Annex A

Note: This annex, kindly provided by CEDA, describes current portal back end systems created and supported by the Centre for Environmental Data Analysis (CEDA) and the current use of GeoNetwork within this arrangement (refer to boxed text)

MEDIN operations at the Centre for Environmental Data Analysis

Overview

CEDA operate the MEDIN Metadata Service that underpins the MEDIN portal. This service is the link between the portal and the harvested metadata provided by the MEDIN metadata providers. There are three components that are required for the successful operation of this service:

- Metadata harvest (csw1.cems.rl.ac.uk)
- Metadata catalogue and ingest procedure (csw1.cems.rl.ac.uk)
- Discovery Web Service (DWS) (ceda-discovery.ceda.ac.uk)

The hostnames in brackets denote the CEDA VM on which that component service is run.

The MEDIN portal interacts with the metadata catalogue comprised of harvested metadata via the DWS service. The DWS provides varying level of interaction and detail depending on the operation requested.

In addition to the metadata harvest, catalogue and DWS operation CEDA provides an OGC Catalogue Service for the Web 2.0.2 (CSW) endpoints by which MEDIN records can be queried. This enables the MEDIN metadata collection to publish to clients as the UK Location Program portal (data.gov). The CSW also provides an interface by which MEDIN can manage the harvest of metadata from the various providers.

The procedure of the MEDIN system run by CEDA can be found at
https://csw-medin.ceda.ac.uk/medin/MEDIN_metadata_processes.pdf

The separate components of the CEDA metadata operations are described in further detail below:

Metadata Harvest

Records are harvested from official MEDIN metadata providers by any of three methods: OAI-PMH, CSW and WAF.

- CEDA maintains an OAI-PMH (Open Archives Protocol for Metadata Harvesting) for the few providers who need to publish their metadata this way. These are harvested into a local file system where they are ingested into the MEDIN catalogue database.
- CSW (Catalogue Service for the Web): CEDA uses a GeoNetwork 2.10.2 instance to manage the harvesting of records via the CSW 2.0.2 specification. A number of providers use this method. Records are routinely harvested from the provider CSW into the dedicated CEDA MEDIN GeoNetwork. Local scripts are then used to retrieve these from the MEDIN GeoNetwork into the local file system, from where they are ingested into the MEDIN catalogue database.

- WAF (Web Accessible Folder): CEDA use the GeoNetwork WEB-DAV facility for harvesting of records into the MEDIN CSW. Such records then follow a similar path into the MEDIN catalogue database as records harvested from CSW providers. However, not all providers WAF's are WEB-DAV (a WAF specification) conformant. Such records are harvested into the local filesystem using dedicated scripts. Certain providers maintain an XML document naming convention based on the resource title. Further scripts rename these according to the metadata UID as well as removing duplicate records. This is required for a successful harvest into the MEDIN catalogue database.

All records harvested from the various MEDIN providers are inserted into the MEDIN GeoNetwork instance at: https://csw-medin.ceda.ac.uk/geonetwork/srv/eng/catalog.search#/home The GeoNetwork instance provides useful management functionality for all MEDIN metadata records and is the source from which ultimately all records are placed into the MEDIN catalogue database.

An important point to note is that this GeoNetwork CSW is used to both harvest records (by CSW, OAI, WEB-DAV etc) *and* provide various CSW endpoints by which the internal catalogue can be queried by.

This GeoNetwork instance has been configured for access by multiple administrative users. GeoNetwork categories and groups have been created to help utilise and partition metadata from the various providers.

The GeoNetwork CSW is configured to offer the standard CSW interface that exposes all records within the catalogue. This CSW endpoint can be found at:

https://csw-medin.ceda.ac.uk/geonetwork/srv/eng/csw?SERVICE=CSW&VERSION=2.0.2&REQUEST=GetCapabilities

A further CSW endpoint has been configured using the GeoNetwork virtual-CSW feature to provide a catalogue of a subset of the records based on the presence of the NDGO0005 keyword. This keyword is found only in records that the original provider wished to be made available to the UK Location Program (data.gov). The following endpoint is used by data.gov to harvest all MEDIN INSPIRE compliant records:

https://csw-medin.ceda.ac.uk/geonetwork/srv/eng/csw-MEDIN?SERVICE=CSW&VERSION=2.0.2&REQUEST=GetCapabilities

Other CSW endpoints have been configured for each MEDIN metadata provider. These are based on the GeoNetwork categories, that have been created for each provider. Each endpoint is based on the harvesting name i.e.:

https://csw-medin.ceda.ac.uk/geonetwork/srv/eng/csw-MEDIN-UKHO?SERVICE=CSW&VERSION=2.0.2&REQUEST=GetCapabilities

MEDIN catalogue database

The MEDIN catalogue holds all metadata records that have been harvested and successfully ingested into the database. Not all records that have been harvested make it into the database due to either configuration issues, formatting errors or xml content issues. Therefore the catalogue is not a direct

content mirror of the MEDIN GeoNetwork catalogue. Reports are created for problem records and placed in the reporting portal (see http://csw1.cems.rl.ac.uk/medin/MEDIN_metadata_processes.pdf).

Records are extracted for each MEDIN provider from the MEDIN GeoNetwork CSW and placed in a local harvest directory. From this location a daily catalogue ingest job is run to place them in the database. New records are added to the database and existing database entries are compared against the recently harvested equivalent records. If any changes are detected the record is updated in the catalogue. If any records are detected in the database with no harvested equivalent, then that record is removed from the database. Thus, the database strives to be an accurate mirror of records found on the provider's metadata publication points. This catalogue ingest procedure is performed automatically daily. Manual runs can be requested.

Records within the database are identified by the metadata UID and qualified by the provider's namespace. The database is comprised of a number of tables. The primary table contains pertinent information used by the DWS to perform queries requested by the MEDIN portal. Other tables contain spatio-temporal information as well as copies of the metadata record in original and converted formats.

The database is a Postgres 8.4.2 installation with a PostGIS 1.3.6-1 enabled database running on a dedicated CEDA database server, and is routinely backed up and updated according to CEDA/STFC JASMIN/CEMS official procedures. Database connections are limited to the VM's that provide the harvesting and ingest functionality and that which supports the DWS.

The Discovery Web Service (DWS)

The DWS is a SOAP (Simple Object Access Protocol) enabled service that receives structured XML requests to access service and catalogue database information. The DWS interprets these requests and formats an appropriate response depending on the nature of the request and returns a SOAP response in a structured format.

In the MEDIN DWS context, the MEDIN portal constructs a request from information submitted by a portal user and submits this to the DWS. The DWS parses this request and interacts with the catalogue database to produce a response. This response is parsed and presented by the portal to the user. The DWS is located at CEDA on the VM ceda-discovery.ceda.ac.uk and is IP locked for security issues to the client MEDIN portal run at BODC. All portal operations and associated issues are out of context of this document.

The DWS operates within a JBOSS 5.10 servlet container on ceda-discovery.ceda.ac.uk port 8280. This VM is a Red Hat Enterprise Linux Server release 6.8 installation and operates within the CEDA STFC JASMIN/CEMS VM cloud. This server is dedicated to the operation of the MEDIN DWS and as such, is configured to run JBOSS solely on port 8280.

The DWS endpoint is http://ceda-discovery.ceda.ac.uk:8280/discovery/dws?wsdl and is restricted by IP address. The contractor will be given access once contract has been signed.