Discovery Metadata Service Collection Interface Specification

Version 0-8
March 2013

uk location

Document Control

Change Summary

Version	Date	Author/Editor	Change Summary
0-1	10/11/2010	Tim Manning	Initial draft version, based on internal UKLP design position paper.
0-2	19/11/2010	Peter Parslow/Tim Manning	Revised draft following internal technical review. Addition of WAF interface specification from John Bywater.
0-3	23/11/2010	Tim Manning	Minor changes only, following the creation of draft Edition 2 of DMS Operational Guide.
0-4	04/02/2011	Rod Kedge	Changes incorporating comments from CoI following development of harvesting functionality
0-8	13/03/2013	Ian James / David Read	Updated to clarify harvesting mechanisms for WAF, and include a single GEMINI2 record option. Other minor changes, clarifications and corrections, including creation of the "Processing the records" section, with an elaboration on validation process.

References

Ref.	Author/Title/Version/Date of Publication
[1]	UK Location, "Getting Started Guides"
[2]	UK Location, "Discovery Metadata Service Operational Guide"
[3]	Open GIS® Catalogue Services Specification OGC 07-006r1 / 2.0.2 with second corrigendum / 2007-02-23 / Ed Douglas Nebert et al
[4]	UK Location, "UK GEMINI Encoding Guidance"
[5]	UK Location, "UK GEMINI Schematron Schema Constraints"
[6]	UK Location, "UK GEMINI Schematron Schema Guidance"

Glossary

The following definitions apply in relation to this document:

Term	Definition
CSW	See OGC CSW
Data.gov.uk	The UK Government data portal, used by UK Location to publish location metadata, and by Data Publishers to register their published data and services.
Discovery Metadata	Information about a data or service resource, used to discover and access its suitability for sharing or re-use.
INSPIRE	Infrastructure for Spatial Information in Europe
OGC CSW	Open Geospatial Consortium's Catalogue Services for the Web. The HTTP protocol binding for OpenGIS® Catalogue Services.
UK Location (UKL)	Brand name and abbreviation for the UK Location Information Infrastructure.
UKLP	UK Location Programme. The cross-government business change programme for the design, build and deployment of the central components of the UK Location Information Infrastructure and the coordination of data publishing by data providers.
URL	Uniform Resource Locator.
WAF	A Web Accessible Folder (WAF) is an HTTP accessible directory of files, typically metadata files in XML format, in which all files and their time-stamps are visible to a web browser or client. Crawlers are able to parse the file listings and date-time stamps and provide a search interface on these documents.
XML	eXtensible Markup Language
XML Document	A collection of data represented in XML.

Contents

Introduction

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

2 The Discovery Metadata Service underpins the coordinated and regulated publishing of public sector location information to the INSPIRE standards and UK Location application profiles. It allows data users to evaluate and use public sector location information using on-line services - to view, download and invoke as part of an end business application.

3 This specification defines the mechanism by which discovery metadata resources will be collected, following their registration with UK Location.

4 This specification is intended to support engagement with UK Location Data Providers, Publishers and their suppliers in establishing their initial operating capability.

5 The XML samples in this document have not been validated, and are provided for illustration purposes only.

6 Please note that whilst the Devolved Administrations of Scotland, Wales and Northern Ireland are part of UK Location, they may have different publishing mechanisms to those described here. Therefore in the first instance, please use the appropriate contact information listed under "Where to obtain more information" below.

Target Audience

7 The primary audience for this specification are those responsible for establishing a data publishing capability within Data Provider and Publisher organisations, and solution suppliers to these organisations.

Assumed Knowledge

8 This document assumes that the reader is familiar with the UK Location "Getting Started" series of guides [1]. This specification should also be read in conjunction with the DMS Operational Guide [2].

DMS Resources

9 The latest versions of all the UK Location resources referred to in this guide can be found via the UK Location Resource Centre:

http://location.defra.gov.uk/resources/discovery-metadata-service

Where to Obtain More Information

10 The latest information, and additional resources, can be obtained by visiting the UK Location web site.

11 If you would like to contact the UK Location Helpdesk please use the contact form at: http://location.defra.gov.uk/resources/contact-us/

12 If you are looking to publish location information specific to Scotland, Wales or Northern Ireland, please contact them as detailed at http://data.gov.uk/location/contact_points.

Publishing Discovery Metadata

13 Publishing location information into the UK Location Information Infrastructure is achieved by creating and publishing discovery metadata resources. These describe the data and the associated on-line services through which the data is published. These discovery metadata resources make the data and on-line services discoverable.

14 The publishing of these resources into UK Location is a two stage process:

1. **Publish discovery metadata resources** to a master repository, from which the resources can be machine-accessed from the Internet
2. **Register the published discovery metadata resources** with data.gov.uk, such that they can be collected and incorporated into the data.gov.uk discovery metadata catalogue and subsequently used as part of data.gov.uk and UK Location discovery services.

15 Discovery metadata resources can be collected from one of three publishing mechanisms:

- OGC Catalogue Service for the Web (CSW); *or*
- Web Accessible Folder (WAF); *or*
- Single GEMINI2 document

16 This document specifies how the UK Location central registration client application, hosted on data.gov.uk will interact with these interfaces to collect discovery metadata resources from the registered source.

17 This interface specification does not include any mechanism to withdraw a metadata resource from the DMS. Details of the process for withdrawing metadata records are detailed in the DMS Operational Guide [2].

Registering the Collection Source

18 The Data Publisher will enter resource details on the relevant page of data.gov.uk. The following details will be required:

- URL for source of metadata
    - For a CSW, this will be the root URL of the OGC CSW, e.g. http://www.someserver.com/csw/csw.cgi or .../geonetwork/srv/en/csw.
    - For a WAF, this will be the URL of an HTML document which contains links to one or more Gemini XML metadata documents (e.g. http://www.someserver.com/waf/index.html). The referenced metadata documents MUST be directly contained in the same folder as the HTML document, and the links to the metadata document MUST be relative (e.g. "metadatarecord.xml") and not full paths.
    - For a Single GEMINI2 document, this will be the URL of a single Gemini XML metadata document (e.g. http://www.someserver.com/waf/metadatarecord.xml)
- Source type - corresponding to the type of source for which the URL was provided - either CSW, WAF or Single GEMINI 2 Document.
- Publisher - the harvested data will be filed under this organization, which corresponds to the 'Provider' in INSPIRE terminology. The list of options in the form includes only those publishers for which the logged-in user has 'editor' or 'admin' privilege.
- Registration Description – an optional free text field for recording any relevant details about the registration. This is to assist Data Publishers to manage their Registrations on data.gov.uk.

This is equivalent to the OGC CSW Harvest operation and response. The operation will be asynchronous, in OGC CSW terms – the DMS does not start to collect records immediately after registration.

19 Sources will be harvested in response to a 'Refresh' request by the user. Harvest requests are batched by data.gov.uk, and processed at regular intervals.

Collecting from OGC CSW

20 Data.gov.uk will act as a client interfacing to OGC Catalogue Servers supporting CSW 2.0.2 [3]. It does not support earlier OGC CSW versions such as 2.0.1.

21 When harvesting from a CSW data.gov.uk implicitly assumes the following:

- Resource type – fixed as http://www.isotc211.org/schemas/2005/gmd/, the namespace URI identifying ISO 19139 encoded records. Beyond this, the interface requires these to be GEMINI2 records
- Resource format – fixed as application/xml

22 The collection interface may attempt GetCapabilities requests at any time after registration. It will use OGC CSW GetRecordByID operations to collect discovery metadata resources, after an initial GetRecords to retrieve the relevant IDs. The DMS will not use or support any other OGC CSW operations.

23 Amongst the various implementation options described within the OGC Catalogue Services Specification, the DMS will use HTTP GET and/or POST with XML payload.

24 After the initial GetRecords to retrieve identifiers, the DMS will request and expect all the values in the record, that is, we will not constrain by element name.

25 The ‘payload’ to be transferred in response to the GetRecordByID requests will be ISO 19139 XML encoded UK GEMINI records.

Collection Process

26 Collecting discovery metadata resources from an OGC CSW interface will involve the following steps.

GetRecords operation: Retrieving the identifiers

27 The collection interface will issue a GetRecords, to establish the size of the job, and collect the resource identifiers. Note that these will be the target catalogue's internal identifiers, not necessarily the fileIdentifier of the metadata Resource itself.

28 This will have the parameters defined in table 1.

Element / attribute	Value	Notes
service	"CSW"	Fixed
version	"2.0.2"	
REQUEST	"GetRecords"	Fixed
resultType	"results"	
ElementName	"dc:identifier"	The IDs do not need to be encoded in GEMINI/ISO 19139

Table 1: GetRecords Parameters

29 All other (optional) GetRecords elements and attributes will be left out, as the interface will work against the defaults.

GetRecordByID: Collecting the Resources

30 Following the Getrecords operation, the collection interface will request the discovery metadata resources by ID.

Element / attribute	Value	Notes
service	"CSW"	Fixed
version	"2.0.2"	
REQUEST	"GetRecordById"	
outputSchema	"http://www.isotc211.org/schemas/2005/gmd/"	In order to get ISO 19139 encoded records
Id	*comma separated list of IDs*	Precisely the strings returned by that catalogue server.

Table 2: GetrecordsByID Attributes

Other HTTP headers

31 All other http headers are expected to have default values.

Collecting from a WAF

32 A Web Accessible Folder (WAF) is an HTTP accessible directory of files, typically metadata files in XML format, in which all files are visible to a web browser or client, in this case data.gov.uk. The DMS metadata collection interface will use the specified URL of the file server to collect discovery metadata resources published by the Data Publisher.

Collection Process

33 An HTTP GET request will be sent using the registered locator. An error will be raised if there is no response.

34 The returned HTTP response body is parsed as an HTML document.

35 Links are extracted from the HTML document using XPath "//a/@href". Extracted links will be ignored if they contain characters '/', '?', '#' or the string 'mailto:'. That is, only relative links that refer to directly contained files will be used.

36 Locators of contained metadata resources are derived by appending extracted links to the registered WAF location (if necessary, a trailing slash will be appended to the registered WAF location). An error will be raised if zero metadata locators are derived from the WAF locator.

37 An HTTP GET request will be sent for each of the derived locators.

38 The returned HTTP response body is expected to be a metadata resource (XML document). An error will be raised if it is malformed XML.

39 If the response body does look like a GEMINI XML document [4], it will be taken forward for validation and inclusion in the central catalogue.

40 If the response body does not look like a GEMINI XML document [4], it will be ignored.

Collecting from a single GEMINI2 record

41 Collecting discovery metadata resources from a single GEMINI2 document will involve the following steps.

Collection Process

42 An HTTP GET request will be sent using the registered locator. An error will be raised if there is no response.

43 The returned HTTP response body is expected to be a metadata resource (XML document). An error will be raised if it is malformed XML.

44 If the response body does look like a GEMINI XML document [4], it will be taken forward for validation and inclusion in the central catalogue.

45 If the response body does not look like a GEMINI XML document [4], an error will be raised.

Processing the records

Initial checks

46 After receiving the discovery metadata resources, some initial checks are done on the metadata.

- If the Metadata date in any record is not newer than a corresponding existing record in data.gov.uk then harvesting of that record will be discontinued.
- If two records with the same File Identifier are received, the one with the more recent metadata date will be retained.

Validation checks

47 Individual records will then be validated against a set of business rules. This validation is to ensure compliant with the INSPIRE Regulations and conformance with INSPIRE and UK Location guidance.

48 Discovery metadata records will be:

- cross-checked against the associated Data Provider details held on data.gov.uk[1]
- validated against the ISO 19139 schemas - UK Location / data.gov.uk use the XSD files provided by EDEN, see [4]
- 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
- validated against the GEMINI2 Profile Schematron schema [5] [6]

Adding to the metadata catalogue

49 Each resource that passes all these checks (initial and validation) will then be added to the data.gov.uk metadata catalogue. If the resource has passed all these checks and has a fileIdentifier that matches an existing entry, it will replace that entry.

50 If a metadata record is harvested with the same file identifier as an existing withdrawn entry, but with a more recent metadata date, then that newly harvested record will be re-instated at data.gov.uk.

[1] Note that on data.gov.uk Data Provider details are held within a registry labelled "Data Publisher", i.e. the Data Provider is referred to as the "Data Publisher", both when publishing directly and through a third party Data Publisher.

51 The discovery metadata resource XML document itself will be stored for use within the data.gov.uk client application and for subsequent publishing through the Discovery Metadata Service Catalogue Publishing service.

Collection Interface Error handling

52 Errors will be reported in the Harvest Dashboard, visible to the Data Publisher when logged into data.gov.uk. This will include:

- HTTP time outs, that may suggest the CSW is 'down' or the Web Accessible Folder cannot be accessed
- CSW error responses to GetRecords requests
- Discovery metadata resources not associated with the Data Provider identified as part of the registration
- Invalid XML documents, i.e. fail XSD and Schematron validation

53 If a validation step results in a failure then any subsequent XML Validation steps are not carried out.

54 Once a discovery metadata resource has been registered, the resource should never be removed from the source location. UK Location will monitor 'aged records', i.e. discovery metadata resources that have not been updated in-line with the "Frequency of Update" value for the dataset or series.