MOLES3: Implementing an ISO standards driven data catalogue
(It’s all about context & structure)

IDCC2015, 10th February 2015

Graham Parton¹, Steve Donegan¹, Bryan Lawrence², Stephen Pascoe¹, Ag Stephens¹, Spiros Ventouras¹

1 - Centre for Environmental Archival, STFC Rutherford Appleton Laboratory, U.K.
2 - NCAS, Department of Meteorology, University of Reading, U.K.
What’s in the talk...

Why we implement an ISO standards catalogue?

Lessons learnt getting to this catalogue

... but first,

a little bit of context...
Centre for Environmental Data Archival
20 years of organic growth

- 4 environmental data centres
- >168 million unique files online + physical archives
- > 2Pb online data
- > 3000 “datasets”
- In 300+ collections
Familiar problems and common approaches

- How do we open up these vast, differing archives?  
  **Discovery**

- How will users find, compare, select and use data?  
  **Context**

- Can users trust the source?  
  **Providence**

- Can they reliably reference the data?  
  **Persistence**

Underpinned by metadata
dataset metadata

definitions

A: archive. Usage metadata generated from (or about the data). Normally generated from internal metadata

B: browse. Context, generic, semantic, including a summary of A-type, links to or embeds discipline specific (E-type)

C: character and citation. Post-fact annotations and citations both internal and external (trackback, comments)

D: discovery. Metadata suitable for harvesting into catalogues and federations. Dublin-Core, NASA DIF, ISO 19115/19139

E: extra. Discipline-specific metadata, may or may not be understood at all sites, e.g. SensorML, NumSim

relationships

XML

S: security

XML

O: ontology

RDF

Q: defined and supported textual, semantic and spatio-temporal queries

QUERY

Lawrence, Lowry, Miller, Snaith and Woolf: Information in environmental data grids, Phil. Trans. R. Soc. A (2009)
Data context + browse functionality

- Allows discrimination between datasets
- Connected content via shared records
MOLES2

- Streamlined – 5 classes
- Provided Discovery function
- Provided additional context to data (Browse)
- Reusable, common objects
MOLES2 – Structured Reusability

DE 1
Deployment 1

DE 2
Deployment 2

DE 3
Deployment 3

Project a

Project b

Computer

Aircraft

Where

Why

model

Temp sensor

Pressure sensor

GPS

How

How

How

How
... but MOLES2 had problems

- Key attributes not reusable (e.g. names) = duplicates, inconsistent
- Lack of constraints = use was subverted
- Over-use of free-text fields
- Lacked ISO compliant fields (needed for EU INSPIRE)
- Couldn’t export to downstream services
- Couldn’t support DOI landing pages (granularity & ISO issue)
- System was unmaintainable
Evolving the MOLES Concept

Documentation describes Content model for MOLES concept

Data

A D B


MOLES + ISO 19156 = MOLES3.4

Documentation describes

Data

A

B

Content model for

Observations & Measurements (UML)

Specialises

MOLES 3.4 (UML)

Aligned with

MOLES2
Implementing MOLES3.4
MOLES3.4 + extra= CEDA-MOLES
Implementing CEDA-MOLES

CEDA-MOLES (NewMoon)

>680 tables
500,000 each
Bespoke system

Simplified CEDA-MOLES
29 tables

CEDA-MOLES (Django)

Migration

MOLES2 catalogue

MOLES 3.4 UML

2009-2010
2011
2012
2013
2014

Centre for Environmental Data Archival
National Centre for Atmospheric Science
National Centre for Earth Observation
Science & Technology Facilities Council
Rutherford Appleton Laboratory
Populating the Django database

Q: Construct afresh vs migrate from MOLES2

Migration necessary as:

• Archive metadata of insufficient quality/lack of tools
• MOLES2 ~6000 records = many years of effort to reproduce
• MOLES2 was unique record for some content + connections
• Need to preserve existing, already cited content
The Migration System

Data Entity

Deployment

Activity

Direct

Observation Collection

Observation

Project

Dataset Collection

Dataset

Instrument

Platform

Operation

Direct

Direct

Direct

Normal cases unless...

If satellite

If mobile...

If model/computer

DPT & OBS combined

 Acquisition

 Composite

 Computation

aka in User views

Linked to Process...
Migration issues and opportunities

• Missing objects in MOLES2 required for MOLES3 records
• Incomplete records (only Data Entity well populated)
• Mapping free-text fields to constrained fields
• Inconsistent content – within and across MOLES2 records

• Large “linting” process possible.
• Migration system + checks captured content issues
• Resolved issues both in migration (automated) and at source (manual)
• Migration also extracted/standardised new fields (e.g. Parties)
## Migration Success

<table>
<thead>
<tr>
<th>MOLES 2 Component</th>
<th>No. Records</th>
<th>MOLES3 counterpart</th>
<th>No. Records</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Entity</td>
<td>310</td>
<td>Observation Collection</td>
<td>314</td>
</tr>
<tr>
<td>Deployments</td>
<td>3026</td>
<td>Observation</td>
<td>3052</td>
</tr>
<tr>
<td>Activity</td>
<td>914</td>
<td>Project</td>
<td>915</td>
</tr>
<tr>
<td>Observation Stations</td>
<td>553</td>
<td>Platform</td>
<td>507</td>
</tr>
<tr>
<td>Data Production Tools</td>
<td>1012</td>
<td>Instrument</td>
<td>865</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Computation</td>
<td>337</td>
</tr>
<tr>
<td><strong>Total MOLES2</strong></td>
<td><strong>5815</strong></td>
<td><strong>Total MOLES3</strong></td>
<td><strong>5990</strong></td>
</tr>
</tbody>
</table>

New MOLES3 record types:

- Acquisitions: 2594
- Composite Process: 245
- Party: 1397
- Responsible Party Info: 43,754
Early limitations and Future Work

• Underlying metadata model has limitations (e.g. data quality description, constraining related observations)

• Full archive heterogeneity difficult to capture: non-geo-spatial (e.g. lab) data; physical archives; non-terrestrial data

• Catalogue-archive connection right allows direct harvesting of metadata (41% of archive is suitably formatted)

• Integration of CHARMe methodology allow further metadata annotations (“C”- metadata)

• Connection to deeper faceted search tools (under development)
Conclusions

• Catalogue requirements continue to evolve
• Structure needs to balance strict standard conformity v pragmatic approach
• Shift from object-orientated to relational catalogue (maintainability, use v changeability)
• Migration is essential: maintain traceability; focus on content too = opportunity to clean records!
• Migration emphasises value of constraining content where possible (free-text v ad hoc mark up v constrained fields)
• Structure now right – focus is now on content and functionality to ensure we provide data context.
Any questions?

CEDA Catalogue: catalogue.ceda.ac.uk

CEDA: www.ceda.ac.uk
Twitter: @cedanews

Email: graham.parton@stfc.ac.uk
Twitter: @gaparton
Web curation/blog: www.scoop.it/t/windgatherer
Bonus material – Catalogue

MOLES3 catalogue example
MST radar facility data – Dataset collection
16 datasets (including 3rd party datasets)
1 project directly connected (3 via datasets)
3 Authors

http://catalogue.ceda.ac.uk
1- citation constructed from: Title, authors, publication date, publisher, UUID fields

2 – temporal + geographic ranged superset of underlying datasets
Implementing CEDA-MOLES
The “NewMoon” approach

Model iteratively improved

CEDA-MOLES (UML) → NewMoon → CEDA-MOLES Database

- Including all ISO 19156 & MOLES3.4
- >680 tables (50 active)
- > 500,000 records each
- High performance and maintenance costs
Implementing CEDA-MOLES
The Django approach

- Off-the-shelf web-framework solution
- Model/View/Control environment with sophisticated DB management
- CEDA expertise

- However, couldn’t use with full CEDA-MOLES UML model – again a structure issue
Implementing CEDA-MOLES
The Django approach

• Simplified CEDA-MOLES UML profile:
  • Dropped unused/difficult to fill classes + attributes
  • Flattened (overcome inheritance issues)

• Resulting database: 29 tables (cf 680!)