Metadata Guidelines for Geospatial Data Resources - Part 1

Introduction

September 2010

Rob Walker and Les Rackham, Rob Walker Consultancy Ltd have written this document, with funding from the residue of the National Geospatial Data Framework (NGDF).

Contents

Preface

This is the first part of a set of guidelines for metadata for geospatial data resources. These Guidelines are intended for general use in the UK geographic information environment. They are primarily concerned with geospatial data (i.e. that which references data to a location on the surface of the Earth), and which has a limited geographic extent (i.e. is restricted to a defined territory). They have been developed within the context of a national geospatial metadata service, and the UK GEMINI2 metadata standard. However, they are sufficiently broadly based to be applicable in a wider context of geospatial metadata creation and management.

The Guidelines are aimed at data managers and creators of metadata, providers of metadata services and general data users. They include guidance on quality management such that they could be used in the context of a national metadata service.

This Part of the Guidelines covers the basics of metadata and provides an introduction to the other two parts. Part 2 provides a set of detailed guidelines for creating metadata to the UK GEMINI2 standard. Part 3 deals with metadata quality and covers quality evaluation and quality management of metadata including guidance on establishing acceptable quality levels.

The Guidelines have been revised during 2010 to take into account other developments, including UK GEMINI2. Any comments on these Guidelines or on the UK GEMINI2 metadata standard should be sent to standards@agi.org.uk.

1. INTRODUCTION

This is the first part of a set of guidelines for the creation, maintenance and quality management of metadata for geospatial data resources. It provides a general introduction to the principles and concepts of metadata. It is aimed at data managers and creators of metadata, providers of metadata services and general data users.

The data resources may be datasets, dataset series, services delivering geographic data, or any other information resource with a geospatial content. This includes datasets that relate to a limited geographic area. The data resources may be graphical or textual (tabular or free text), hardcopy or digital. Geospatial data is data containing a positional or locational element relative to the Earth. Many data resources that at first sight do not appear to be geospatial nevertheless do have a geospatial component, in that they apply to a limited geographic area, for example statistics for a local authority area. Geospatial 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.

Metadata is data about data. It provides additional information about the data resource, to enable it to be better understood and used to good effect.

In an organisation, metadata is required for both internal and external purposes. It can add significantly to the value of an organisation’s data holdings, and failure to create it at the appropriate time can lead to greater hidden costs later. Lack of knowledge about available data can lead to costly duplication of effort.

Within an organisation, metadata is required to organise and maintain the internal investment in data. As staff change, institutional knowledge leaves the organisation. Undocumented data can lose its value, and subsequent staff may have little understanding of the contents and use of the data and may find that they cannot trust results generated using this data.

Externally, metadata is required to provide information about an organisation’s data holdings. Data resources are a major national asset, and information of what data resources exist within different organisations, particularly in the public sector, is required to improve efficiencies and reduce data duplication. Data catalogues and data discovery services enable potential users to find, evaluate and use that data, thereby increasing its value. In addition, metadata received from an external source requires further information supplied by metadata in order to process and interpret it.

The requirements for metadata for internal and external purposes are different. Metadata for external purposes will be at a general level, providing basic information about the data resources. Having this in a standardised form enables metadata services to be set up for widespread use. Standardisation of this aspect is thus important, and most metadata standards are designed for this purpose. Metadata for internal purposes will be much more detailed. Internal standards for metadata will generally be extensions of those required for external purposes, and often will contain items particular to the organisation or business sector. Whilst the details of these Guidelines relate to external metadata, and in particular the UK GEMINI2 metadata set designed for discovery metadata services, the principles are equally applicable to internal metadata.

A glossary of terms is provided in Annex A.

2. FUNDAMENTALS OF METADATA

2.1 The nature of metadata

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 to the users 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 of the information) in order to ascertain whether a candidate resource is fit 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.

2.2 Metadata services

Metadata services provide one of the fundamental parts of a Spatial Data Infrastructure (SDI). 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 Category and Extent, to carry out a search to discover data resources that might be suitable and return information about their source, content and availability. A discovery metadata service is being established under the UK Location Programme[1].

2.3 Metadata as a business process

A simple process model is presented at Figure 1 and shows the flow from metadata creation through to metadata query with quality related processes built-in. It can represent the processes within one organisation or split between separate organisations. The flow is idealised and could relate to any discovery-level metadata service. This model is developed further in Part 3 of the Guidelines.

Although metadata creation and maintenance are shown as separate operations, it cannot be over-emphasised that they will be far more successful if they are fully integrated into the other business processes of an organisation. Thus, when a dataset or other form of data resource is produced, it should be routinely documented as metadata as part of the production or distribution process. Further, when the data resource is updated or subject to some form of change, the metadata should also be updated.

[1] see http://location.defra.gov.uk

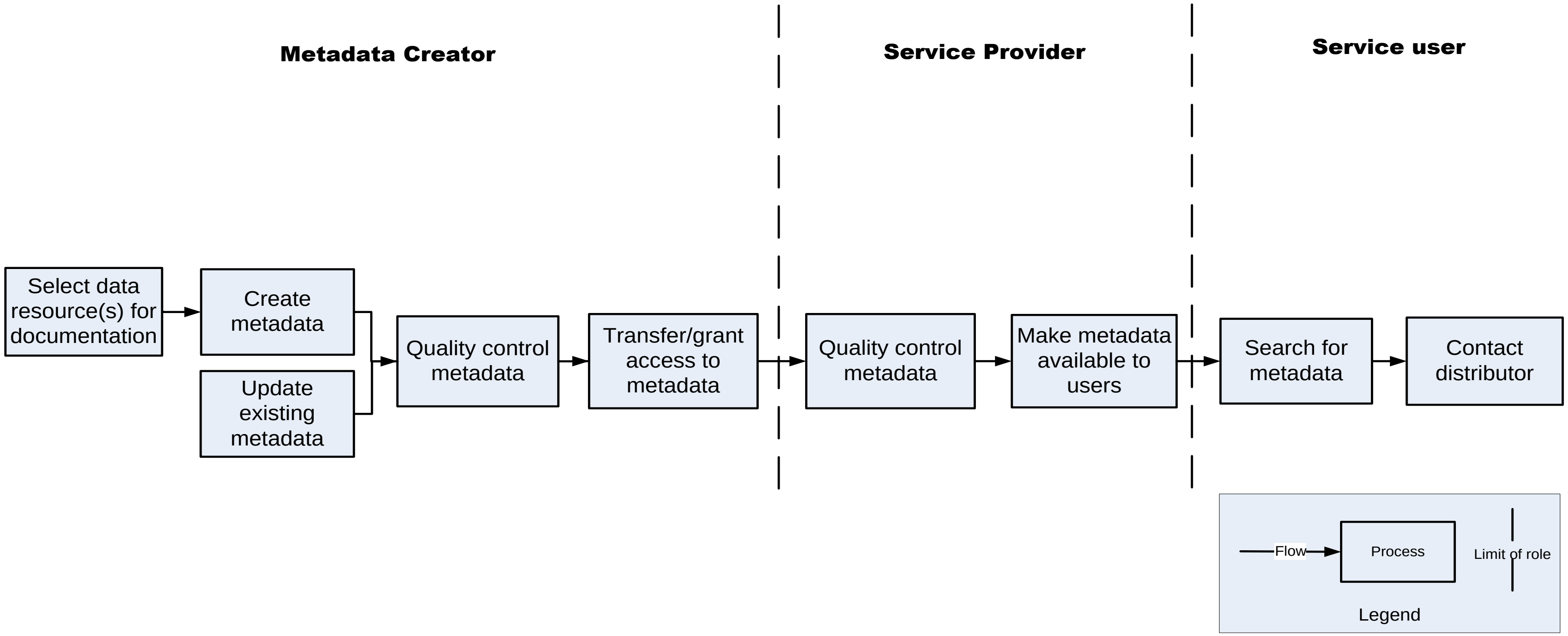

Figure 1. Simple process model for metadata creation, maintenance, service provision and use

If the main purpose of creating metadata is to document and enable discovery of an organisation's own data resources and the exposure of all or part of that metadata to some external service is secondary, then there will be far greater chance of support and resources for metadata in the business.

However, if metadata is seen as an additional activity to support some external service, it is likely to be ignored or forgotten. When it is finally picked up, it may well be assigned to someone who has no knowledge of the data resource and no interest in the quality of the metadata.

2.4 Metadata roles

Three generic roles in the metadata process can be recognised, these are illustrated in Figure 1:

- **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, and may be part of the same organisation as the metadata creator or quite separate. A broad view of the role is taken here which 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 finds data resources meeting their requirements or, at least, meriting further investigation.

2.5 Metadata standards

There are many metadata standards in existence. These have been produced at different times by different bodies for different purposes. The main ones that are relevant to geospatial data resources are:

1. **Dublin Core**

   This was originally developed by librarians for cataloguing information resources. It uses free-text fields, which makes automatic searching difficult. Consequently, it not ideally suited for discovery purposes using electronic data services. It is severely limited in its ability to handle the geospatial aspects of data, and also in how it handles the geographic extent of non-geographic data, e.g. data that applies to one country or region rather than another.

2. **e-GMS**

   The UK e-Government Metadata Standard. This forms part of the e-Government Interoperability Framework (eGIF) and is mandated for use in central and local government in the UK. Although it was intended to support resource discovery, it is based upon Dublin Core, so is not ideally suited for discovery metadata services, and its capabilities to handle geospatial data are limited.

3. **FGDC**

   This was developed by the US Federal Geographic Data Committee for discovery services for geospatial data. It has been largely superseded by ISO 19115.

4. **ISO 19115**

   This describes all aspects of geospatial metadata and provides a comprehensive set of metadata elements. It is designed for electronic metadata services, and the elements are designed to be searchable wherever possible. It is widely used as the basis for geospatial metadata services. However, because of the large number of metadata elements and the complexity of its data model, it is difficult to implement.

5. **INSPIRE**

   The INSPIRE metadata Implementing Rules defines the minimum set of metadata elements necessary to comply with the INSPIRE Directive. In essence it is a profile of ISO 19115 for discovery purposes. It allows a variety of possible implementations.

6. **UK GEMINI**

   The UK Geospatial Metadata Interoperability Initiative. It was originally produced by a collaboration between the Association for Geographic Information (AGI), the e-Government Unit (eGU) and the UK Data Archive. It was designed for use in *gigateway* to replace the existing metadata set with a set that is compatible with the core elements of ISO 19115. Version 2.1 complies with the INSPIRE metadata set, and has been adopted for the UK Location Programme metadata service.

Formal references for these standards are given in the Bibliography.

3. GENERAL PRINCIPLES OF METADATA

3.1 Introduction

This section describes some general high-level rules that apply to the creation of metadata. Whilst primarily applying to UK GEMINI2, they are also relevant to other metadata specifications.

3.2 Free text

The aim of a discovery metadata specification such as UK GEMINI2 is to define metadata in a form that provides easily understood information for potential users of the data resource, and is searchable in a computerised discovery metadata service such as that being developed for the UK Location Programme. It is difficult to carry out searches on free text, since the same thing can be written in different ways, for example “OS”, “Ordnance Survey”, “The Ordnance Survey of Great Britain”, “OSGB”. Hence, free text entries are discouraged where the metadata is searchable. Where only a limited number of options is available, code lists are specified. For some metadata elements, such as Topic, these may not correspond exactly to the theme of the dataset, and the nearest option should be chosen. Where code lists are used, input and output facilities in a metadata service should identify the corresponding element value.

3.3 Obligation

As with other metadata specifications, elements in UK GEMINI2 are given as mandatory, conditional or optional. Mandatory elements must be completed, and most software tools do not allow the metadata to be entered if any mandatory elements are missing. Conditional elements have a condition associated with them, and should be supplied when that condition is fulfilled. Such conditions usually define the applicability of the element. Optional elements may not always be completed. This can be for a number of reasons, for example if the element is not relevant or its value is not known. In practice, optional elements are really conditional, and there is a set of circumstances in which a value should be given, roughly corresponding to “is it relevant?” and “is its value known?” Optional elements should not be ignored.

3.4 Element domains

Each metadata element has a range of allowable values, called its domain. These may be values in a given range (e.g. positive integers) or a set of code values chosen from a list. Specifying the domain helps with both the initial creation of data and quality checking[2]. In some cases, default values may be used, for example for elements relating to the metadata itself.

3.5 Spatial references

By definition, geographic data contains some form of spatial reference. The spatial reference identifies the position of the object of interest in the real world. Spatial references can be given in a number of forms, not just as Grid References. They can take the form of the name or identifier of a geographic location which can be

[2] For further details on quality checking, see Part 3 of these Guidelines

described in a gazetteer. Examples are property addresses, postcodes and census areas. These spatial references are a key means of searching the data by location, not only within the data resource, but also positioning the data resource in the world (e.g. data for Scotland). A consistent set of spatial references enables spatial searches to be made for datasets in a metadata service.

3.6 Date fields

Metadata can contain a multitude of dates identifying the different stages in the life cycle of the data. Key dates are included in metadata. To enable searches to be carried out (e.g. date relating to the period 1990 to 1999), these dates need to be recorded in standardised form. Unfortunately, different standards are used in different places. Here it is recommended that the extended format defined in ISO 8601[3] (YYYY-MM-DD) is used.

4. DATA RESOURCES IN SCOPE

4.1 What is stated in UK GEMINI2

The purpose of this section is to provide guidance on (i) what types of geospatial data can be documented using UK GEMINI2, (ii) the applicability of the Standard and (iii) how to find an appropriate level for the individual documentation of data resources.

UK GEMINI2 "specifies a set of metadata elements for describing geographic data resources" but provides no other information about what types of data are in or out of scope.

In practice it can be difficult to decide what data resources are in scope and how they should be documented. For example, what constitutes a suitable body of data for individual documentation? Should it be an individual map or the whole map series?

There is no simple answer to these questions; it is likely to be a compromise. However, some general guidance can be given.

4.2 Characteristics of data resources in scope

The general nature of geospatial data is described in the Introduction. As is emphasised there, many data resources that at first sight do not appear to be geospatial nevertheless have a geospatial component. They are geographically constrained in some way, in that the data only refers to certain areas or locations on the surface of the Earth, whether on land or sea. The way that these locations are referenced can take many forms such as coordinates (e.g. latitude and longitude, National Grid), or geographic place names (e.g. street, locality, town, administrative area).

Although it is common to think of geospatial data as being synonymous with maps, this is far from the case. Geospatial data can equally well be textual - whether tabular or free text – or images taken from the air or from the ground. The data can be as hardcopy or in digital or even video form.

All these types can be documented using UK GEMINI2.

[3] ISO 8601 Data elements and interchange formats – Information interchange – Representation of dates and times.

4.3 Applicability of UK GEMINI2

The primary purpose of UK GEMINI2 is to provide a standard for documenting data resources within the United Kingdom that complies with the INSPIRE Metadata Implementing Rules. This does not mean that it cannot be used for documenting data resources that reference locations outside the United Kingdom. Indeed the standard expressly allows for this. Domains may be extended using the rules of ISO 19115.

4.4 Levels of data resources for documentation

There are no absolute rules for deciding on an appropriate level for the individual documentation of a data resource. The overriding consideration is that the data resource has been documented with sufficient granularity to yield a useful result if discovered using a metadata service. Too coarse a granularity will result in too generalised a result, too fine a granularity is likely to overwhelm the user (and the metadata creator!). The granularity can relate to geographical extent, temporal extent or subject.

There are some questions that can be posed which may help the metadata creator to find an answer.

- How is the data resource used and how is it made available? Is it a product, dataset, document in its own right that may be used and combined with other datasets or is it an integral part of a larger data resource?
- Has the data resource been captured using a single data specification? Are there other data resources captured using the same data specification?
- Is the data resource part of a time series? Is the data resource covering the same extent periodically updated to the same specification?
- Does the data resource relate to, or reference, a continuous area or contiguous areas, or does it reference specific locations?
- Does the data resource relate to one or many subjects, topics or themes?

An approach to resolving these questions is found at Table 1.

The key points are:

- Finding the appropriate level will be a compromise between what appears to be fit for purpose for the user and the granularity which can be supported by the metadata creator.
- If the metadata creator has to aggregate what is documented for practical reasons then they should ensure that the resource is adequately documented in terms of Extent, Dataset reference date and Topic Category. There should also an Additional Information Source included so that the service user can understand the finer granularity lying behind the single entry.

Table 1: Guidance on levels of data resources for individual documentation.

Nature of data resource	**How to document**	**Examples**
Stand-alone product or identifiable dataset or document and; a) not part of a series produced to the same specification; b) not part of a time series.	Individually with one metadataset. **Notes:** • If the data resource references a number of separate locations then at least ensure that this is reflected in a multiple entry for Extent. • If the data resource covers a number of topics or subjects then ensure that each of these is reflected in multiple entries for Topic Category and Keyword.	• Stand-alone soil map of the New Forest. • One-off report on a Site of Special Scientific Interest. • Historical index of streets in Cambridge. • One-off table of statistics for Oxfordshire.
Products or identifiable datasets or documents forming a series with a common specification; a) referencing one location or contiguous area; and b) may be part of a time series.	Together with one metadataset. **Notes:** • If the time series is not regular with periodic updates, consider documenting by individual dates or durations referring to particular periods of update. • If the data resource covers a number of topics or subjects then ensure that each of these is reflected in multiple entries for Topic Category and Keyword.	• Unrevised geological map series of all of Scotland. • Topographic map series covering all of GB updated periodically. • Series of statistical reports on the crime by ward in Surrey for 1973. • Street gazetteer of Hampshire.
Products or identifiable datasets or documents forming a series with a common specification; a) referencing more than one discrete location or contiguous area; and b) may be part of a time series.	Each location or area individually with a metadataset. **Notes:** • If the time series is not regular with periodic updates, consider documenting by individual dates or durations referring to a particular period of update. • If the data resource covers a number of topics or subjects then ensure that each of these is reflected in multiple entries for Topic Category and Keyword. • If it is not feasible to individually document each location or contiguous area, e.g. borehole records or Sites of Special Scientific Interest, then at least ensure that is reflected in a multiple entry for Extent. Also include Additional Information Source.	• Geological map series of different parts of Great Britain – document each part separately. • Topographic maps of National Parks – document each Park separately. • Aerial photography of parts of Devon, Cornwall and Somerset taken in 1957 and not re-flown to same specification – document each part separately. • Statistical tables for metropolitan areas in UK – document separately if feasible.

Annex A Glossary of Terms

acceptable quality level (AQL)

threshold value applied to the results of testing data quality to determine whether the data meets criteria determined from a standard, specification or user requirements

aggregated AQL

acceptable quality level for aggregated results from a number of tests, e.g. 100% correct

data resource

dataset, collection of datasets or service supplying data

dataset

identifiable collection of data

dataset series

collection of datasets sharing the same specification, e.g. a 1:1250 scale map series

granularity

resolution in terms of density or frequency

location

identifiable geographic place

metadata

data about data

metadatabase

collection of metadata about a set of data resources

metadata element

individual item of metadata relating to a data resource, e.g. Extent, Topic

metadata service

service that supplies information about data resources, e.g. *gigateway*

metadataset

identifiable set of metadata relating to a single data resource

quality assessment

review of quality of a data resource

quality assurance

process to ensure that quality is of an acceptable level

quality control

process of checking items to ensure that they are of an acceptable level of quality

quality management

overall process for assessing and controlling quality

quality result

value of a quality measure

universe of discourse

view of the real or hypothetical world that includes everything of interest

Bibliography

e-Government Metadata Standard. Cabinet Office e-Government Unit. (see http://www.cabinetoffice.gov.uk/govtalk)

Federal Geographic Data Committee "Content Standard for Digital Geospatial Metadata" Version 2 - 1998. (see www.fgdc.gov/standards/standards_publications)

INSPIRE Metadata Regulation 03.13.2008 (see http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32008R1205:EN:NOT)

INSPIRE Metadata Implementing Rules: Technical Guidelines based on EN ISO 19115 and EN ISO 19119 v1.2 2010-06-16 (see http://inspire.jrc.ec.europa.eu/documents/Metadata/INSPIRE_MD_IR_and_ISO_v1_2_20100616.pdf)

UK GEMINI2 Standard Version 2.1.AGI August 2010. (see www.gigateway.org.uk)

ISO 8601: 2004 Data elements and interchange formats – Information interchange – Representation of dates and times.

ISO 8402: 1994 Quality management and quality assurance - Vocabulary

ISO 15836: 2003 Dublin Core metadata element set

ISO 19115: 2003 Geographic information — Metadata