

Creating a UK web-based service for discovery of location-based data resources

Rob Walker
Independent Consultant.

Rob Walker Consultancy
64 Histon Road
Cottenham
Cambridgeshire
CB24 8UD

Tel 01954 251003
Email robwalker@cix.co.uk

12th October 2007

Contents

	<u>Page</u>
1. Introduction.....	4
2. Metadata	4
3. gigateway	6
4. External drivers	7
5. Operation of a metadata service	8
5.1 Discovery	8
5.2 Usage	8
5.3 Data input	9
5.4 Data quality	9
6. The case for continuing gigateway	10
7. Plan for implementation.....	11
8. Recommendations	12
Annex A Draft INSPIRE discovery metadata elements	13

Summary

This report looks at the issue of how a discovery metadata service for spatial information might operate, and identifies some of the decisions that need to be taken in order to set up such a service. It is intended to assist government, including the Department for Communities and Local Government to determine the way ahead with this.

A metadata service provides access to high level information about data resources, to enable potential users to find out what information exists, and to determine whether it is suitable for the intended use. The EU INSPIRE Directive mandates making information available about spatial data and making it possible to discover, inventory and use them. It envisages a “metadata service”, operating on a common basis across Europe.

The UK currently has this type of service in *gigateway* which has been running for several years. However, *gigateway* contains information about only a limited amount of data resources, and uses old technology. It should be regarded as a prototype for a full national metadata service.

The report reviews the current situation, including the existing *gigateway* service, and describes how a new service might be implemented. It builds upon earlier reports to ODPM/CLG by Les Rackham.

The report describes aspects of the operation of a metadata service, and provides an outline plan for implementation. The first step of this is to set up a broadly-based Steering Group to oversee implementation. Any implementation should be tied in with implementation of INSPIRE. It is recommended that in the meantime, *gigateway* should continue to be run on a “care and maintenance” basis, and a case for this is provided.

1. Introduction

This report looks at the issue of how a discovery metadata service for spatial information might operate, and identifies some of the decisions that need to be taken in order to set up such a service. It is intended to assist government, including the Department for Communities and Local Government to determine the way ahead with this.

The EU INSPIRE Directive [1] mandates making information available about spatial data and making it possible to discover, inventory and use them. "Spatial data" in this context means any data with a direct or indirect reference to a specific location or geographical area, and includes not just traditional map-based information, but also data classified by area, for example neighbourhood statistics¹.

A metadata service provides access to high level information about data resources, to enable potential users to find out what information exists, and to determine whether it is suitable for the intended use. The INSPIRE programme envisages a "metadata service", operating on a common basis across Europe.

The UK currently has this type of service in *gigateway* which has been running for several years. However, *gigateway* contains information about only a limited amount of data resources, and uses old technology. It should be regarded as prototype for a full national metadata service.

The report reviews the current situation, including the existing *gigateway* service, and describes how a service might be implemented. It builds upon a report to ODPM/CLG [2] by Les Rackham which described the arrangements, future governance and specification of a metadata service for geographic information. That in turn built on an earlier report on the sustainability of a UK Metadata Service for Geographically Related Information [3].

2. Metadata

Metadata is defined as "data about data". It means information describing spatial data sets and spatial data services and can be associated with any data resource, such as document, a dataset or other source of information. Metadata provides additional information about the data resource, to enable it to be better understood and used to good effect, for example what it is about, how it was captured and its currency.

¹ Many datasets that at first sight do not appear to be spatial nevertheless do have a spatial content, in that they apply to a limited geographic area, for example statistics for a local authority area. Spatial data contains spatial references which may take the form of coordinates, for example in latitude and longitude, or references to geographic place names, for example street data.

There are a range of uses for metadata, for discovery, for evaluation and for use:

- **discovery**: the user aims to find out what available resources are potentially able to satisfy a specified set of requirements. This is typically what a search engine can process, using basic search criteria to identify the available resources corresponding more or less to the user requirements and providing basic metadata (name, content description, geographic area of applicability, etc) about the candidate resources.
- **evaluation**: the user needs to go deeper in the metadata (e.g. looking at the quality information) in order to ascertain whether a candidate resource fits for the intended purpose.
- **use**: the user has selected a candidate resource, but needs to access it and to configure a system or software to process it.

Metadata services provide one of the fundamental parts of a Spatial Data infrastructure (SDI)². They are used within organisations as part of the information management facilities, and on a national basis for discovery purposes. Essentially, they work on the basis of a user defining parameters such as topic and geographic extent, to carry out a search to discover data resources that might be suitable and return information about their source, content and availability.

There are three generic actors in the metadata service:

- **Metadata creator** – responsible for creating and maintaining the metadata and for its quality. Although the metadata creator is usually the data producer or distributor, this is not always so.
- **Service provider** – runs the metadata service. They may be part of the same organisation as the metadata creator or quite separate. A broad view is taken here of the role, which in fact may be made up of several actual roles in the real world - for example:
 - a contractor hosting the service and providing IT support, who is contracted to
 - a service owner who is ultimately responsible for the quality of the service and has service level agreements (SLAs) with the contractor and metadata creators.
- **Service user** - the consumer of the service who selects the search criteria matching their requirements, performs the searches and hopefully finds data resources meeting their requirements or, at least, meriting further investigation.

² An SDI is defined by INSPIRE as metadata, spatial datasets and spatial data services; network services and technologies; agreements on sharing, access and use; and coordination and monitoring mechanisms, processes and procedures.

A metadata service is not just about discovering maps or known spatial data, but may include all datasets with a geographic dimension, and those whose existence is not widely known.

3. *gigateway*

gigateway is a discovery metadata service operated by the Association for Geographic Information (AGI) on behalf of the Department for Communities and Local Government (CLG). It was previously funded under the National Interest Mapping Agreement (NIMSA). Since April 2006, it has operated on a “care and maintenance” basis, whereby the service is maintained and data providers and users supported, but no attempt is made to expand the service through inclusion of additional data, to improve the quality of existing data, or encourage usage.

EDINA, the national academic data centre based at the University of Edinburgh, provide and maintain hardware for hosting the *gigateway* website and host the *gigateway* central catalogue node. However, the data is not held centrally, but is distributed across a number of “nodes”, some hosted by metadata providers and others by EDINA. There is no requirement for data providers to have their own node. The *gigateway* catalogue is available to metadata providers that do not have a separate node.

The nodes are:

- AGI Cymru Environmental Data
- British Atmospheric Data Centre
- British geological Survey (BGS)
- Central Government (IGGI)
- Centre for Ecology and Hydrology (CEH)
- GeoHub-NI (formerly Mosaic)
- Gigateway catalogue
- GVAC – UK Higher Education
- NERC Earth Observation Data Centre
- Ordnance Survey
- QinetiQ
- UK Local Authority Catalogue

A detailed description of *gigateway* is given in Reference 1.

The report by Les Rackham [2] into the arrangements, future governance and specification of the service in November 2005, written before the impact of INSPIRE became clear, identified the following lessons from the operation of the current service:

- The service cannot operate in a vacuum – it needs to operate in the context of a broader national strategy for geographic information;
- The strategy cannot be contracted out to the service provider – as there will be little strategic direction;
- The need for effective governance – to provide drive and direction;
- Metadata quality – required for successful and accurate searches;

- Core datasets - these must be included, and a strategic approach to their identification and acquisition is required;
- Support for metadata providers – required to improve the quality of the metadata;
- Meaningful monitors and measures – to gauge the success of the service;
- Promoting the service – at all levels, including high levels of government;
- Contracting out on the open market will not be straightforward – the service provider role needs to be broad, and they need to work in partnership with the owners who provide leadership and direction;
- Transition from the current service – the metadata providers need to be kept on board to ensure their cooperation and goodwill.

Glgateway is built on an old technology that cannot support current requirements primarily because the information model (the structure and semantics of the underlying metadata) cannot be adapted. It also lacks the ability to perform searches based on a gazetteer.

4. External drivers

The EU INSPIRE Directive [1] mandates making up-to-date metadata available for describing spatial datasets to facilitate their discovery. Details of how this will be implemented will be defined in a set of Implementing Rules (IR). These Implementing Rules are being developed by Drafting Teams comprising sponsored individuals from the user community, and will be based on existing International Standards. Implementing Rules for Metadata will define the minimum set of metadata to be recorded, and are due to be adopted by May 2008. Metadata must be created within 2 years of this for INSPIRE Annex I and Annex II spatial data themes. Implementing Rules for Network Services will define the way in which a metadata service will operate in order to achieve interoperability across Europe.

The Draft Implementing Rules for Metadata [4] are at a fairly advanced stage of development, and the set of elements that it specifies is fairly stable. It follows the core metadata set defined in ISO 19115 [5], with additional elements to cover services. These elements are listed in Annex A. The set is close to that defined in UK GEMINI [6], the UK metadata profile. Version 2 of this, currently available in draft form, provides a revised set of metadata elements that is compatible with the INSPIRE set. It will be finalised when the final Implementing Rules for Metadata are available.

The Implementing Rules for Network services have yet to appear. However, it is understood that they will be based upon the Web Map Server principles of the OpenGIS Catalogue Services Implementation Specification [7] with the ISO 19115/ISO 19119 Application Profile. Further details are given in a recent report by Atkins Ltd, commissioned by AGI [8].

In addition, the UK Location Strategy (not yet published), is likely to identify a metadata service as one of the key components.

5. Operation of a metadata service

5.1 Discovery

A discovery metadata service contains high-level information about data resources to enable users to search on a set of criteria to discover what data exists to meet that requirement. It is likely to be web-based. The central service will handle queries, but the data is likely to be held on a set of nodes, possibly held by major data providers or data brokers. There will be facilities for data providers to load new metadata and maintain existing metadata. The central service must be able to carry out basic quality assessment on the incoming metadata.

The user will be able to ask basic queries such as what data is available on a specified topic, for a specified geographic area, for a given time period, for example, find all datasets relating to incidents of ground water pollution in Yorkshire during 2006. The service will return information about all data resources meeting these criteria. This information should be sufficient for the user to understand the data, and determine whether it is suitable, and how it may be obtained.

5.2 Usage

Many organisations keep detailed metadata for internal data management purposes. This enables them record the extent of their data holdings and to keep track of updates. A simplified view of this may be offered externally as discovery metadata, and this is done by some of the *gigateway* node holders (e.g. British Geological Survey). However, most holders of data on *gigateway* provide metadata primarily for external purposes.

Users of a metadata service are seeking to find out what data exists that might be relevant to them for a particular application. Examples of general queries to a discovery metadata service that might be available on an enhanced service with comprehensive data are as follows:

- **incident response and investigation:** when an incident occurs, for example an outbreak of foot and mouth disease, there is an immediate requirement to find data relating to things like pollution history, animal movements, incidence of notifiable diseases, and recent building work. This information is continually changing and access to the most up-to-date version is essential.
- **emergency planning:** much attention is now being paid to emergency planning for disaster recovery. Knowledge of what information exists about a range of phenomena including provision of emergency response resources, location of at-risk residents, location of hazardous facilities, and safe access routes needs to be obtained.

- **strategic planning:** this involves access to information on a range of topics, that may not be obvious before the search exercise began.

5.3 Data input

The success of any metadata service will depend on the quantity and quality of the metadata that it contains. The crucial issue becomes how to populate the service with suitable data. The requirement of a national metadata service is to include all public datasets, starting with those in key areas such as those defined in some of the INSPIRE Annex1 data themes:

- geographic names (areas, regions, places, geographical or topographical features of public or historic interest);
- administrative units;
- addresses (of properties);
- cadastral parcels (registered land);
- transport networks (road, rail, air and water).

This will then be supplemented by other data, including that from other non-public-sector data suppliers.

The service should provide basic tools to input data such as the MetaGenie tool provided with *gigateway* and assist data suppliers by providing support services to them.

Although the onus will be on the data suppliers to provide their own data, there may be a case for the service to create metadata for a few key datasets.

Metadata should be considered as part of the essential documentation of a data resource. However, there is often a reluctance for data providers to create and provide metadata as this is seen as an additional burden for which resources are not made available. However, organisations that manage their information holding in a professional manner, such as British Geological Survey, keep metadata as part of their routine operations. The primary use of this is for internal management purposes, but a high-level extract of the information is provided externally to assist their users. When maintained as part of their internal management process, the cost becomes included in the overall costs. For organisations that create many one-off datasets, such as many government departments, metadata is often not created. This may be because at the time of creation of the dataset, when it would have been easy, it was not considered important. However, the hidden costs of not knowing about existing datasets, necessitating costly re-capture, or decision making based upon less-than optimum information can be considerable, though hidden from the data owner.

5.4 Data quality

In any metadata service, data quality is a major issue. If the data is not of a suitable quality, in terms of completeness and accuracy, then the service will fall into disrepute. The quality of the metadata on *gigateway* is seen as an issue. AGI have produced a set of guidelines for metadata [9], including consideration of quality, to assist users with this. An attempt was made to test

the quality of the incoming data, but this is no longer happening with the downgrading of the operational activities.

The main quality issues that need to be considered are as follows:

- Completeness – is there metadata for all the major datasets within scope, and is that metadata complete?
- Up-to-dateness – does the metadata represent the current state of the dataset?
- Consistency – are similar datasets described in a similar way?
- Understandability - is it possible to get a clear idea of the dataset from the metadata? (e.g. does the abstract give a good picture of the dataset?)
- Accuracy – are the metadata items accurate? (this is largely up to the data suppliers).

For a more detailed discussion of data quality for metadata, see Part 3 of the Metadata Guidelines.

A metadata service needs to establish quality evaluation procedures to check that all data being added to the service meets acceptable quality levels.

6. The case for continuing *gigateway*

There is no doubt that a geospatial metadata service will be required in the future to meet the INSPIRE and other requirements. The timescales for this will depend on availability of funding, and resolution of issues of governance. *gigateway* is currently funded until the end of March 2008. It is quite likely that arrangements for a successor service will not be finalised for it to take over at that point. Therefore there is the question of what to do about *gigateway* in the interim. The options would be either to continue on a “Care and Maintenance” basis, or else to abandon it altogether.

If *gigateway* were closed down, the following would happen:

- there would be no central service for locating data;
- users would be discouraged from searching for existing data;
- individual nodes (listed in 3) would carry on, but they would be unsupported;
- data suppliers users would not collect metadata and existing metadata would not be maintained and would eventually reach the state where it is unusable;
- expertise in running a metadata service would be lost;
- There would be a saving of approximately £100,000 per annum (this might be only for a period of six months if INSPIRE implementation proceeds to the current schedule).

If *gigateway* were to continue on a “Care and Maintenance” basis until a successor service was available to take over from it, the following would happen:

- The existing data discovery service, for all its limitations, would be maintained;
- Individual nodes would continue to be supported;
- Existing data suppliers would be encouraged to continue to provide and maintain their metadata;
- Expertise in running a metadata service would be retained, to aid setting up a replacement service.

7. Plan for implementation

1. Establish a governance structure

The main tasks to be carried out are:

- Strategic management;
- Securing and managing the funding;
- Operational and contract management;
- Managing change;
- Measuring and monitoring performance of the service;
- Communication;
- Promotion of the service;
- Selling and marketing the service.

The recommended governance structure is to have a Management Board representatives from the major stakeholders, chaired by the sponsoring Department, and operational management provided by the Department.

2. Statement of technical requirements

Any metadata service will be web-based. The technology will not be the same as that used for *gigateway* (Z39.50, which is now obsolete), but will use that recommended by the INSPIRE Implementing Rules. The likely technology will follow the OpenGIS Catalogue Services Implementation specification version 2.0 with the ISO 19115/ISO19119 Application Profile, as described in reference 7. The metadata data structure will similarly be different, not based upon the NGDF specification, but on UK GEMINI [5]. This is being revised to conform to the evolving INSPIRE metadata set.

3. Tender for services

How this is done is up to the sponsoring Department. Details of how this could be done are given in the earlier report *Arrangements, future governance and specification of service* [2]. However, this will require review and updating in the light of technical developments mentioned above, and INSPIRE requirements.

4. Data collection

There are three aspects to data collection

- **migration of existing data** – the existing data in *gigateway* can be transferred into the new service, as there is a direct mapping from the old data structure into the new. Most of the additional metadata elements required (e.g. language) can be provided automatically with default values. Quality control will need to be applied to ensure that

data is of an acceptable quality, and some additional manual data cleansing may be required.

- **updating of existing data** – much of the existing data in *gigateway* is out of date, and a programme of updates will need to be initiated. This can be done on a managed supplier-by-supplier basis.
- **collection of other data** – additional data providers can be targeted to get their data, with support services provided to assist them.

This will need to be controlled by the service manager, to ensure that the data is of acceptable quality.

An outline programme to achieve this to meet a target date of Quarter 2 2009 (in line with the initial INSPIRE requirements) is as follows:

<u>Period</u>	<u>Activity</u>
Q1 08	Set up Steering Group – broadly based, user community
Q2 08	Produce Statement of Requirements for service
Q3 08	Tender for service contract
Q4 08	Decision on award of service contract
Q1 09	Implementation including data migration
Q2 09 –Q2 10	Collection of data for INSPIRE Annex I & II themes

8. Recommendations

The following course of action is recommended:

1. Commit to a UK geospatial metadata service.
2. Set up a Steering Group that is broadly based, to oversee the implementation of a service.
3. Tie the implementation of the metadata service into the UK Location Strategy and the implementation of INSPIRE.
4. Continue *gigateway* on a care and maintenance basis in the interim, until the replacement service is available.

Annex A Draft INSPIRE discovery metadata elements

The following are the draft set of INSPIRE discovery metadata elements:

Resource title – the identifier of the resource;

Temporal reference – the time or time period covered by the resource;

Geographic extent of the resource – the area to which the resource applies;

Resource language – the language(s) used within the resource;

Resource topic category – a high-level classification using standard categories;

Keyword – commonly-used words to describe the subject;

Service type – interface specification (applies only to services);

Resource Responsible party – for establishment, management, maintenance or distribution;

Abstract – brief narrative summary of the content;

Resource locator – where the resource may be found;

Constraint – any constraint(s) placed upon the access or use of the resource;

Lineage – general information about the history of the resource;

Service type version – the interface specification (applies only to services);

Operation name – associated with the service type (applies only to services);

Distributed computing platform – on which the service is deployed (applies only to services);

Resource identifier – an unambiguous reference;

Spatial resolution – the level of detail of the data;

Conformity – how well the resource conforms to the INSPIRE Implementing Rules.

References

1. European Union. Directive of the European Parliament and of the Council establishing an Infrastructure for Spatial Information in the European Community (INSPIRE). see www.ec-gis.org/inspire
2. Les Rackham. Arrangements, future governance and specification of service. November 2005
3. Rackham L J (2004) An independent Review of the sustainability of a UK Metadata Service for Geographically Related Information. See <http://www.gigateway.org.uk/aboutus/aboutus.html>
4. INSPIRE Metadata Drafting Team. Draft Implementing Rules for Metadata. See www.ec-gis.org/inspire
5. ISO 19115 Geographic information - Metadata
6. UK GEMINI – A UK Metadata Standard for discovery of geographic data resources. Available at <http://www.gigateway.org.uk/metadata/standards.html>
7. Open Geospatial Consortium. Open GIS Catalogue Services Specification v2.0.2 2007.
8. Atkins Ltd. Future technology for a UK Metadata service. Available at <http://www.gigateway.org.uk/aboutus/aboutus.html>
9. AGI. Metadata Guidelines for Geospatial Datasets in the UK
 - Part 1: Introduction to metadata
 - Part 2: Creating metadata using UK GEMINI
 - Part 3: Metadata qualityAvailable from www.gigateway.org.uk/standards.html