



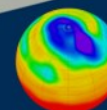
# MOLES3: Implementing an ISO standards driven data catalogue (It's all about context & structure)

IDCC2015, 10<sup>th</sup> February 2015

Graham Parton<sup>1</sup>, Steve Donegan<sup>1</sup>, Bryan Lawrence<sup>2</sup>,  
Stephen Pascoe<sup>1</sup>, Ag Stephens<sup>1</sup>, Spiros Ventouras<sup>1</sup>

1 - Centre for Environmental Archival, STFC Rutherford Appleton Laboratory, U.K.

2 - NCAS, Department of Meteorology, University of Reading, U.K.





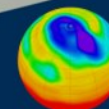
# What's in the talk...

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

The screenshot shows the website's header with the Centre for Environmental Data Archival logo and navigation links. The main content area is titled 'Data Centres' and lists four data centres:

- British Atmospheric Data Centre**: The British Atmospheric Data Centre (BADC), NERC's designated data centre for the UK atmospheric science community, covering climate, composition, observations and NWP data.
- NERC Earth Observation Data Centre**: The NEODC is NERC's designated data centre for Earth Observation data and is part of NERC's National Centre for Earth Observation.
- The UK Solar System Data Centre**: The UK Solar System Data Centre, co-funded by STFC and NERC, curates and provides access to archives of data from the upper atmosphere, ionosphere and Earth's solar environment.
- IPCC Data Distribution Centre**: The Intergovernmental Panel on Climate Change (IPCC) DDC provides climate, socio-economic and environmental data, both from the past and also in scenarios projected into the future. Technical guidelines on the selection and use of different types of data and scenarios in research and assessment are also provided.

- 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

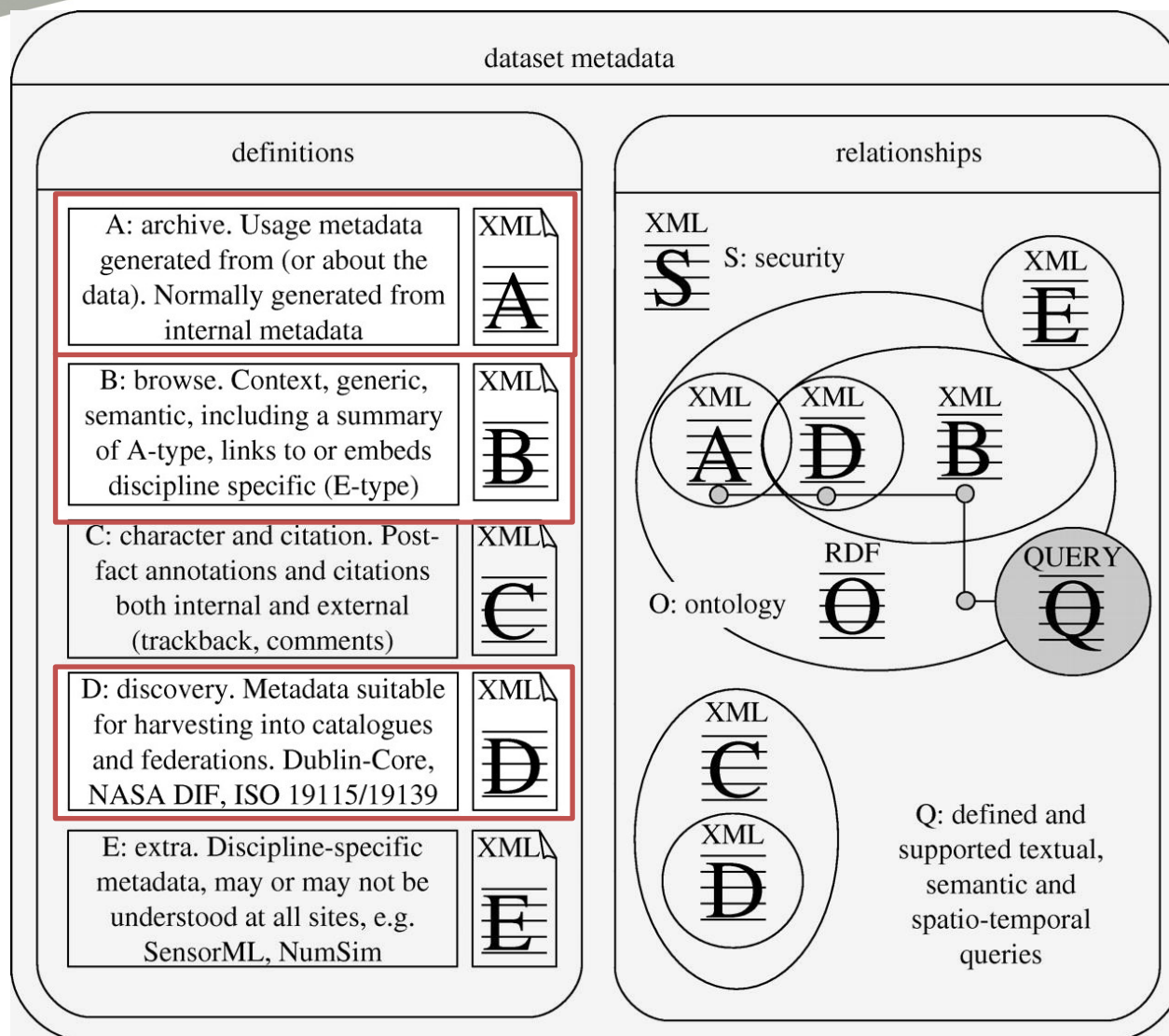
- Can users trust the source?

Providence

- Can they reliably reference the data?

Persistence

**Underpinned by metadata**

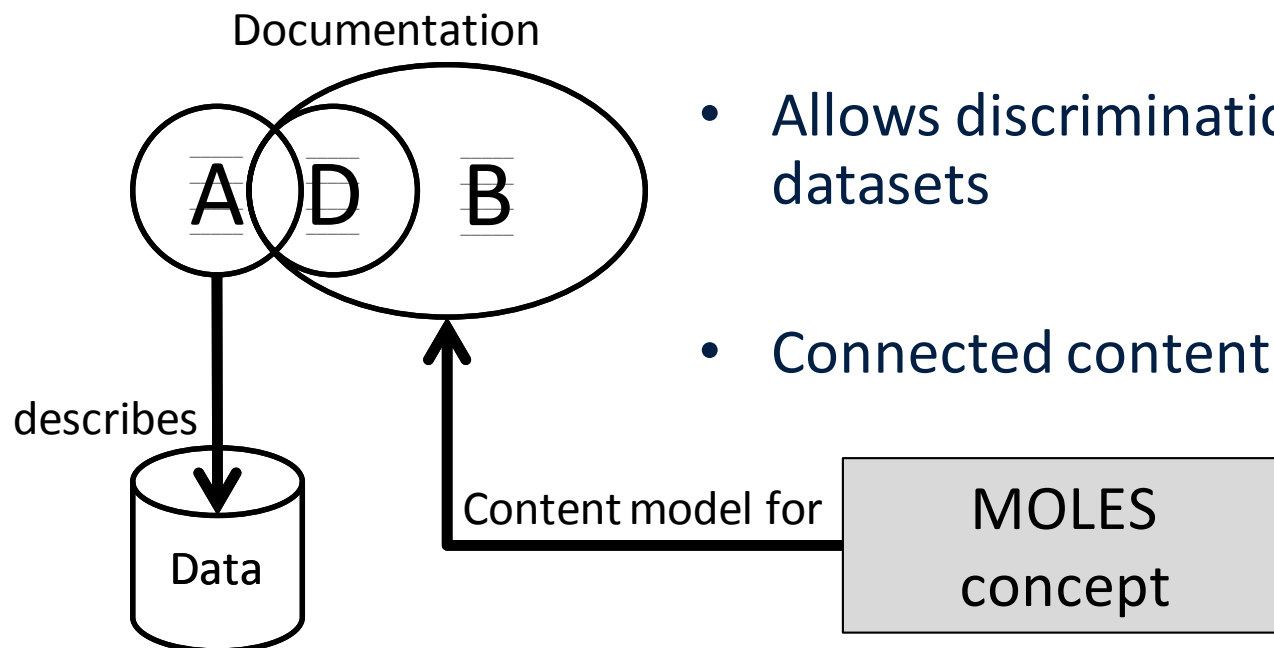


Lawrence, Lowry, Miller, Snaith and Woolf: Information in environmental data grids, *Phil. Trans. R. Soc. A* (2009) doi:10.1098/rsta.2008.0237

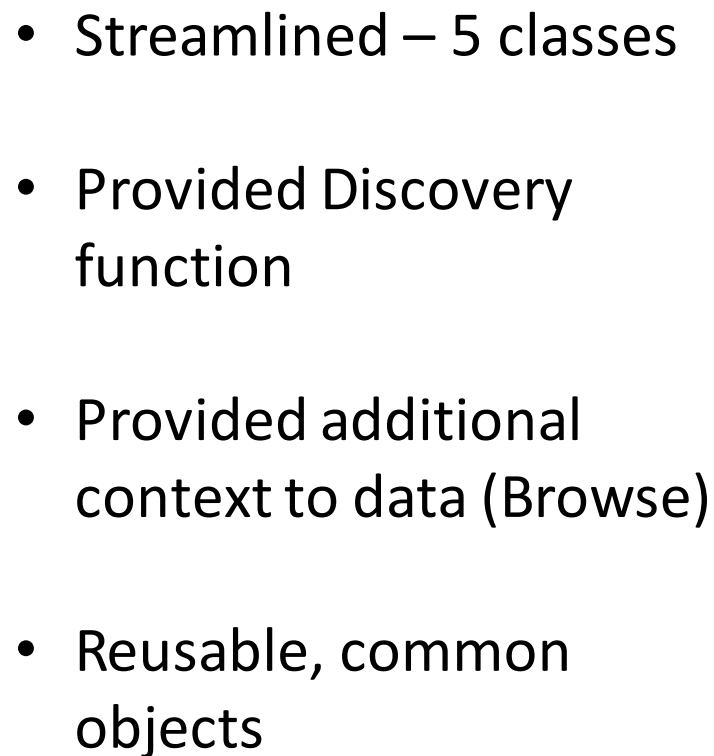


# Metadata Objects Linking Environmental Sciences

- Data context + browse functionality
- Allows discrimination between datasets
- Connected content via shared records

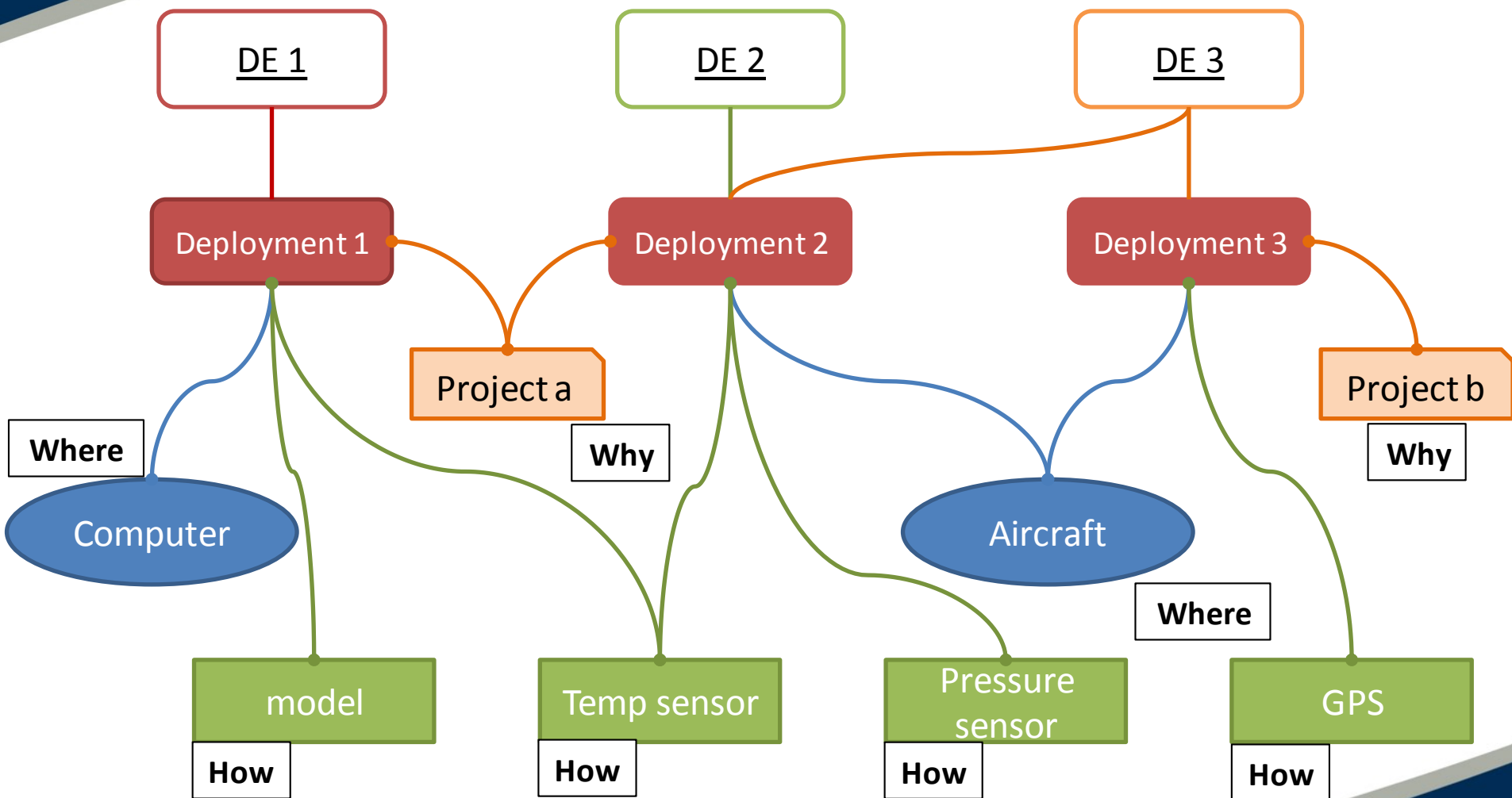








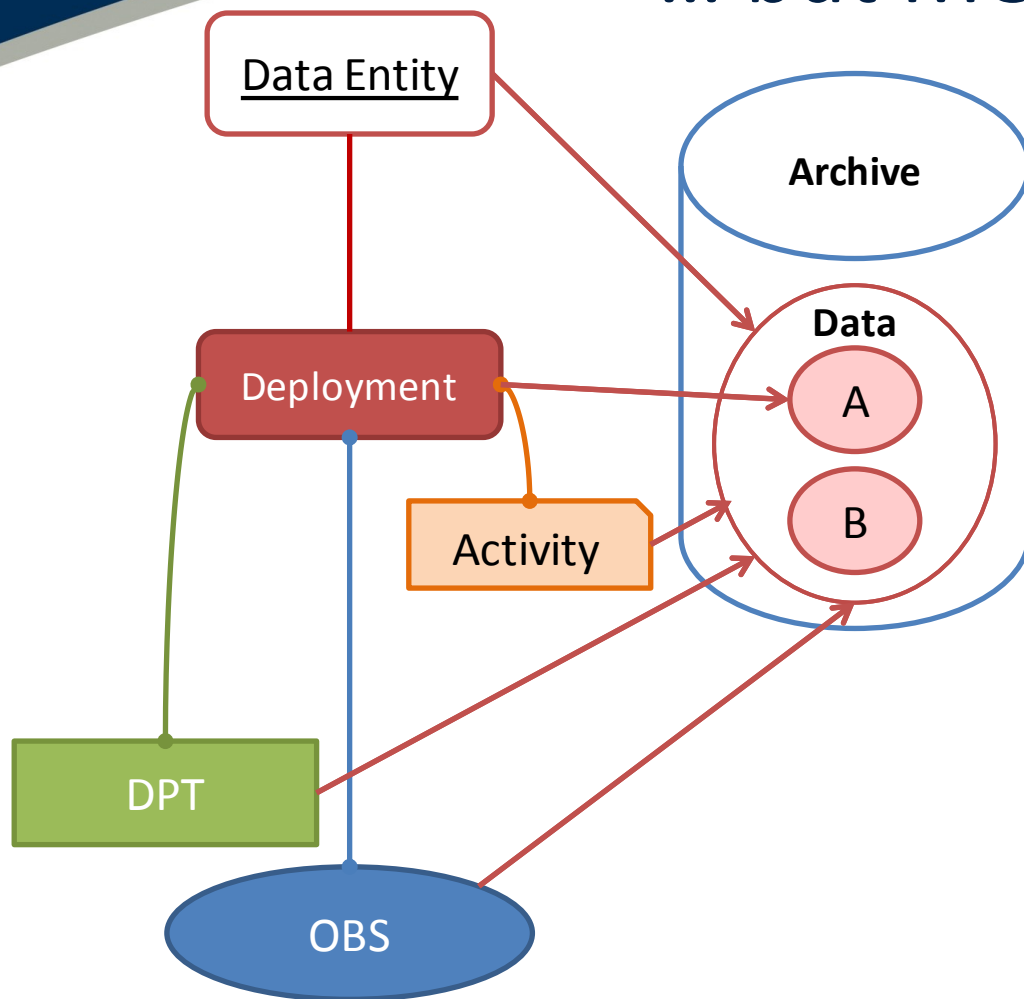
# MOLES2 – Structured Reusability







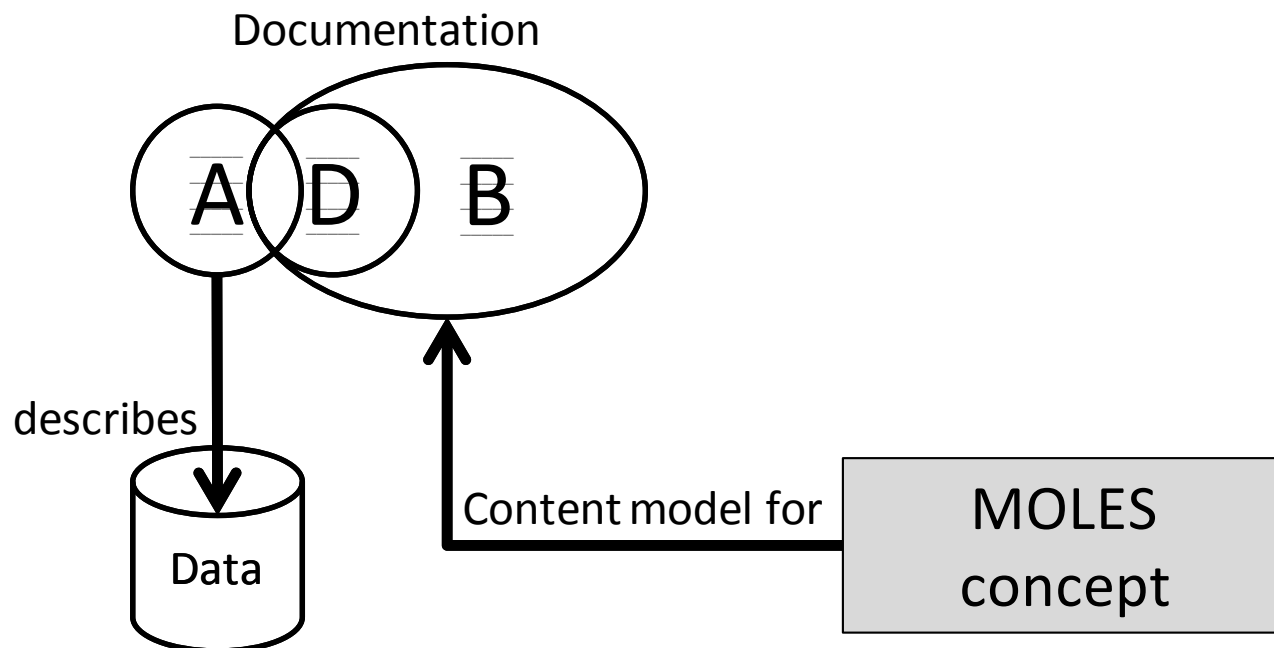
## ... but MOLES2 had problems



- 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

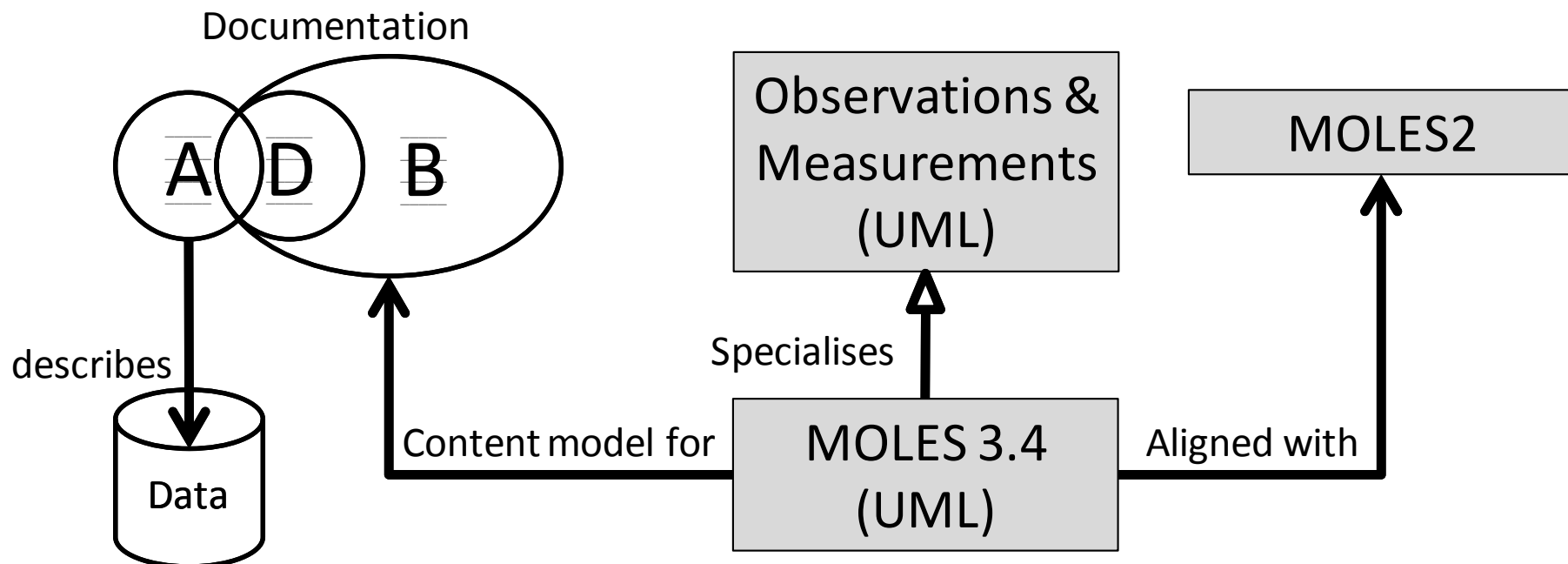


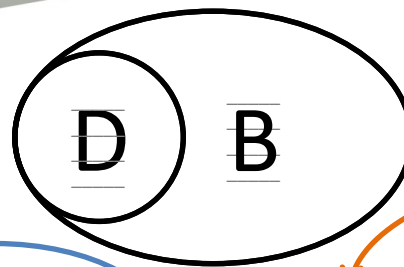
# Evolving the MOLES Concept





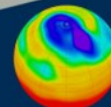
# MOLES + ISO 19156 = MOLES3.4





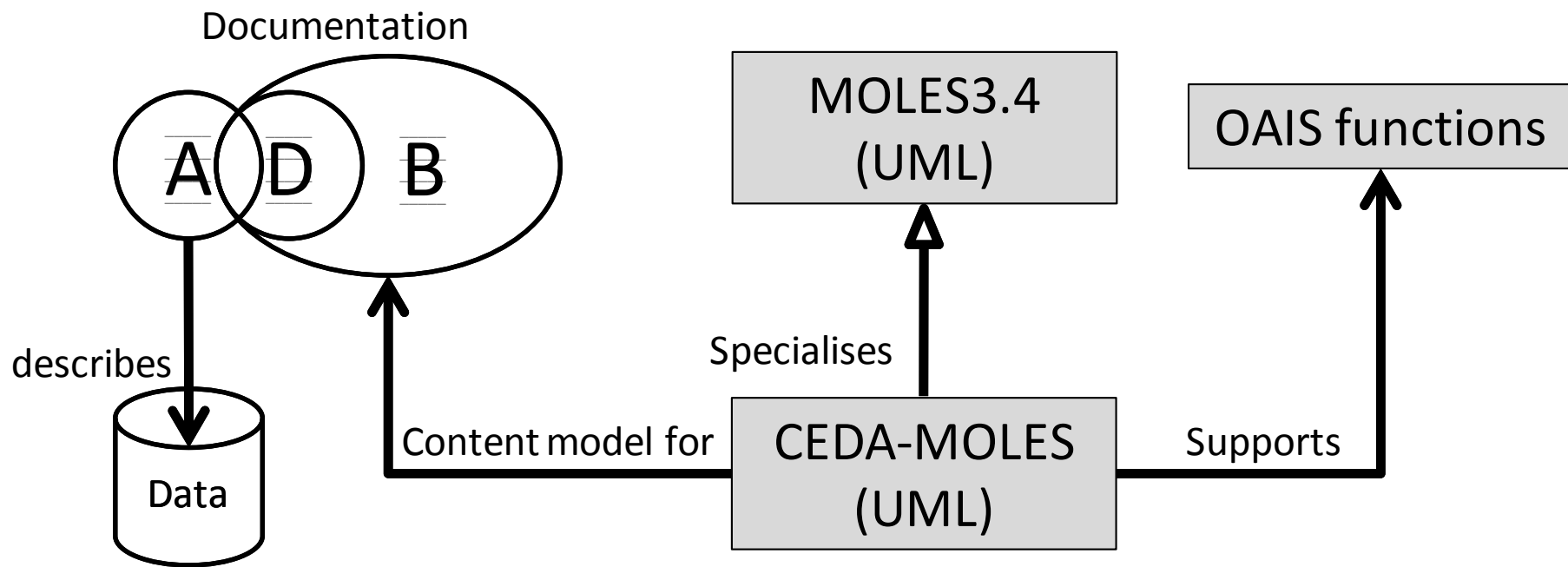


# Implementing MOLES3.4





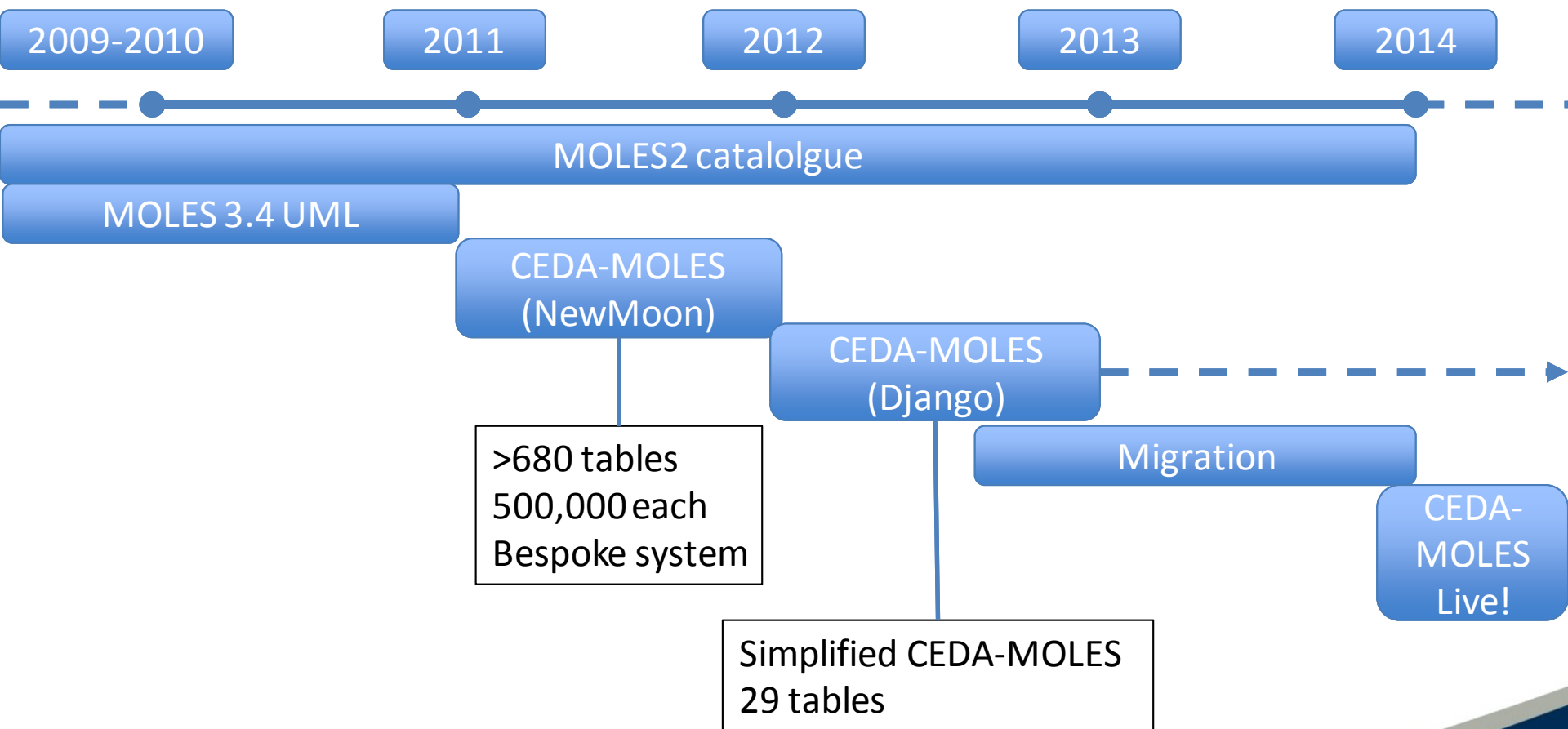
# MOLES3.4 + extra = CEDA-MOLES







# Implementing CEDA-MOLES



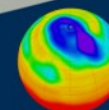


# Populating the Django database

Q: Construct afresh v migrate from MOLES2

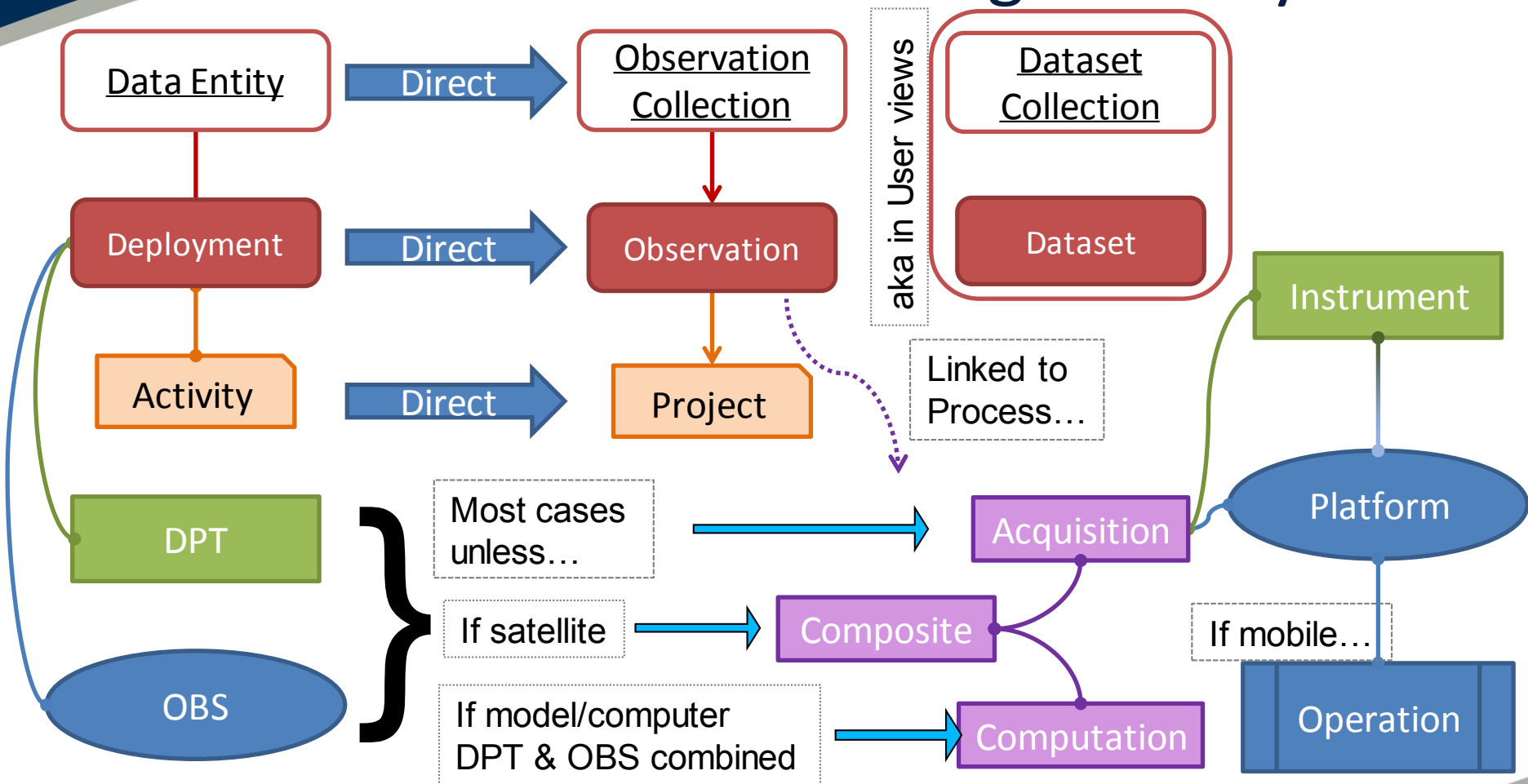
Migration necessary as:

- 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





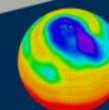
# The Migration System





# 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)





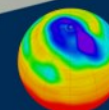
# Migration Success

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

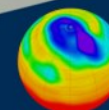






# Conclusions

- 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](http://catalogue.ceda.ac.uk)

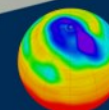
CEDA: [www.ceda.ac.uk](http://www.ceda.ac.uk)

Twitter: @cedanews

Email: [graham.parton@stfc.ac.uk](mailto:graham.parton@stfc.ac.uk)

Twitter: @gaparton

Web curation/blog: [www.scoop.it/t/windgatherer](http://www.scoop.it/t/windgatherer)





# Bonus material – Catalogue

MOLES3 catalogue example

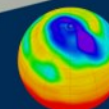
MST radar facility data – Dataset collection

16 datasets (including 3<sup>rd</sup> party datasets)

1 project directly connected (3 via datasets)

3 Authors

<http://catalogue.ceda.ac.uk>





## Dataset Collection

Publication State: Published  
Publication Date: 2005-12-10  
[ Edit Record (Admin only) ]



### The Natural Environment Research Council (NERC) Mesosphere-Stratosphere-Troposphere (MST) Radar Facility at Aberystwyth Data

#### Abstract

This collection contains data from the Natural Environment Research Council (NERC) Mesosphere-Stratosphere-Troposphere (MST) Radar Facility at Capel Dewi, near Aberystwyth in West Wales. The principal measurements made by the MST radar, a 46.5 MHz pulsed Doppler radar, ideally suited for studied of atmospheric winds, waves and turbulence. It is run predominantly in the ST mode (approximately 2 - 20 km altitude) for which MST radars are unique in their ability to give continuous measurements of the three dimensional wind vector at high resolution (typically 2 - 3 minutes in time and 300 m in altitude). Under certain circumstances they can additionally provide information about humidity, static stability (thus allowing monitoring of the altitude and sharpness of the tropopause) and turbulence of at least moderate intensity. Surface meteorological measurements from the radar site, ceilometer data, sky camera images and wind speed and direction recorded from a 10m tower located at Frongoch (6km away) are also available. Other instruments at

**Citable as:** Natural Environment Research Council Mesosphere-Stratosphere-Troposphere Radar Facility; Natural Environment Research Council Mesosphere Stratosphere Troposphere Radar Facility at Aberystwyth; Hooper, D.A. (2008): The Natural Environment Research Council (NERC) Mesosphere-Stratosphere-Troposphere (MST) Radar Facility at Aberystwyth Data. NCAS British Atmospheric Data Centre, date of citation. <http://catalogue.ceda.ac.uk/uuid/bd095d88e4a9f0c708b08058dbad3b31>

#### Datasets (16)

Surface Meteorological Data from the NERC MST Radar Facility, Capel Dewi, Wales

Aberporth Radiosonde Data

Surface Pressure, Temperature and Relative Humidity Data from the Vaisala WXT...

Precipitation Data from the Vaisala WXT510 instrument deployed at the NERC MS...

Met Office GPS Integrated Water Vapour (IWW) Data from the NERC MST Radar Fac...

Natural Environment Research Council (NERC) Mesosphere-Stratosphere-Troposphere...

Laser Ceilometer Data from the Natural Environment Research Council (NERC) Me...

Met Office 915 MHz UHF Radar Data deployed at the NERC MST Radar Facility, Ca...

#### Temporal Range

1989-07-01 00:00:00

Present

#### Geographic Extent



-4.559444° 52.4° -4.0°  
52.114722°

#### Related People and Organisations (6)

Natural Environment Research Council Mesosphere-Stratosphere-Troposphere Radar Facility (MSTRF) (Author)

David A. Hooper (Author)

Natural Environment Research Council Mesosphere Stratosphere Troposphere Radar Facility at Aberystwyth (Author)

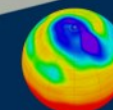
NCAS British Atmospheric Data Centre (NCAS BADC) (Publisher)

NCAS British Atmospheric Data Centre (NCAS BADC) (Curator)

NCAS British Atmospheric Data Centre (NCAS BADC) (Curator)

1- citation constructed from: Title, authors, publication date, publisher, UUID fields

2 – temporal + geographic ranged superset of underlying datasets





## Dataset

Update Frequency: Daily  
Latest Data Update: 2014-09-28  
Status: Ongoing  
Publication State: Published  
Publication Date: 2012-09-28  
[ Edit Record (Admin only) ]

### Surface Pressure, Temperature and Relative Humidity Data from the Vaisala WXT510 instrument located at the NERC MST Radar Facility, Capel Dewi, Wales

Apply for access

Download

#### Abstract

Surface pressure, Temperature and humidity data (PTU) are collected by a Vaisala WXT510 instrument located at the Natural Environment Research Council's (NERC) Mesosphere-Stratosphere-Troposphere (MST) Radar Facility, Capel Dewi, near Aberystwyth in West Wales. Rainrate data from this instrument are also available as a separate dataset within the MST Radar Facility dataset collection. Independent surface meteorological data are also collected from a suite of instruments by a Campbell Scientific CR10 Climate Data Logger and are also available as a separate dataset within the MST Radar Facility dataset collection.

**Citable as:** Natural Environment Research Council Mesosphere Stratosphere Troposphere Radar Facility at Aberystwyth; Hooper, D. (2012): Surface Pressure, Temperature and Relative Humidity Data from the Vaisala WXT510 instrument located at the NERC MST Radar Facility, Capel Dewi, Wales. NCAS British Atmospheric Data Centre, date of citation.

<http://catalogue.ceda.ac.uk/uuid/8d7a920827e8137145f75dfe08d322dc>

#### Additional Information

**Dataset is part of:** Dataset Collection: The Natural Environment Research Council (NERC) Mesospher...

**Process that generated the data:** Acquisition Process for: Natural Environment Research Council (NERC) Mesosphere-Stratosphere-Troposphere (MST) Radar Facility Pressure, Temperature and Relative Humidity Data

**Observed/simulated phenomena:** ATMOSPHERICSTABILITY  
SPECTRALWIDTH  
SURFACEWINDS  
TROPOPAUSE  
TURBULENCE  
UPPERLEVELWINDS  
UZONALWINDSPEED  
VERTICALWINDMOTION  
VERTICALWINDSPEED  
VMERIDIONALWINDSPEED

#### Temporal Range

2007-12-21 00:00:00

Present

#### Geographic Extent



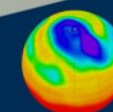
-4.0° 52.4° -4.0°  
52.4°

#### Related People and Organisations (7)

David Hooper (Author)  
Natural Environment Research Council Mesosphere Stratosphere Troposphere Radar Facility at Aberystwyth (Author)  
NCAS British Atmospheric Data Centre (NCAS BADC) (Publisher)  
NCAS British Atmospheric Data Centre (NCAS BADC) (Curator)  
NCAS British Atmospheric Data Centre (NCAS

1- Download link

2 -Links to other records and other A, B and D metadata

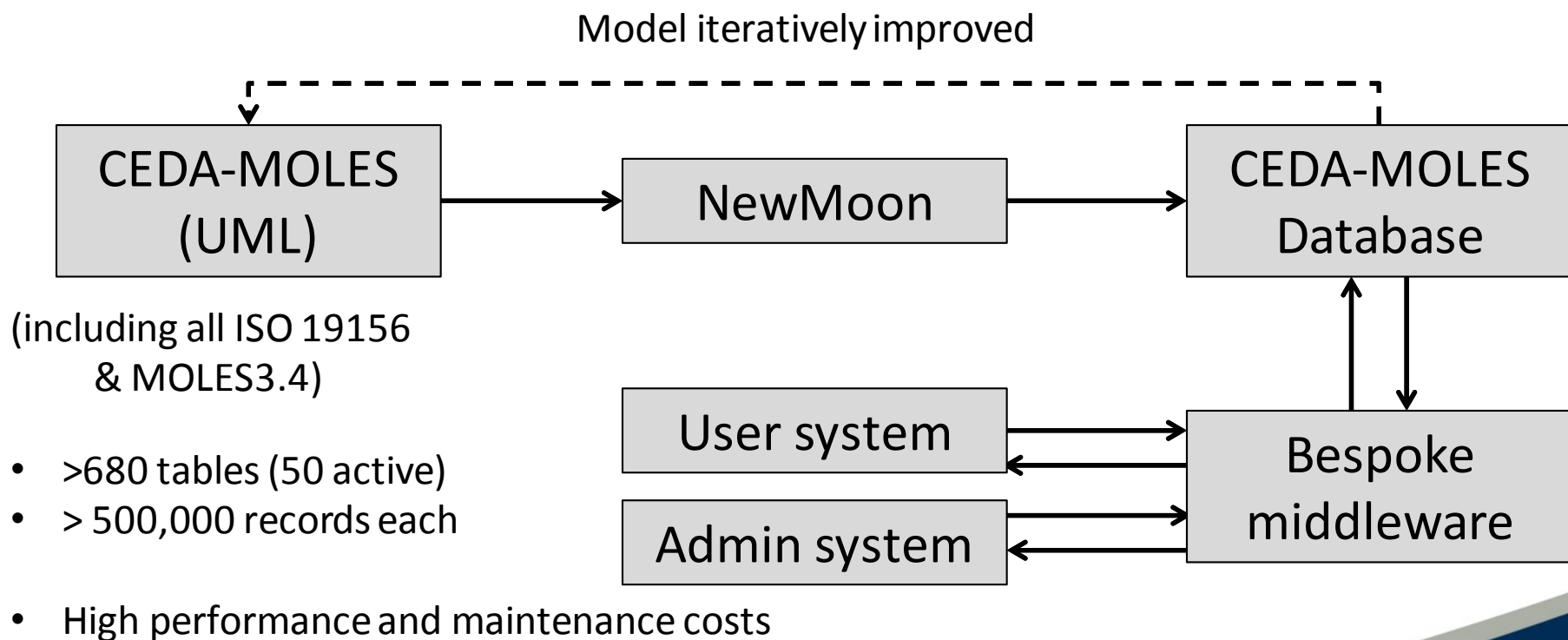






# Implementing CEDA-MOLES

## The “NewMoon” approach



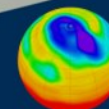




# Implementing CEDA-MOLES

## The Django approach

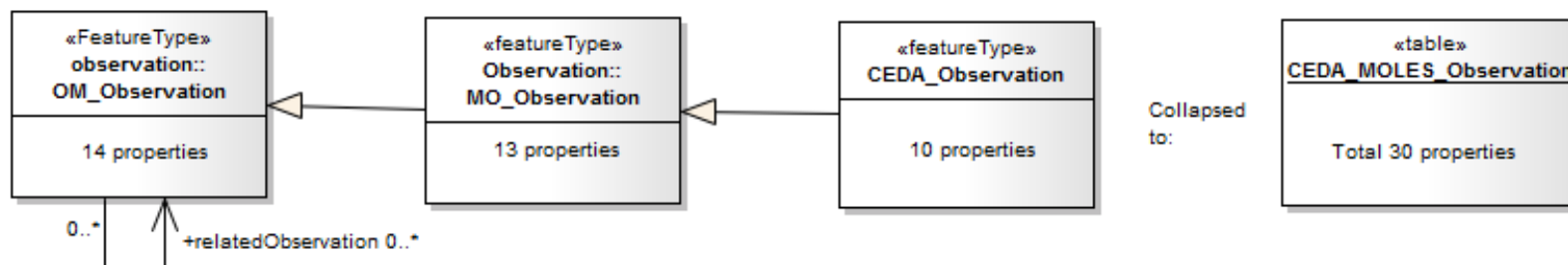
- Off-the shelf web-framework solution
- Model/View/Control environment with sophisticated DB management
- CEDA expertise
- However, couldn't use with full CEDA-MOLES UML model – again a structure issue



# 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!)