

Three use case scenarios for geospatial metadata automation

Tony Mathys and James Reid,
EDINA National Data Centre,
The University of Edinburgh

Scenario 1.

Title	Assisting a spatial data developer with metadata creation.
Author	Tony Mathys
Narrative	<p>Fiona is a researcher at the University of Edinburgh. She has created a spatial dataset which maps midge density patterns on the Isle of Skye in Scotland. As a responsible data developer, she recognises the importance of documenting her Midge Map dataset for purpose of practising good data management, and for publishing this information as a metadata record on a geo-portal. She requires a tool which can facilitate this task, especially as she and many other data developers find metadata creation tedious and time consuming.</p>
Tools	<p>GeoDoc Metadata Editor Tool and My Go-Geo!, part of the JISC-funded Go-Geo! Portal service run at EDINA, a UK National Data Centre at The University of Edinburgh, offer the functionality to support partial automation of geospatial metadata creation and publication on a geo-portal. GeoDoc also supports a range of standards for the export of metadata records into XML and PDF files.</p> <p>My Go-Geo! is a service for registered Shibboleth users of the Go-Geo! Portal website. My Go-Geo! offers users customisation options and access to extra services on Go-Geo!. My Go-Geo! also provides a form for users to enter and store their contact details which are automatically transferred to the relevant element fields with the creation of each metadata record in GeoDoc. These contact details are repeated for data custodian, distributor and metadata creator, which in many instances, represents the same person. GeoDoc uses elements of the UK Academic Geospatial Metadata Application Profile (UK AGMAP), an ISO 19115-compliant application profile which comprises 97 elements made available to the UK academic community for describing spatial data and providing information for those wishing to acquire the dataset. This automation function populates 24 of the 97 contact element fields in GeoDoc. UK AGMAP also includes 36 mandatory elements, 15 of which will have been completed as soon as a user has started a new metadata record.</p> <p>Another GeoDoc feature which simplifies metadata creation is a map interface which allows a user to define the area or coverage of a dataset's study area extent. This is achieved through the use of the mouse and dragging it across corresponding area on the map, which then is displayed as a bounding box. Once the extent has been defined, the co-ordinate values are accepted and automatically entered into the relevant fields corresponding to the northern, southern, eastern and western extent of the dataset's study area.</p> <p>Capturing and documenting a dataset's spatial extent is a requirement for any metadata record submitted to a geo-portal for publication. The Go-Geo! Portal offers users coordinate-based spatial searching to retrieve metadata records. A Go-Geo! user can define a spatial extent which covers a specific area of interest, and metadata records with coordinates falling within this defined extent will be retrieved from the Go-Geo! catalogue or harvested from other portals. This is useful for those wishing to discover which datasets are available for a specific area. This offers the prospect of expanding research designs to incorporate new spatial datasets and eliminating the need to invest time and cost in creating datasets which already exist for that area of interest.</p> <p>GeoDoc also allows users to automatically submit their metadata records for publication on the Go-Geo! portal or academic institutional node, and export to various formats including the UK Academic Geospatial Metadata Application Profile (UK AGMAP), ISO 19115, the Federal Geographic Data Committees (FGDC) Content Standard for Digital Geospatial Metadata (CSDGM), Data Documentation Initiative (DDI) and Dublin Core.</p>

	<p>The Go-Geo! portal, is a JISC-funded service which includes My Go-Geo! and GeoDoc. Go-Geo! 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 within UK tertiary education and beyond. 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. The portal is a key component of the UK academic Spatial Data Infrastructure (SDI).</p>
--	---

<p>Standards</p>	<p>ISO 19115: 2003</p> <p>ISO 19115:2003 defines the schema required for describing geographic information and services. It provides information about the identification, the extent, the quality, the spatial and temporal schema, spatial reference, and distribution of digital geographic data.</p> <p>ISO 19115:2003 is applicable to:</p> <ul style="list-style-type: none"> • the cataloguing of datasets, clearinghouse activities, and the full description of datasets; • geographic datasets, dataset series, and individual geographic features and feature properties. <p>ISO 19115:2003 defines:</p> <ul style="list-style-type: none"> • mandatory and conditional metadata sections, metadata entities, and metadata elements; • the minimum set of metadata required to serve the full range of metadata applications (data discovery, determining data fitness for use, data access, data transfer, and use of digital data); • optional metadata elements - to allow for a more extensive standard description of geographic data, if required; • a method for extending metadata to fit specialized needs. <p>Though ISO 19115:2003 is applicable to digital data, its principles can be extended to many other forms of geographic data such as maps, charts and textual documents as well as non-geographic data.</p> <p>http://www.iso.org/iso/catalogue_detail.htm?csnumber=26020</p>
	<p>Federal Geographic Data Committees (FGDC) Content Standard for Digital Geospatial Metadata (CSDGM)</p> <p>The standard was developed from the perspective of defining the information required by a prospective user to determine the availability of a set of geospatial data, to determine the fitness the set of geospatial data for an intended use, to determine the means of accessing the set of geospatial data, and to successfully transfer the set of geospatial data. As such, the standard establishes the names of data elements and compound elements to be used for these purposes, the definitions of these data elements and compound elements, and information about the values that are to be provided for the data elements. The standard does not specify the means by which this information is organized in a computer system or in a data transfer, nor the means by which this information is transmitted, communicated, or presented to the user.</p> <p>http://www.fgdc.gov/standards/projects/FGDC-standards-projects/metadata/base-metadata/index.html</p>
	<p>Data Documentation Initiative (DDI)</p> <p>The Data Documentation Initiative is an international effort to establish a standard for technical documentation describing social science data. A membership-based Alliance is developing the DDI specification, which is written in XML.</p> <p>http://www.icpsr.umich.edu/DDI/</p>

	<p>Dublin Core (ISO 15836:2003)</p> <p>The Dublin Core metadata element set is a standard for cross-domain information resource description. It defines conventions for describing things online in ways that make them easy to find. Dublin Core is widely used to describe digital materials such as video, sound, image, text and composite media like web pages. http://dublincore.org/</p>
Key Issues	<p>Despite the development of this automation function to support the creation of geospatial metadata records, there is still resistance to this effort within UK academia, though more uptake has been experienced since Go-Geo! became a service in November 2008.</p> <p>Experience dictates that a dedicated geo-data repository is required to hold spatial data associated with the metadata created with GeoDoc, but this poses a problem with regards to open access to these datasets as many are created in UK academia, which are often derived from Ordnance Survey data products. The OS has strict policies with regards to residual IPR and licensing restrictions. EDINA, with initial financial support from the JISC has responded to this with the creation and support of ShareGeo, a spatial data repository; however restricted to only Digimap Collections subscribers.</p> <p>Other issues which affect success include the existence of legacy datasets and concerns expressed releasing too much information. A spatial data audit conducted in 2006 revealed the existence of 600+ spatial datasets at four participating universities. In addition, hundreds more ‘orphan’ datasets were identified. These are datasets with no provenance and would require further investigation for curation purposes. Academics note that there are no funds and time to endeavour in such investigations. Further automation of GeoDoc would prove useful to address this problem. Extraction functionality to capture a dataset’s spatial, date stamp (creation) as well as lineage, data processing and format would encourage more metadata creation. The GeoDoc map tool does offer limited automation functionality for entering the co-ordinate values of a dataset’s study area extents, but still requires user manipulation to define that area on the map using a mouse.</p> <p>Some academics and students have expressed concerns about IPR and copyright with the publication of their metadata records. A metadata record can reveal considerable information about a research project, a student’s dissertation or thesis; data requests can be made which requires copying data to a medium for dissemination and leaves any dataset left open for scrutiny.</p> <p>In short, the best automation cannot overcome some of these salient issues, but further automation of GeoDoc, and the inclusion of an open access spatial data repository would go some way to encourage more metadata creation across UK academia. Additional funding to address legacy datasets could play a primary role in supporting metadata creation. The EC INSPIRE Directive (currently being transposed into UK law) should also instigate more metadata creation as this Directive mandates that any dataset with an environmental aspect must be documented and made known to the public; this includes datasets created within academia.</p>

Scenario 2.

Title	Facilitating the process for a spatial data developer to upload datasets to a repository.
Author	Tony Mathys

Narrative	<p>Fiona is a researcher at the University of Edinburgh. She has created a spatial dataset which maps MP expense claims across parliamentary constituencies. As a responsible data developer, she recognises the importance of data management and sharing her dataset with other researchers within UK academia. She would like the opportunity to deposit her data on a repository.</p>
Tools	<p>ShareGeo, a spatial data repository with initial funding support from the JISC, is available under Digimap Collections at EDINA, a UK National Data Centre at The University of Edinburgh. ShareGeo facilitates submission of data with metadata automation function which extracts information from the dataset during the submission process and transfers this to the dataset's metadata record. This information includes the spatial extent or coverage of the dataset. An example is the area where fieldwork was conducted. This extent is expressed as a bounding box which is based on spatial co-ordinates for the farthest northern, southern, eastern and western points of the study area where fieldwork was conducted.</p> <p>Capturing and documenting a dataset's spatial extent is a requirement for any metadata record submitted to a geo-portal for publication or deposit on ShareGeo. ShareGeo offers users coordinate-based spatial searching to retrieve datasets. A ShareGeo user can define a spatial extent which covers a specific area of interest, and datasets with coordinates falling within this defined extent will be retrieved from ShareGeo. This is useful for those wishing to discover which datasets are available for a specific area. This offers the prospect of expanding research designs to incorporate new spatial datasets and eliminating the need to invest time and cost in creating datasets which already exist for that area of interest.</p> <p>ShareGeo also captures the file format of the dataset and transfers this to the metadata record. Dataset files submitted to ShareGeo must have the following format extensions:</p> <p>Vector GML (Geography Markup Language) GPX (GPS Exchange Format) KML (Keyhole Markup Language) MIF/MID (MapInfo Interchange Format) TAB (ID MAP DAT) (MapInfo Native format) SHP (SHX DBF) (ESRI Shapefile) E00 (ArcInfo Export)</p> <p>Raster ASC (ArcInfo ASCII) IMG (ERDAS Imagine) TIF (GeoTIFF) TIF Should also have .tfw or .tab world files (TIFF with World Files)</p> <p>Tabular CSV (Comma Separated Variable)</p>

Standards	<p>Dublin Core (ISO 15836:2003)</p> <p>The Dublin Core metadata element set is a standard for cross-domain information resource description. It defines conventions for describing things online in ways that make them easy to find. Dublin Core is widely used to describe digital materials such as video, sound, image, text and composite media like web pages.</p> <p>http://dublincore.org/</p>
Key Issues	<p>Many datasets created in UK academia are often derived from Ordnance Survey data products and the OS has strict policies with regards to residual IPR and licensing restrictions. ShareGeo has been developed to ingest these datasets, but access is restricted to only Digimap Collections subscribers. There are datasets deposited on ShareGeo which have no IPR or licensing restrictions, but until an open access version of ShareGeo is developed, access to these datasets will only be available to Digimap subscribers.</p> <p>Metadata for the datasets submitted to ShareGeo are based on Dublin Core, hence providing a limited amount of information. ShareGeo users can use GeoDoc to create descriptive level metadata and point to ShareGeo in their metadata records. ShareGeo offers users a standard xls spreadsheet for documenting metadata, which can be bundled into the dataset when downloaded.</p> <p>ShareGeo could also be enhanced to support extraction of other dataset information such as date stamp (creation), lineage and processing. Automation could also be employed to hold and transfer contact details to metadata records for those submitting multiple datasets to ShareGeo.</p> <p>Note: JISC has funded the development of a Geospatial Application Profile (GAP) that maps international (ISO) standards onto DC elements using a Dublin Core Application Profile, specifically for application within repositories. GAP represents a 'middle ground' between the fully descriptive UK AGMAP metadata standard (see above) and Dublin Core. JISC are deciding how this and other commissioned Application Profiles (e.g. for Images, time Based Media, Scientific Data sets), may be aligned consistently within the repository context in order to assist richer resource description and resource discovery.</p>

Scenario 3.

Title	A university wants to add tools and geo-referencing information to its repository collection to enhance search capabilities.
Author	James Reid and Tony Mathys
Narrative	<p>The University of Leicester would like to develop an institutional repository and recognize that many of the items likely to be deposited contain implicit geospatial references such as place names, postcodes, parish names etc. Ideally, when a resource is first ingested it can be geoparsed (i.e. natural language processing techniques used to mine the geospatial references from the resource) for geographic place names and georeferenced against an authority gazetteer – that is the textual references to places can be transformed into spatial coordinates (such as latitude/longitude) which provides a consistent and persistent spatial referencing for the resource and enables complex spatial relationships between resources to be performed e.g. 'show me all the resources within Cambridge for this resource'. Additionally, once the explicit georeferences are embedded with the resource, any geographical entry point to the resource can be used e.g. postcode searching when used in conjunction with the JISC funded shared service, GeoCrossWalk. GeoCrossWalk has the ability to translate one geographical reference into another so e.g. a postcode can be obtained from a coordinate, a place name from a postcode, a list of parish names from a county name etc. In practices this means that geographical entry points to resources can be whatever is appropriate to the application and is not restricted to the native geographic indexing term originally used (if at all!) with the resource.</p> <p>Furthermore, with geographic referencing added to a repository's content, e.g. text-based documents, images, etc., the portal can be enhanced to allow for geographical searches using a user-specified buffer which could define a specific range from the target place name and retrieve all photographs or documents falling within the defined range e.g. "find all pictures and documents associated with archaeological surveys conducted within 10 miles of the River Soar").</p>
Tools	<p>The GeoParser, a JISC-funded service run at EDINA, a UK National Data Centre at The University of Edinburgh and developed by the University's expert Language Technology Group, can assist with adding, enhancing or validating geographic referencing to text based documents, such as web sites, word processing documents and metadata records, which have textual references to geographical features in their contents. Such a tool could help add geographical coordinates to the vast amount of existing metadata which contains textual references to place (geographic features) yet has no explicit geographic referencing (e.g. the inclusion of latitude and longitude coordinates).</p> <p>Integration of the GeoParser via published APIs with, for example, repository services, could ensure that relevant metadata is extracted from resources as they are ingested into repositories. Such metadata will enable advanced geographic searching.</p> <p>Though the GeoParser can be treated as a distinct service, it is associated with GeoCrossWalk, a database of geographical features (like towns, rivers, woodlands and counties), their name and location. In other words, a digital gazetteer. GeoCrossWalk does not just store a feature's location as a point, it stores the features "footprint" where possible.</p>

	<p>In addition, GeoCrossWalk can make data, metadata and services 'geo-smart' in order to enhance resource searching, so users can search for resources based on geography. Many existing on-line services contain data, or records about the data (called "metadata"), which refer to geographical features but contain no information about the geographic location of the features. So although a user can define a search by specifying "what", "when" and "who" they are searching for, they can not specify "where". GeoCrossWalk allows such services to make use of the under-utilised geographical information in data and metadata and adds the "where" to resource searching.</p> <p>Finally, GeoCrossWalk can translate between different geographic reference or search terms to enhance portal cross-searching. e.g. a portal may search many resources, all with potentially different types of geographic referencing. A user is searching by Postcode but the resources are referenced by National Grid Reference, place name or parish code - <i>not</i> postcode. With the GeoCrossWalk middleware between the user and the resources being searched, these inconsistencies don't matter: GeoCrossWalk translates the user's postcode into whatever type of geographical reference is needed to search resources natively in the target resources. All this occurs invisibly to the user, who will enter a postcode and receive a list of results, unaware that the geographical cross-walking has taken place.</p>
Standards	<p>ISO 19112:2003</p> <p>ISO 19112:2003 defines the conceptual schema for spatial references based on geographic identifiers. It establishes a general model for spatial referencing using geographic identifiers, defines the components of a spatial reference system and defines the essential components of a gazetteer. Spatial referencing by coordinates is not addressed in this document; however, a mechanism for recording complementary coordinate references is included.</p> <p>ISO 19112:2003 assists users in understanding the spatial references used in datasets. It enables gazetteers to be constructed in a consistent manner and supports the development of other standards in the field of geographic information. It is applicable to digital geographic data, and its principles may be extended to other forms of geographic data such as maps, charts and textual documents.</p> <p>http://www.iso.org/iso/catalogue_detail.htm?csnumber=26017</p>
	<p>Shared Infrastructure Services are an important building block of the JISC Information Environment architecture¹. To enable portals and other presentation services to make sense of the diverse digital resource, machine-readable information about services, content, rights and users is required. This information must be provided in standard ways. Shared services is a common set of infrastructural services, that provide this information through machine-to-machine (m2m) interfaces, on which other elements of the environment, such as portals, brokers and aggregators, can draw. The result will be that a user can submit a search and receive details of the relevant resources in return, by matching the user's requirements, personal profile and institutional profile to descriptions of the content and the rights required to access it. In this way the user is guided to the most appropriate materials and at the same time the rights of publishers and other data owners are properly protected.</p> <p>¹http://www.ukoln.ac.uk/distributed-systems/jisc-ie/arch/standards/ http://www.ukoln.ac.uk/distributed-systems/jisc-ie/arch/ http://www.jisc.ac.uk/whatwedo/programmes/reppres/sharedservices</p>

Key Issues	The GeoCrossWalk gazetteer is built, in the most part, from OS data licensed under the HEFCE/OS agreement. Use of these OS datasets within HFE must only be within institutions which have subscribed to the Digimap service (and by necessity, also signed the OS/HEFCE sub-licence agreement) and persons within those institutions who are registered as Digimap users. This arrangement significantly restricts the use of GeoCrossWalk within HFE. However, work is currently under way to expand GeoCrossWalk with non-IPR restricted content which will also provide global geographical coverage for place names.
------------	---

Recommendations

Both the GeoDoc Metadata Editor Tool and ShareGeo Spatial Data Repository complement each other and are critical to the sustainability of the JISC-funded Go-Geo! Portal service. Portal users are being made aware that these comprise the Spatial Data Infrastructure (SDI) for UK academia and serve to provide resources for data management and sharing.

There is a growing awareness of the importance of an SDI in UK academia, but concerns persist with regards to IPR and data licensing restrictions as well as time and cost required to create and publish metadata. Metadata automation is intended to facilitate this to support concerns about the latter. The Go-Geo! Service also offers workshops and support for metadata creation, and institutional nodes are available upon request. An institutional node can address some concerns about IPR as metadata records published here are only accessible to those affiliated with the institution.

In conjunction with these efforts, ShareGeo has been developed to hold spatial datasets to reduce impact of data requests to data creators as well as hold data as part of data management. This reduces demands of disk space and storage on various media, hence providing cost and time-effective benefits to data developers.

The resources are in place and funding support for further automation will encourage further uptake of these tools, hence leading to more metadata record creation and publication on the Go-Geo! Portal, and submission of datasets to ShareGeo. The creation of an open access version of ShareGeo will instigate more spatial data submissions as many in UK academia conduct research outside the UK. More researchers are turning to the use of Global Positioning Systems (GPS) as well to map field data; this means more datasets without IPR or licensing restrictions.

JISC can offer support to this through with a more aggressive marketing and promotion strategy. In addition, the JISC can promote and mandate metadata creation based on the EC INSPIRE Directive. The objective of the EU INSPIRE Directive is to provide a better and more integrated information base to support the development and delivery of better European Community policies. The focus of INSPIRE is spatial information – defined as data referenced to a specific location or to an area. The INSPIRE Directive will create a European Spatial Data Infrastructure by standardising and harmonising the structure, content, format and access mechanisms for spatial information. It will be based on Member States' Spatial Data Infrastructures (SDIs). The directive will facilitate improvements in the sharing of spatial information between public authorities and provide improved public access. The common basic datasets such as administrative units, transport networks and property ownership will be among the first themes covered and will extend to commonly used themes later such as agriculture, energy, minerals and demographics. The Directive mandates that spatial information is made available to the public, hence, metadata, published on geo-portals.

The UK Government has agreed to the adoption of a joint transposition approach for the EC INSPIRE Directive; this action is taking place now with finalisation set for the latter part of 2009 according to INSPIRE at the GSDI/INSPIRE 2009 Conference in Rotterdam last month.

GeoCrossWalk has now reached a level of technical maturity that would enable it to support the tasks originally envisaged for it – enhancing geographic searching within the JISC IE. It is worth remembering that most resources have some geographic component to them and can thus be indexed geographically (and hence searched geographically). GeoCrossWalk provides the facility to resolve dialectical differences between services and can assist in ‘geo-enabling’ legacy resources. Geography is a very powerful synergistic means of querying resources and to date has been something of a Cinderella in the digital library community. As Google note, geography is a powerful organizing principle for information resources.

To build upon the achievements of the Project to date, ongoing cultivation of the potential services that could fruitfully exploit GeoCrossWalk (in its various guises – middleware server, gazetteer database, and the GeoParser) remains a significant hurdle. Whilst the growing popularity of a ‘spatial approach’ to resource discovery, access and synthesis is evident outside the JISC IE, the opportunity to realise the same innovation within the IE requires strategic impetus from JISC allied with a more aggressive marketing and promotion strategy. While the server interface can remain constant, there is no single, generically applicable client solution possible – each service must be considered in turn and implementation planned in sympathy with ongoing service operation and upgrade.

The GeoParser should be marketed in its own right. Although it was built and tested using GeoCrossWalk as the reference gazetteer, it can in fact use any suitable digital gazetteer and as such, should not be promoted solely as a GeoCrossWalk product per se. Giving the GeoParser a life outside of the GeoCrossWalk brand will help clarify both its function and utility and that of GeoCrossWalk.