Metadata for Data Discovery: The NERC Data Catalogue Service

Steve Donegan
Introduction

- NERC, Science and Data Centres
- NERC Discovery Metadata
- The Data Catalogue Service
- NERC Data Services
- Case study: generating Metadata and doing something useful with it!
- Main UK body for funding research, training, knowledge exchange in environmental sciences
- Annual budget £388m (2011)
- Covers atmosphere, earth, terrestrial, aquatic sciences
- Research ships and aircraft, satellite technology
What sort of data do we deal with?

A variety of environmental measurements, along with the results of model simulations.
NERC Designated Data Centres

The UK’s Natural Environment Research Council (NERC) funds eight data centres which between them have responsibility for the long-term management of NERC's environmental data holdings.
The role of the data centres

- NERC funds research projects, which produce data.

- It is essential that these data are properly managed to ensure their long-term availability.

- NERC’s network of data centres provide support and guidance in data management to those funded by NERC, are responsible for the long-term curation of data and provide access to NERC's data holdings.

- The NERC Data Policy details their commitment to support the long-term management of data and also outlines the roles and responsibilities of all those involved in the collection and management of data.

We are also involved in externally funded projects in informatics, e-Science and domain specific areas.
Changing and conflicting user demands

There is a tension between the requirements of different users.

- Scientists / NERC
  - Want raw data in its original format
  - Require long-term stewardship of data
  - Want as much contextual detail as possible

- Government Agencies / Knowledge Exchange:
  - Use environmental information to drive policy making
  - Prefer real time data delivery
  - Require derived products that address specific questions
  - Need to synthesise data from many different sources in order to reach a decision

Quality control is critical!
Legislation and technical changes

- Open standards for geospatial data and services promises a new level of interoperability between data providers
- EU INSPIRE directive requires us to provide data discovery, view and download services
- INSPIRE is an Infrastructure for Spatial Information within Europe for the purposes of Community environmental policies and policies or activities which may have an impact on the environment.
- As NERC data is within the UK public domain and many of its data holdings have a geospatial component, then by law NERC must produce metadata that is compliant with the EU INSPIRE directive (http://inspire.jrc.ec.europa.eu).
- Data interoperability and data sharing are prime objectives for INSPIRE and these are underpinned by a specification for metadata used for Data Discovery within INSPIRE. INSPIRE discovery metadata is based on the ISO19115/19119 Application Profile (metadata for geographic information) with a definition of core metadata elements from this required for INSPIRE compliance.
Discovery Metadata to satisfy all requirements

- NERC requires research/data to be able to generate on demand consistent discovery metadata describing NERC’s data assets.

- Compliance to this standard helps to ensure that NERC’s data assets are consistently discoverable, and aids in the generation and operation of services that utilise these assets across the NERC disciplines.

- NERC metadata must also accommodate and comply with international standards and directives.

- Metadata providers must have the capability to produce metadata conforming to this standard, under
Discovery Metadata to satisfy all requirements

- NERC Data Management Advisory Group (DMAG)
  - The ISO standards 19115 and 19119 define metadata schema definitions adequate for describing data resources held by NERC. For communication purposes, the ISO19115/19119 metadata can be serialised and encoded as XML using the ISO standard 19139.
  - NERC produced a profile of the ISO19115
  - But before official adoption...

- NERC SIS Group review MEDIN Discovery Metadata Standard:
  - MEDIN: Marine Environment Data Information Network: Some NERC DDC’s MEDIN partners
  - MEDIN largely conformant with INSPIRE and Gemini2 but with specialism's for the marine community (i.e. Seadatanet keywords
  - Decided to base NERC “standard” on MEDIN Discovery Standard but with exceptions/additions for NERC specific areas (i.e. How do you define vertical extent for Butterfly counts?)
  - MEDIN community has published schematron and metadata tools to support standard
  - Datasets, Series and Services!

- Adopt straight UK Gemini?
The NERC Data Catalogue Service

- The NERC DCS aims to provide a searchable interface to “published” discovery records from NERC DDC’s

- Provides the ability to conduct a simple text, geographic and/or temporal search.

- Advanced search option allows structuring of complex queries: search for the term “ozone” but NOT if associated with the term “depletion”

- Results returned with basic information rendered from the discovery metadata – links back to DDC, further information, download service etc
  - Currently datasets & series..

- Uses NERC Vocabulary Service for added content/dissemination
Data Services: Services need discovering too!
Developing a Portal

- NERC needed to replace the previous NERC Discovery Service: limited by metadata content (GCMD DIF, interoperability issues – services etc)
  - Developed as part of the NERC Data Grid (NDG) activity – consisted of a portal connected to a metadata catalogue all located at NEODC/CEDA
- NERC SIS recommended not only adopting the MEDIN Discovery Standard but also using the existing MEDIN Discovery Portal and web service
- MEDIN portal uses the Discovery Web Service (DWS) developed by NEODC/CEDA to search a metadata catalogue derived from discovery metadata harvested from data providers
  - Based on previous generation NERC Discovery Portal but adapted for ISO19139 rather than GCMD DIF
  - More powerful “targeted” keyword and text searches
  - Distributed architecture: DWS runs on catalogue at NEODC/CEDA whilst portal located at Geodata in Southampton
- NERC Data Catalogue Service adapted for NERC style MEDIN records but with added targeted text search etc
  - DWS/Catalogue runs at NEODC/CEDA and DCS portal at BODC
Developing a Portal (NERC model)

Metadata Catalogue (PostgreSQL)

Discovery Web Service (DWS)

Data Providers Web Service (DPWS)

OAI-PMH

OGC CSW

WAF

Metadata Providers
Harvesting the metadata.

OAI-PMH (Open Archive Initiative: Protocol for Metadata Harvesting):

• Providers and Harvesters

• A harvester takes full XML metadata and returns a copy to the local environment

• Any format – however, Dublin Core must be provided to be OAI-PMH compliant

• Support for deleted records, detection of changed records, regular harvesting

• Works via HTTP
Developing a Portal (MEDIN model)
Portal future developments

- UK Location is implementing the UK’s response for INSPIRE
- All in-scope records must be published to the UK Location Portal
- OGC CSW (Catalog Service for the Web) or via WAF (Web Accessible Folder)
- CSW: The Catalog Service defines common interfaces to discover, browse, and query metadata about data, services, and other potential resources.
  - Opensource solution: Geonetworks
- MEDIN solution for compliance is to run a parallel CSW to the MEDIN DWS with identical content
- NERC solution is for all DDC’s to replace OAI-PMH with local Geonetworks with one “core” CSW that supports a Discovery portal using a federated search. The core CSW is also the publishing point to UK Location
Developing a Portal (future NERC model)

Metadata Providers

- OGC CSW

Metadata Providers

- OGC CSW

Federated Searches

OGC CSW (Geonetworks)

Metadata Providers

- OGC CSW

Metadata Providers

- OGC CSW
CEDA Case Study

• CEDA: Centre for Environmental Data Archival:
  • NERC Earth Observation Data Centre (NEODC)
  • British Atmospheric Data Centre (BADC)
  • UK Solar System Data Centre (UKSSDC)

• Located at STFC Rutherford Appleton Laboratory, Oxfordshire

• Actively participates in NERC e-infrastructure projects:
  • NERC Data grid
  • INSPIRE LMO, OGC, ISIC, and much much more.

• Data Centres publish data to NERC DCS but also runs the Harvesting, catalogue and DWS operations supporting the portal (BODC)

• But how does a data centre generate metadata and get it published?
CEDA Metadata Catalogue

• All of CEDA data holdings are catalogued in a database according to a data model (MOLES2/3).

• This model quantifies various aspects of the data:
  • What is it? (i.e. instrument, format, model, service)
  • Where and when is it? (i.e. spatial coverage, date range/times)
  • Who owns it/where did it come from? (i.e. Who created the dataset? Restrictions on usage – UK only?)
  • What can it do? (i.e. Is it available in a visualisation service? Any legal aspects?)
  • Any associated resources? (i.e. Keyword or Parameter names, Links to original data provider site, documentation, Web Service Endpoints)
CEDA Metadata Catalogue

Information in the data catalogue is created by a combination of manual entry by Data Scientists as well as information taken from the data itself during the ingestion process and placement on the CEDA archive.

Metadata in the catalogue is used for a variety of purposes:

- Provide a resource to generate metadata for external consumption i.e. to aid data “discovery”, allow data/CEDA services to be used in external resources (i.e. WFS, WMS etc)
- Provide an accurate up to date description of each dataset and any related issues as a resource for the community
- Reference – allow citation of dataset (DOI)
- Dataset management
Data Suppliers

3rd Party Data providers

Catalogue
- MOLES
- CEDA Info

Automatic Update

Discovery XML
DataCite XML
CSML/WMS/WFS Service metadata

XML Generation

OAI PMH
OGC CSW
Web Accessible Folder (TBC)

Publicly Visible

External Users
NERC Catalogue Service, DataCite, UK Location Portal, Go-Geo, MEDIN Portal, INSPIRE.... All use metadata from CEDA metadata publishing layer