Common Metadata Errors

UK Location Discovery Metadata Service

Edition 1.0
February 2013

uk location

Document Control

Change Summary

Version	Date	Author/Editor	Change Summary
1-0	25/02/2013	Peter Parslow	First version of document.

References

Ref.	Author/Title/Version/Date of Publication
1	UK Location / Getting started – series of guides
2	UK Location / DMS Operational Guide / 2.2 / February 2013
3	AGI / Metadata Guidelines for Geospatial Data Resources, Introduction
4	ISO 19139 Geographic information – Metadata – XML schema implementation – published by ISO (and national standards bodies)
5	AGI / UK GEMINI / 2.2 / December 2012
6	ISO 19115 Geographic information – Metadata, published by ISO (and national standards bodies)
7	EC / INSPIRE Metadata Regulation; Commission Regulation (EC) No 1205/2008 of 3 December 2008
8	UK Location / UK GEMINI Encoding Guidance / Version 1.4 / February 2013

UK Location resources listed above can be found on data.gov.uk/location

INSPIRE resources can be found at: http://inspire.jrc.ec.europa.eu/

AGI GEMINI documents are at http://www.agi.org.uk/uk-gemini

Glossary

The following definitions apply in relation to this document:

Term	Definition
AGI	Association for Geographic Information
DMS	Discovery Metadata Service
INSPIRE	Infrastructure for Spatial Information in Europe
ISO	International Organisation for Standardisation
Schematron	A language for making assertions about patterns in XML documents; ISO 19757 Document Schema Definition Language
XML	eXtensible Markup Language
XML Document	A collection of data represented in XML.

Contents

Introduction

The UK Location Discovery Metadata Service (DMS) lies at the heart of UK Location and the delivery of the UK Location Strategy and INSPIRE - 'to know what data we have'.

The DMS operates by collecting valid metadata records from data publishers and making them available on data.gov.uk, and for further collection to the European INSPIRE Geoportal.

Target Audience

The primary audience for this document are those responsible for creating and maintaining metadata records, located within data provider organisations, and their technical partners.

Background

This document is largely based on an analysis of validation failures when metadata validation was first run on the records that had already been collected into data.gov.uk. Approximately half the records present in data.gov.uk at that time initially failed.

Assumed Knowledge

This guide assumes that the reader is familiar with the creation and management of metadata and has read the UK Location "Getting Started' series of guides [1], and the DMS Operational Guide.

Readers requiring an introduction to discovery metadata for geospatial data resources are referred to the UK GEMINI guide "Metadata Guidelines for Geospatial Data Resources, Introduction – Part 1" [3].

Other useful reading includes:

- UK GEMINI Encoding Guidance
- GEMINI Schematron Schema technical documentation
- GEMINI Schematron Error Descriptions
- AGI's Guidelines for UK GEMINI Part 2 Creating Metadata using UK GEMINI
- AGI's Guidelines for UK GEMINI Part 3 Metadata Quality

Where to Obtain More Information

The latest information, and additional resources, can be obtained by visiting the UK Location web site. Search:
http://data.gov.uk/search/apachesolr_search?filters=type%3Aresource

If you would like to contact the UK Location Coordination Unit, contact details are at http://data.gov.uk/location/contact

Validation

Validating a UK Location metadata record is a three stage process. First, a metadata set is validated against the ISO 19139 schemas. UK Location / data.gov.uk use the XSD files provided by EDEN; see *UK GEMINI Encoding Guidance* for details. If the metadata set proves to be schema valid, it can then be validated against the ISO 19139 Table A.1 Constraints Schematron schema. The Schematron schema relies on hardcoded XPath statements which will only work effectively on a schema valid XML set. Finally, if the XML is still valid it can be validated against the GEMINI2 Profile Schematron schema.

Any failure of this validation will result in the record being rejected by data.gov.uk

The European INSPIRE Geoportal carries out its own validation. This uses basically the same rules, and no records are rejected. However, there are some "soft errors" that are commonly reported by the geoportal.

Beyond that are "soft errors", where the metadata is valid and will be collected and made available, but may be misleading.

Schema errors

Confusion of Date and DateTime

In ISO 19115, Date and DateTime are distinct types. Although in many elements, either is allowed, the XML encoding needs to be explicit about which is given. It is an error to put a date (such as 2010-05-12) in a DateTime element.

Example of invalid structure:

```
<gmd:dateStamp>
      <gco:DateTime>2012-11-15</gco:DateTime>
</gmd:dateStamp>
```

This should either include the time, e.g.

```
<gmd:dateStamp>
      <gco:DateTime>2012-11-15T13:50:38</gco:DateTime>
</gmd:dateStamp>
```

Or be explicit that it doesn't:

```
<gmd:dateStamp>
      <gco:Date>2012-11-15</gco:Date>
</gmd:dateStamp>
```

Elements in the wrong order

Although there is no "logical" ordering of the elements in ISO 19115, INSPIRE, or GEMINI, the XML pattern adopted by ISO 19139 means that the elements must appear in the correct order. The order enforced is the order the elements appear in ISO 19115; the GEMINI Encoding Guidance contains an appendix giving the correct order for the subset of elements used by GEMINI.

The example found contained the ISO 19115 metadata characterSet element before the metadata language element.

ISO 19139 Schematron errors

See also "Empty strings".

No level description

ISO 19115 requires that a "level description" is given for any quality statement that is not describing the "dataset" or "series" level. INSPIRE and GEMINI use the quality statement for both lineage and conformity. This means that any "service" record must provide a gmd:DQ_Scope/gmd:levelDescription element, such as:

```
<gmd:DQ_Scope>
	<gmd:level>
		<gmd:MD_ScopeCode
codeList="http://standards.iso.org/ittf/PubliclyAvailableStandards/ISO_19139_Schemas/resources/codelist/gmxCodelists.xml#MD_ScopeCode"
codeListValue="service">service</gmd:MD_ScopeCode>
	</gmd:level>
	<gmd:levelDescription>
		<gmd:MD_ScopeDescription>
			<gmd:other>
				<gco:CharacterString>Geographic web
service</gco:CharacterString>
			</gmd:other>
		</gmd:MD_ScopeDescription>
	</gmd:levelDescription>
</gmd:DQ_Scope>
```

A similar rule applies to "hierarchyLevelName", which must be provided for any record that is not describing a dataset.

No Format provided

ISO 19139 requires that a format is given, either within the "distribution" section or the "distributor" section. INSPIRE & UK GEMINI encoding guidance only expects the distribution section, as the place to encode the Resource locator, which effectively renders distribution format mandatory.

UK Location profile schematron errors

Metadata language missing

Metadata language is optional in ISO 19115 and GEMINI, but mandatory for INSPIRE and therefore in the UK Location profile. This is now checked by the UK Location GEMINI schematron.

Processing errors

These errors do not result in invalid records on an individual basis, but may well result in the wrong records being available to the public.

Non-compliant Web Accessible Folder (WAF)

The collection definition is very precise. Just because you can browse to a page that appears to contain the records does not mean that the system can collect from them. For example, the page may just contain links to records that are elsewhere.

Similarly, collection differentiates between a single XML record at the end of a URL, and a WAF.

FileIdentifier

This is the identifier for a metadata record through time, so do ensure that it remains the same when you update a record, and is different in each record you create.

If you accidently change the identifier, the new record will not replace the old one.

If you accidently provide two records with the same identifier, one will replace the other.

Metadata date

This is used to distinguish which of two records with the same fileIdentifier the system will keep.

Even if the records are moved to a different server, if the fileIdentifier and metadata date are the same, the harvester will not collect the new files.

European INSPIRE Geoportal errors & warnings

Conformity statement missing

INSPIRE requires a 'conformity' statement, which can say that the resource conforms, does not conform, or has not been evaluated against, a specification. This is encoded with an ISO 19115 quality report. The INSPIRE encoding guidelines say that "not evaluated" should be encoded by leaving out the quality report. For this reason, GEMINI says that it is conditional, "required if claiming conformance to INSPIRE".

However, the INSPIRE Geoportal reports a validation issue if the element is missing: "The metadata element "Conformity" is missing, empty or incomplete but it is required". This can be ignored.

Missing "coupled resource"

INSPIRE requires "Coupled resource" to be populated "where relevant". This effectively makes it mandatory for View & Download services, and the INSPIRE Geoportal reports this as a validation issue. However, the Geoportal also reports this issue for a Discovery service metadata record, where coupled resource is not mandatory.

"Soft" errors

Empty strings

When first run, approximately half the rejections were down to this error. UK Location decided to remove this check, so the records can be harvested, although it is strongly recommended that this structure is avoided.

ISO 19139 clause A.2.1 states "a property element following the default XCPT pattern is designed to have content (by-value) or attributes (by-reference or NULL with reason)." In context, this states that (with very few exceptions) no metadata element shall be entirely empty: it will either have (valid) content, or it will have a gco:nilReason attribute stating why the content is absent. If the element is optional, it shall be entirely absent from the XML document, not "present but empty".

The most common examples were structures like:

```
<gmd:phone>
      <gmd:CI_Telephone/>
</gmd:phone>
```

(in this case, as gmd:phone is optional, it should be entirely missing)

and

```
<gmd:administrativeArea gco:nilReason="missing">
      <gco:CharacterString/>
</gmd:administrativeArea>
```

Examples of valid structures are:

```
<gmd:administrativeArea gco:nilReason="missing"/>
```

(although again, it would be equally valid to leave the element out entirely)

```
<gmd:deliveryPoint>
      <gco:CharacterString>Explorer House, Adanac Drive</gco:CharacterString>
</gmd:deliveryPoint>
```

Short relative URLs

The issues discussed here have just been noted whilst browsing the records in data.gov.uk

Quite a lot of records use the ISO 19115 element "browseGraphic", which could usefully provide a quick visualisation of the dataset. However, mostly these are populated with a local path, which is then impossible to use once the dataset is harvested to any other location – even if it did work in the data publisher's original location. For example:

```
<gmd:MD_BrowseGraphic>
      <gmd:fileName>
            <gco:CharacterString>Council_Housing_s.png</gco:CharacterString>
      </gmd:fileName>
      …
</gmd:MD_BrowseGraphic>
```

Incorrect code list URL

The INSPIRE encoding guidance, and therefore earlier versions of the GEMINI Encoding Guidance had a minor spelling (capitalisation) mistake in the path for the code list dictionary. This is used a number of times in most records.

```
<gmd:CI_OnLineFunctionCode
codeList="http://standards.iso.org/ittf/PubliclyAvailableStandards/ISO_19139_Schemas/resources/Codelist/gmxCodelists.xml#CI_OnLineFunctionCode"
codeListValue="information">information</gmd:CI_OnLineFunctionCode>
```

Should be

```
<gmd:CI_OnLineFunctionCode
codeList="http://standards.iso.org/ittf/PubliclyAvailableStandards/ISO_19139_Schemas/resources/codelist/gmxCodelists.xml#CI_OnLineFunctionCode"
codeListValue="information">information</gmd:CI_OnLineFunctionCode>
```

(a lower case ‘c’ in the first occurrence of ‘codelist’ within the URL)

Overly precise bounding box

GEMINI states that the bounding latitude and longitude should be stated to two decimal places. Last year, the INSPIRE geoportal reported this as an error; it does not now. For example:

```
<gmd:westBoundLongitude>
      <gco:Decimal>-8.17548890898914</gco:Decimal>
</gmd:westBoundLongitude>
```

Six decimal places of a degree is something like 1cm, so this longitude is being given to the nearest molecule.

Incomplete Web service URLs

For example, in a Resource Locator, a WMS service that does not end in an ampersand is incorrect, for example:

```
<gmd:linkage>
      <gmd:URL>http://maps.bgs.ac.uk/ArcGIS/services/BGS_Detailed_Geology/MapServer/WMSServer</gmd:URL>
</gmd:linkage>
```

Should be:

```
<gmd:linkage>
      <gmd:URL>http://maps.bgs.ac.uk/ArcGIS/services/BGS_Detailed_Geology/MapServer/WMSServer?service=WMS&request=GetCapabilities&</gmd:URL>
</gmd:linkage>
```

Insufficient INSPIRE keywords

INSPIRE requires service metadata records to include a keyword from part D.4 of the INSPIRE Metadata Regulation , for example:

```
<gmd:keyword>
      <gco:CharacterString>infoMapAccessService</gco:CharacterString>
</gmd:keyword>
```

Incorrect INSPIRE keywords

Putting in more than one INSPIRE keyword, where the first keyword in the list is not the correct Annex/Theme

Missing or incorrect constraints information

Licence information should be provided, preferably by a URL.

It is an error to

- State "no limitations" in Limitations on Public Access, when the resource is not open access.
- State "no conditions apply" in Use Constraints, when restrictions do in fact exist.

And the winner?

This table, for interest, gives the approximate error rate for each of the above errors, in the sample tested in late 2012. This does not include those soft errors only found by inspection.

Error	Proportion
Empty strings	50%
No level description	25%
No format	24%
Four other errors, one occurrence each	