for discovering resources in the context of their geographical coverage. The goal of the portal was to fulfil this need by creating a universally applicable system under which resources, and in particular geospatial data, collected and catalogued under other headings, (for example, archaeological and historical data), can be discovered in terms of their geographical coverage. This goal has been met. This ability to geographically cross-search catalogues and resources found within the national digital library, i.e. the JISC Information Environment, distinguishes Go-Geo! from other GI portal developments occurring in the world where the focus is purely on information about datasets.

As far back as 1999, in the original proposal for a geospatial data discovery tool, we stated that to gain maximum benefit from the geospatial data currently existing in UK tertiary education, to ensure that their teaching and research potential is fully exploited, and to minimise the redundancy of data collection, the existence of the data needed be recorded and then promoted to others. The amount of metadata describing geospatial data within academic institutions remains low.

The issue of metadata provision is critical to the success of Go-Geo! as a service. The academic community has expressed a greater interest in metadata than was anticipated at the outset of the Metadata Promotion and Creation Initiative project, showing that the mindset is changing. But it requires real resource to encourage and co-ordinate those who are producing or have already produced geospatial data to produce metadata. Furthermore, there has to be proven benefit in relation to the investment in time and effort required, support services are needed to assist data creators in the production of metadata and, as phases 1 and 2 showed and now again in this phase, concerns about the distribution of data have to be dealt with.

Access to data certainly presents more difficulties than discovery. Data residing in specialist services are easily available for academic use from those services. Indeed the technology and standards exist such that some of this data could be made accessible to the user directly through the portal. Whilst the response to the data audit was disappointing, contact with GIS experts in academic organisations suggests that many, potentially valuable, geo referenced datasets are not being made available for secondary use. The reasons for and the impact of this on the portal have been discussed in this final project report and the accompanying reports. However what has become apparent is that there is a need for broader strategy for the management, preservation and storage, and dissemination of geospatial datasets within the UK tertiary education. This is something that needs to be taken up by the relevant parties with some urgency.

References

Reports from previous phases of the Go-Geo! project are available from the project web site at <http://go-geo.data-archive.ac.uk/>

Medyckyj-Scott D, Newman I, Ruggles C and Walker D (eds), 1989, Metadata in the Geosciences, Group D Publications, Loughborough.

Medyckyj-Scott, D. et al, 2001, "Geo-data Browser – A geospatial data resource discovery tool for UK Further and Higher Education - Project Overview and Recommendations", 12/07/01, 1.1, EDINA/UK Data Archive.

Nebert, D. (ed.), 2001, Developing Spatial Data Infrastructures: The SDI Cookbook, version 1.1, 15 May 2001.

O'Hanlon C., 2001, "Geo-data Browser – Review of Metadata Standards", 12/07/01, 1.0, EDINA/UK Data Archive.

Open GIS Consortium Inc: <http://www.opengis.org/>

Pradhan, A. and Medyckyj-Scott, 2001, Geo-data Browser - Requirements Specification and Solution Strategy, 29th June 2001, 1.0, EDINA/UK Data Archive.

Reid, J., Higgins. C., Medyckyj-Scott, D. and Robson A., 2004, Spatial Data Infrastructures and Digital Libraries: Paths to Convergence, D-Lib Magazine, Volume 10 Number 5, May 2004, doi:10.1045/may2004-reid.

Appendix 1 The People Search tool

The spider is given a starting set of URLs for academic institutions. It visits these in a breath-first search, building up a database of the site URLs, possible locations associated with the site and content. This database is then analysed. A list of department names is used to identify the sites of particular departments or schools. The names in the list are ones judged to have staff working in areas with a geographic focus e.g. geography, geology, environmental science, archaeology, and so on. Each web page within a site that passes the first test is then rated on its likelihood of being a staff page or series of pages for a particular member of staff. This rating is calculated by looking at the format of the site URL, words occurring in the title and page content. It also scans the start of the page, keywords and the title to find a recognised name.

A location is then assigned to a web page in one of two ways. The first is to look for a postcode on the page itself or on the parent page (so it would have the postcode on the main university if no postcode existed on the page itself). The second method is to look for one or more places names mentioned in the page by the staff member. This is very problematic to determine, so another rating system is used. The difference between these two methods is that the first determines the institution that the individual is working at and the second attempts to find the location where that the individual is currently undertaking his or her research.

Sites with a high staff ratings are stored in the search engine database. This stores information gained from the spidering of the page (such as the plain text content, URL, address, keywords and title) as well as extra information gained from the further analysis. The database is arranged to allow quick and easy searches of words and place names, so there is an indexed table of words that occur in all the pages which is used to associate words with integer relevance values (0-255). Each word has a relevance value as words that, for example, appear in the page title and keywords are thought as more relevant to the page as a whole than words that appear at the end of the page content. Each page has latitude and longitude values which are found from the postcode/place name analysis. Extra information about the member of staff such as what the software thinks their name, email and telephone number is can also be stored. In order to record how certain we are that the place name is valid, the database stores a place name rating for each page, which is calculated by looking at the following three factors:

- How certain is it that the word is a place name?

The words which match one of a list of place names in a place names list are analysed further. Those words before and after the place name are examined to determine if the word is being used as a place name or as part of, for example, the name of an individual (e.g. James Devon). The names of individuals can be mostly eliminated by checking to see if the word is preceded or followed by a recognised name (three words before and after are checked, to take account of unrecognised middle names and initials).

- How ambiguous is the place name?

If there are many places with the same name then a low rating is given, as the likelihood of the place mentioned being any one of the locations with the same name is less than if there was only one or two places with that name.

- How certain is it that the place is where the staff member is basing their research?

This is a difficult thing to determine, as we want to eliminate places the staff member may be mentioning for other reasons, for example, because it is the name of the university they work at, or because that is where they like to spend their holidays. We can eliminate colleges and university names by scanning for 'university', 'college', etc in the surrounding words, and give higher ratings to sentences that imply they are researching the area, but this stage is not as accurate as it could be, and requires further investigation.

Figure11 (in the main document) shows the search and result listings screen for the people search tool. Clicking on a staff name results in the relevant staff page being displayed. This page is requested from the university or departmental web server hosting the staff pages. One of the reasons this is done is to ensure copyright is not infringed. It also means that a user sees the most up-to-date page. The user can also view a map showing the relevant locations mentioned by each member of staff found with respect to the users query (Figure 12 in main document).

The staff page, word position and place name ratings are used to affect the results returned by the search engine and their order. The order of all pages returned is affected by the staff page rating, so that pages which are more likely to be staff pages are returned first. If a user searches for a place name only, then pages that are from a nearby location are returned. If he searches for a word only, then pages that contain that word are returned, with those pages which mention the word early in the page being returned first. If the user searches for a word and a location then a rating is worked out based on the position of the word in the page and its assigned location.

The people search tool needs to frequently revisit all the university web pages to ensure reliable searching, probably at least once a month. The number of web pages harvested also means that a lot of space is required to store data, as the tool stores data for all the pages spidered and only identifies staff pages in post processing. Although this space is only required during the processing of the web pages, it does need to be available.

Although the tool demonstrates that it is possible to perform people searching, at the moment it should be considered a proof of concept. A number of improvements are required before it could be incorporated into Go-Geo! (or even offered as a standalone service) as well as a significant amount of further testing.

- * Refine the spidering stage to store information from staff pages on only those points in the text that are thought to be relevant. This would cut down on the space required for processing. This may prove difficult do as the spidering code is complicated.
- * Modify the tool to use the GeoParser, developed in the geoXwalk project, for the identification of place names in a page. This is a more robust and accurate tool but comes at the cost of longer processing times.
- * Currently the spatial searching is point based i.e. distance of a location from a point, which can give some strange results. The tool could be modified to use area polygons in searches.
- * Integrate with the geoXwalk gazetteer server. This would allow more sophisticated spatial searching.
- * More intelligently decide when pages need to be re-spidered. For example, by using a facility to check if webpage links are still valid and to re-spider using this information.

Appendix 2 Final Budget

Total JISC grant **£76,950** excluding metadata project **Co-funding** **None**

This reporting period **Aug 03-Jul 04**

	Forecast budget for this reporting period(from project plan)	Budget for this reporting period (including any underspend or overspend)	Spend for this reporting period	Balance for this reporting period
Staff				
EDINA staff	49,800		51,300	-1,500
Philip Abrahamson AD2 (£21125-£29621)				
Eddie Boyle AD2 (£21125-£29621)				
David Medyckyj-Scott AD5 (£35251-£43067)				
HDS staff	14,500		15,565	-1,065
Julie Missen OR1 pt 49				
Informatics, City University*	2,500		745	1,755
Joint ADS and DA for metadata**	3,000		750	2,250
HDS Consumables	200		100	100
EDINA Consumables	400		454	-54
Travel and Subsistence shared EDINA and HDS	6550		6577	-27
TOTALS	£76,950		£75,491	£1,459

* Used by UKDA after termination of collaboration with City

** Portion used by UKDA

