MOLES3: Implementing an ISO standards driven data catalogue

(It's all about context & structure)

IDCC2015, 10th February 2015

Graham Parton[1], Steve Donegan[1], Bryan Lawrence[2],
Stephen Pascoe[1], Ag Stephens[1], Spiros Ventouras[1]

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?

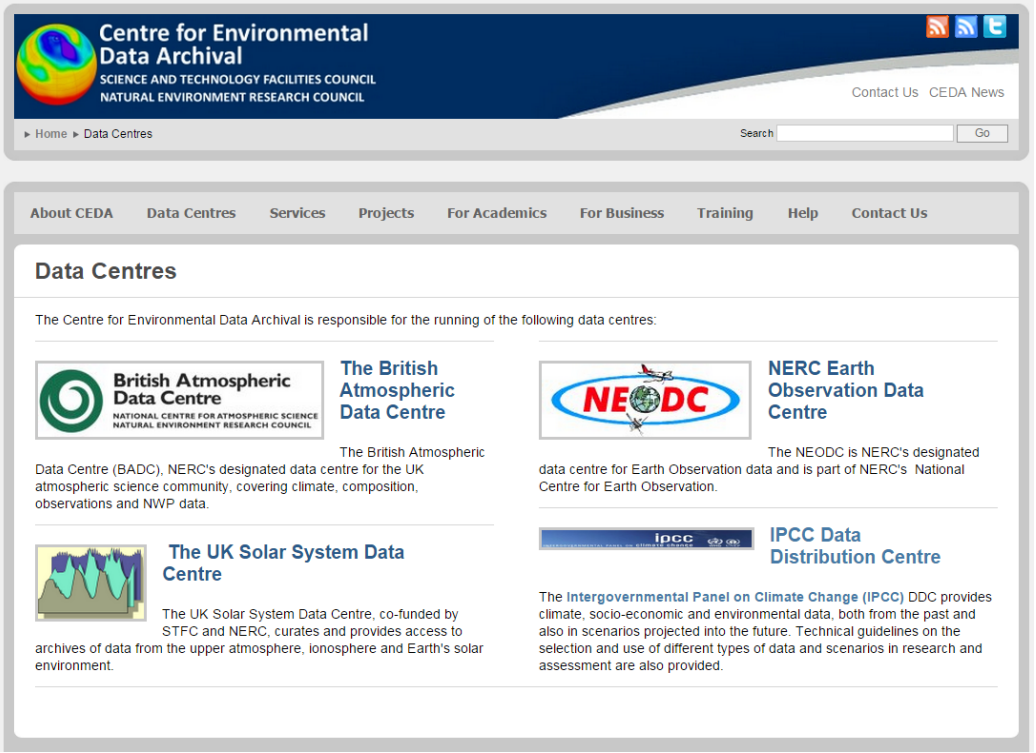
Lessons learnt getting to this catalogue

… 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?

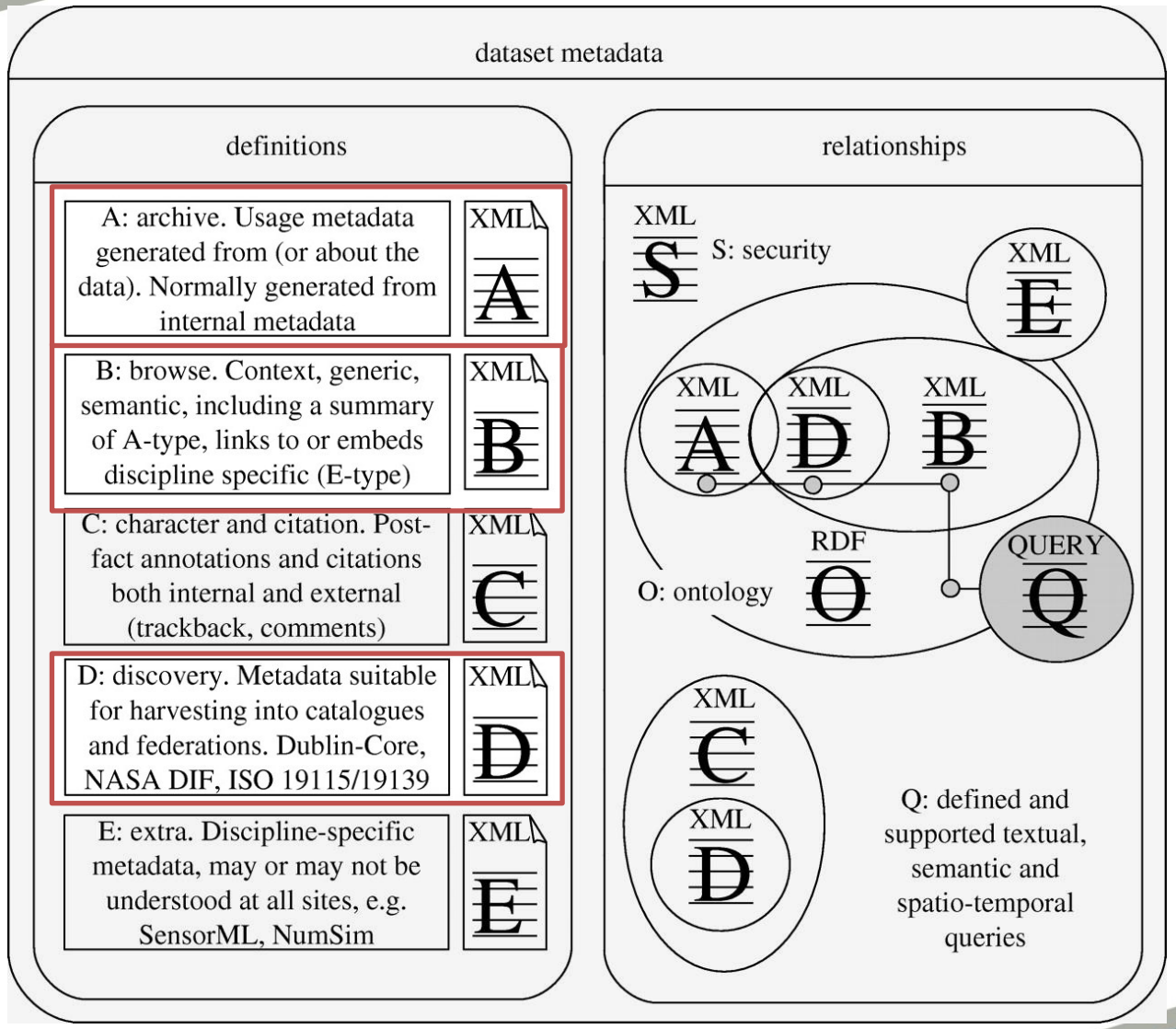
  Discovery
- How will users find, compare, select and use data?

  Context
- Can users trust the source?

  Providence
- Can they reliably reference the data?

  Persistence

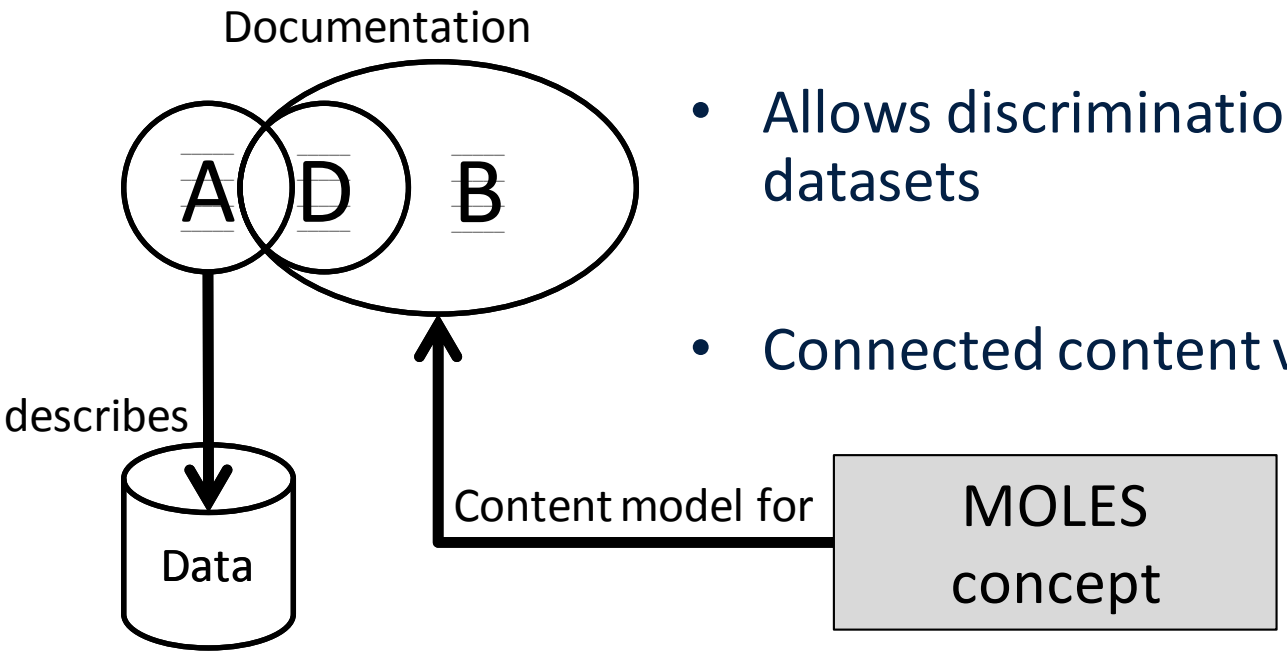
Underpinned by metadata

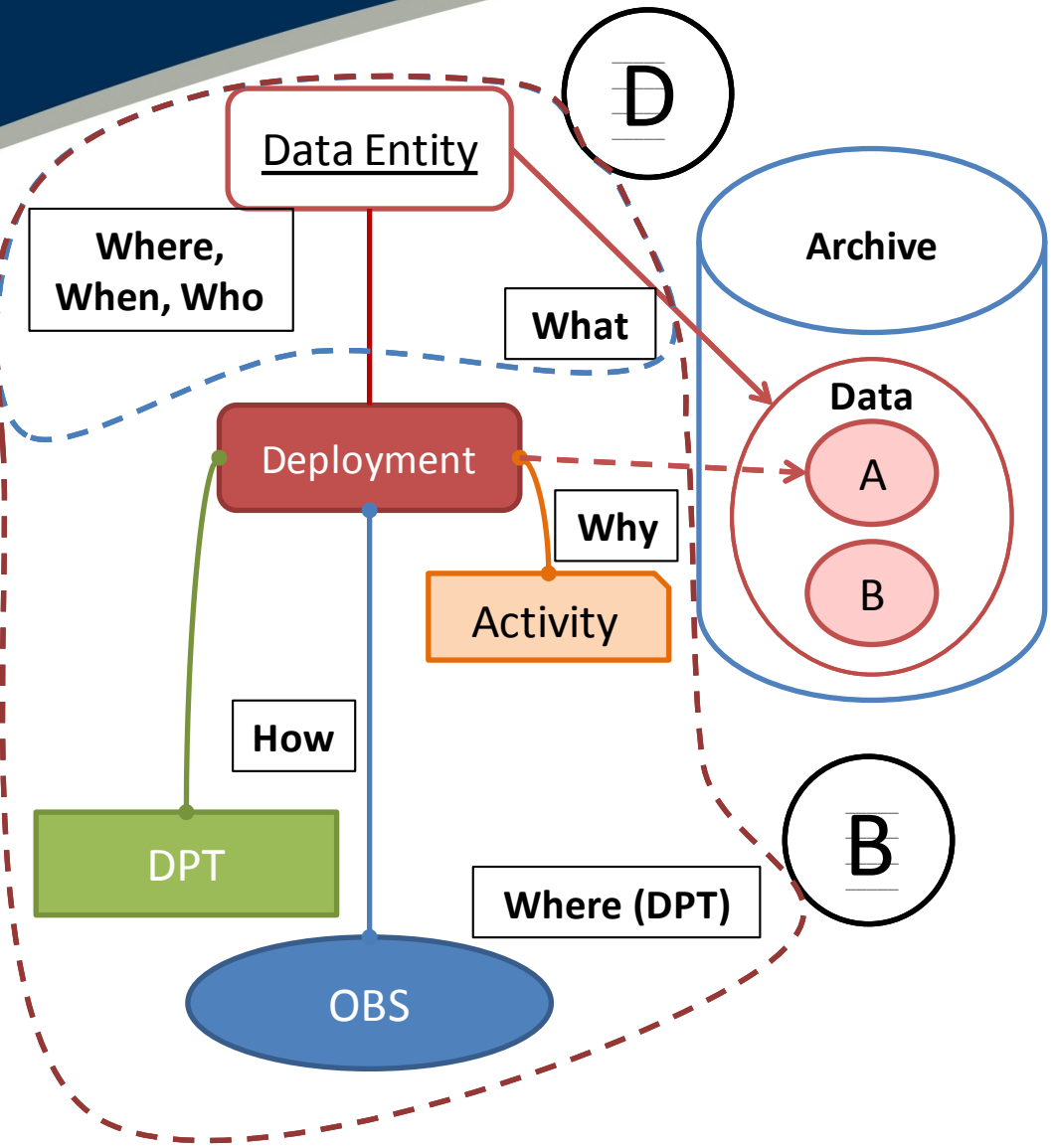


Lawrence, Lowry, Miller, Snaith and Woolf: Information in environmental data grids, Phil. Trans. R. Soc. A (2009) doi:10.1098/rsta.2008.0237

Metadata Objects Linking Environmental Sciences

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

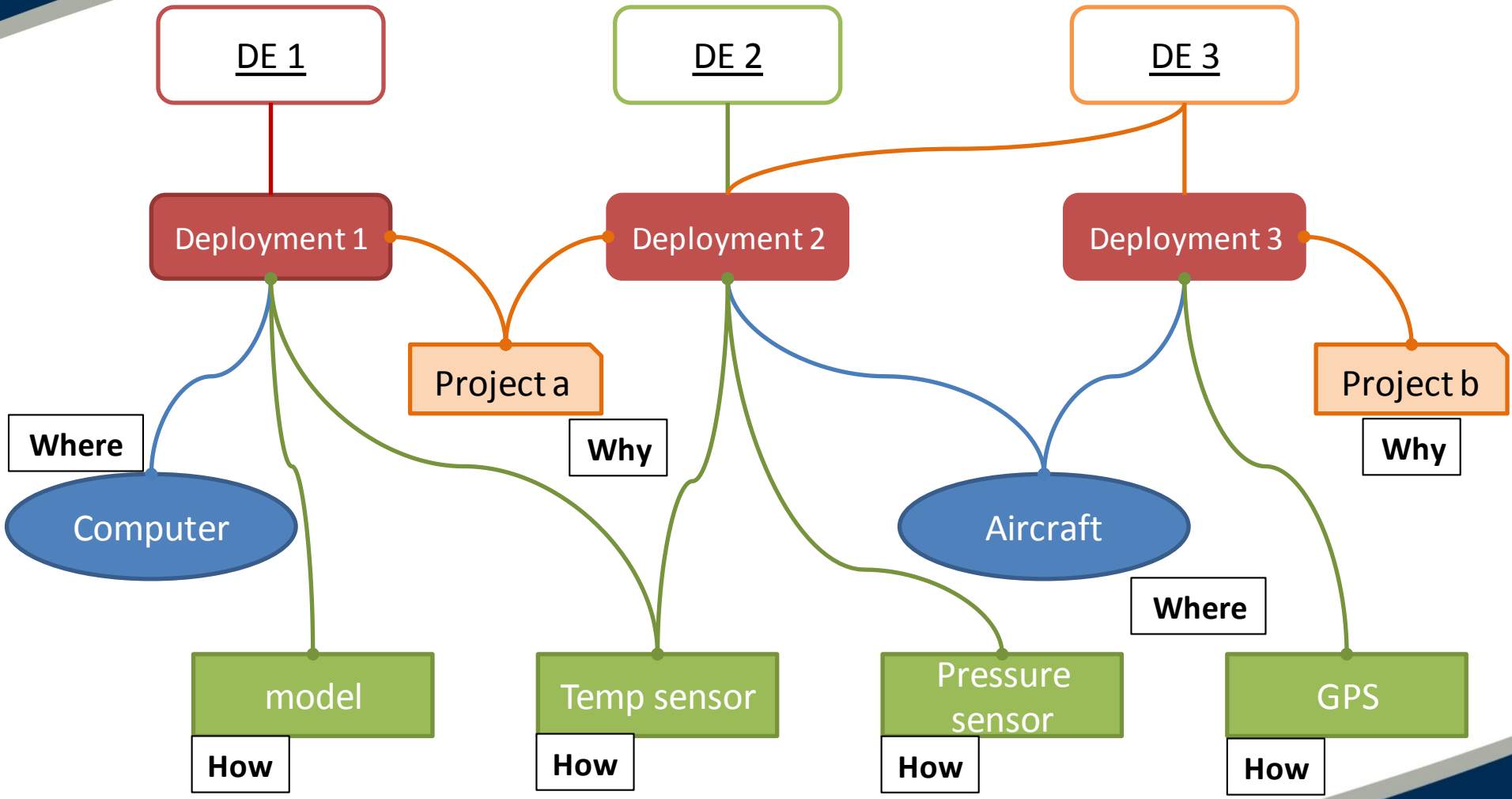




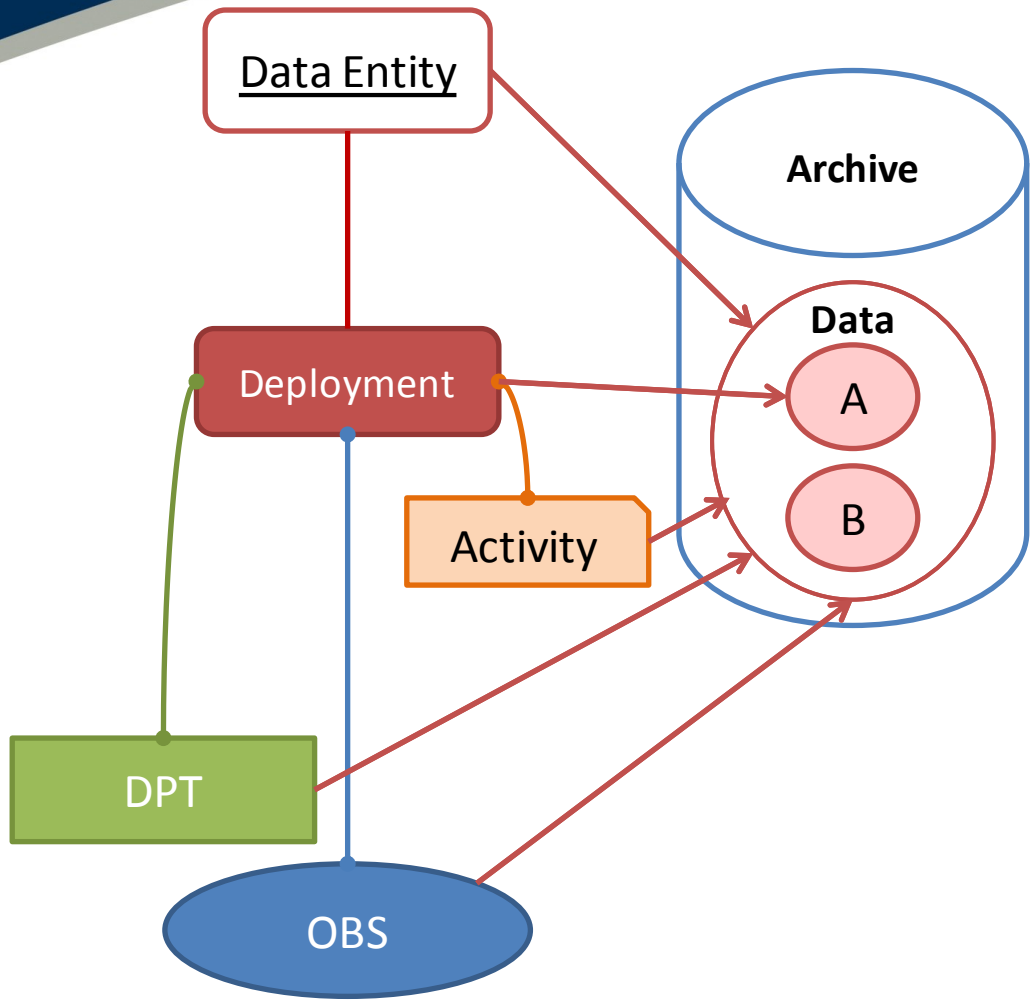
MOLES2

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

MOLES2 – Structured Reusability

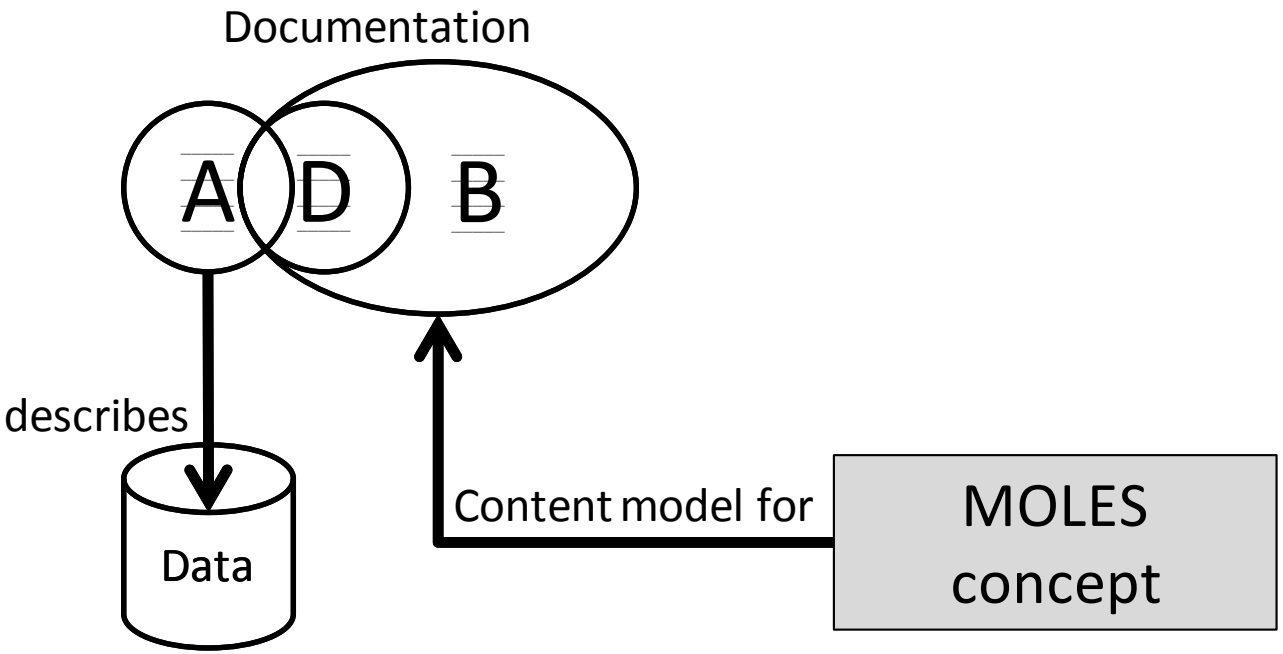


… but MOLES2 had problems

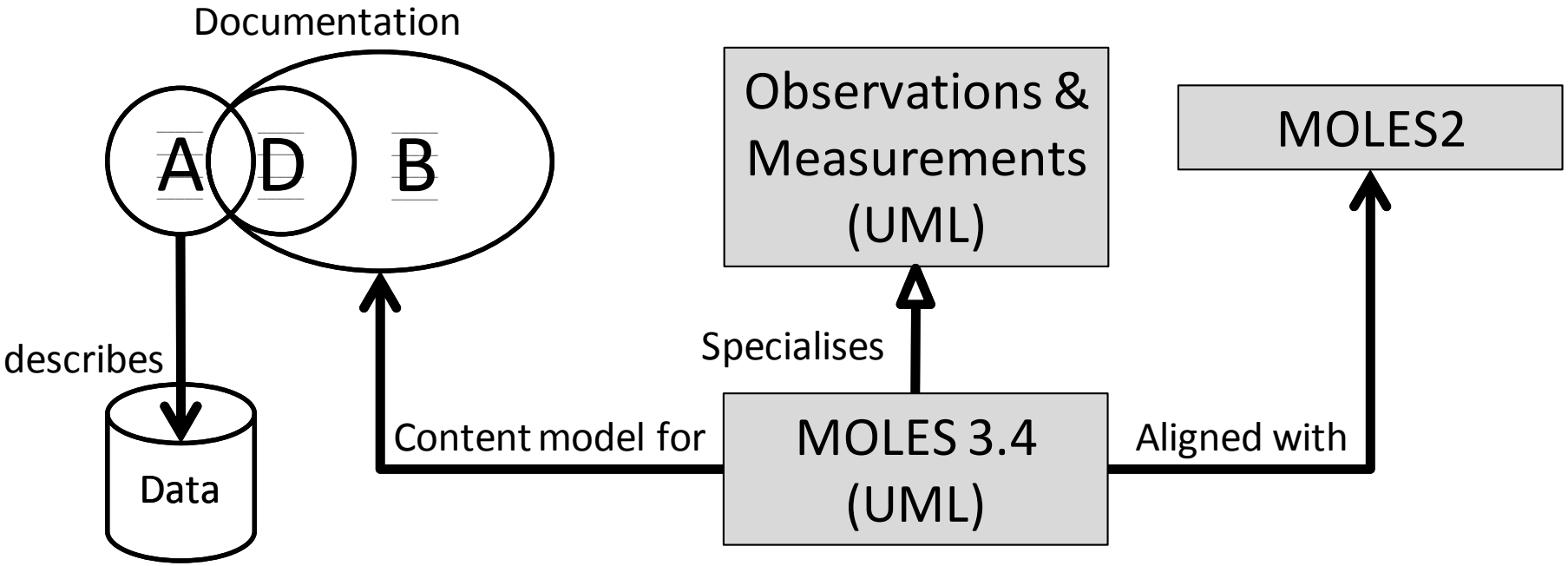


- Key attributes not reusable (e.g. names) = duplicates, inconsistent
- Lack of constraints = use was subverted
- Over-use of free-text fields
- Lacked ISO compliant fields (needed for EU INSPIRE)
- Couldn't export to downstream services
- Couldn't support DOI landing pages (granularity & ISO issue)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

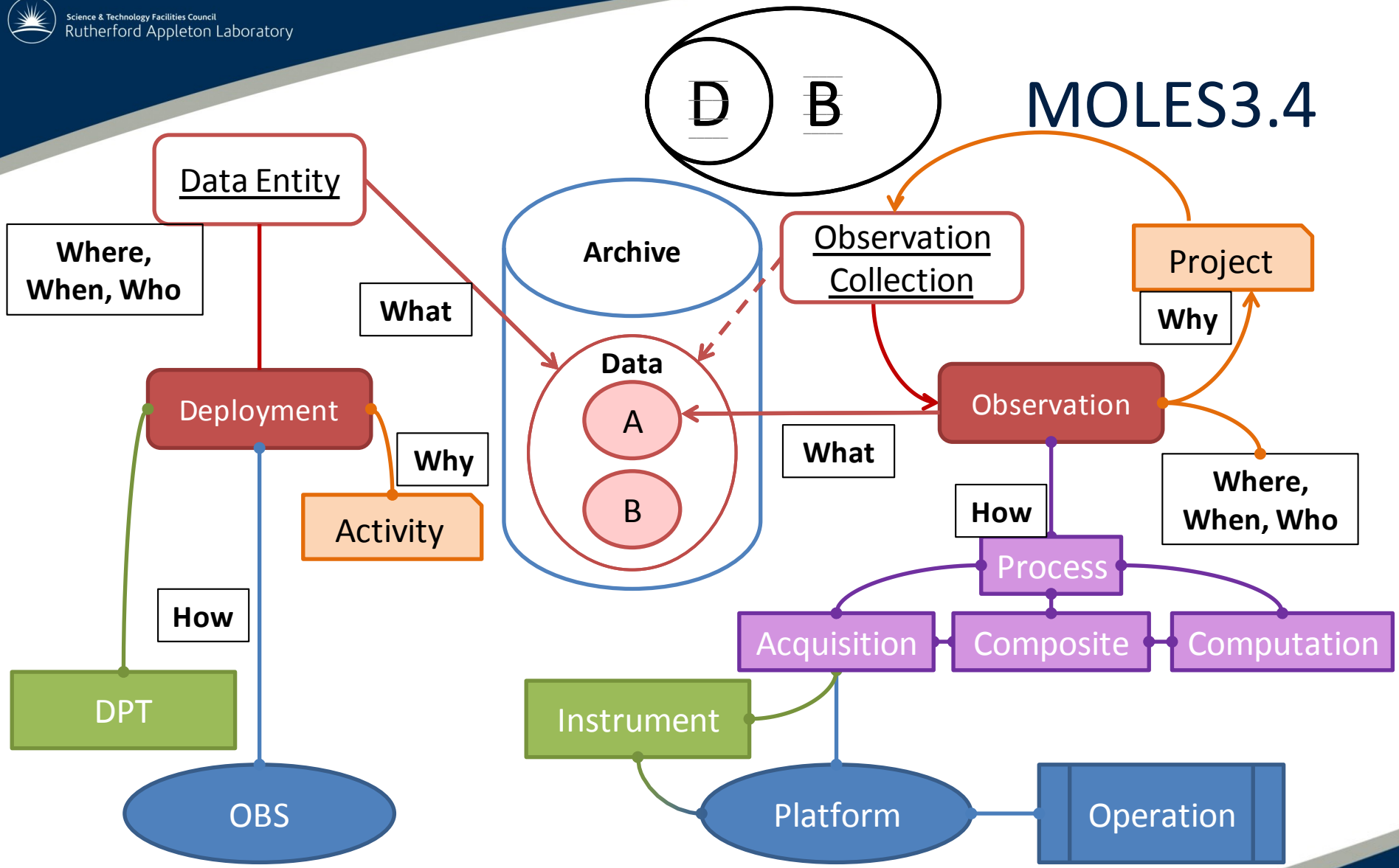
Evolving the MOLES Concept



MOLES + ISO 19156 = MOLES3.4

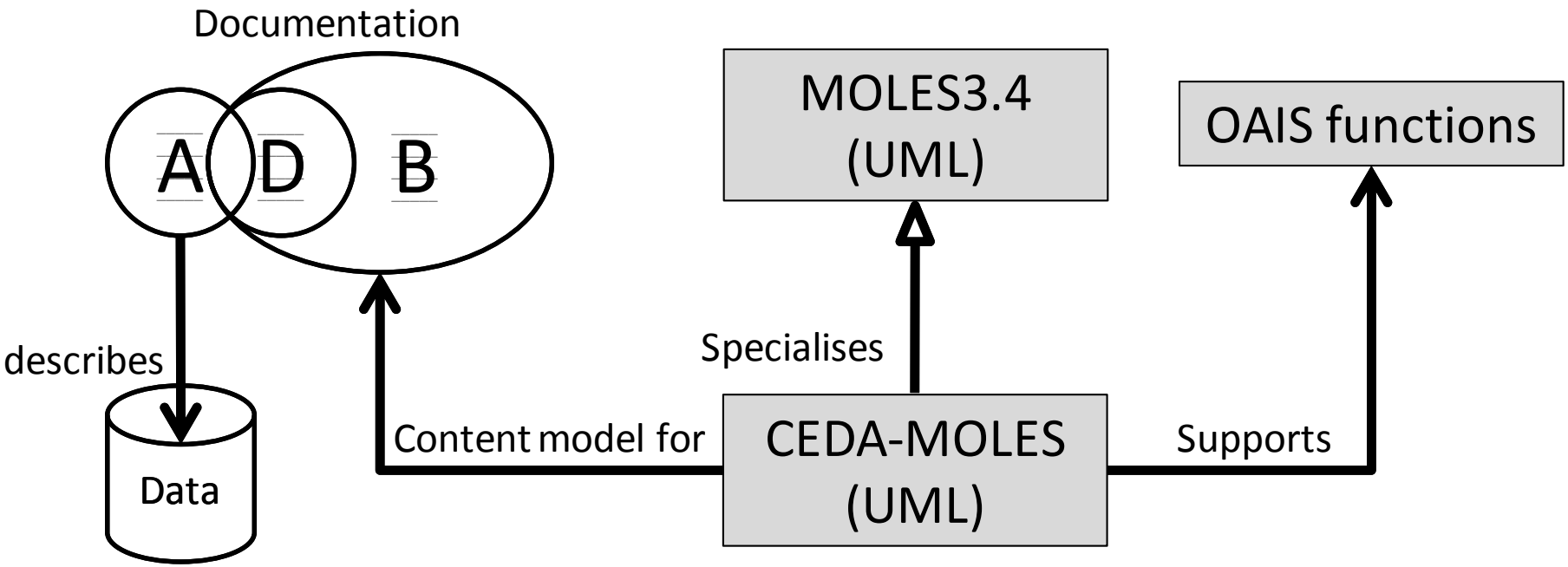


MOLES3.4

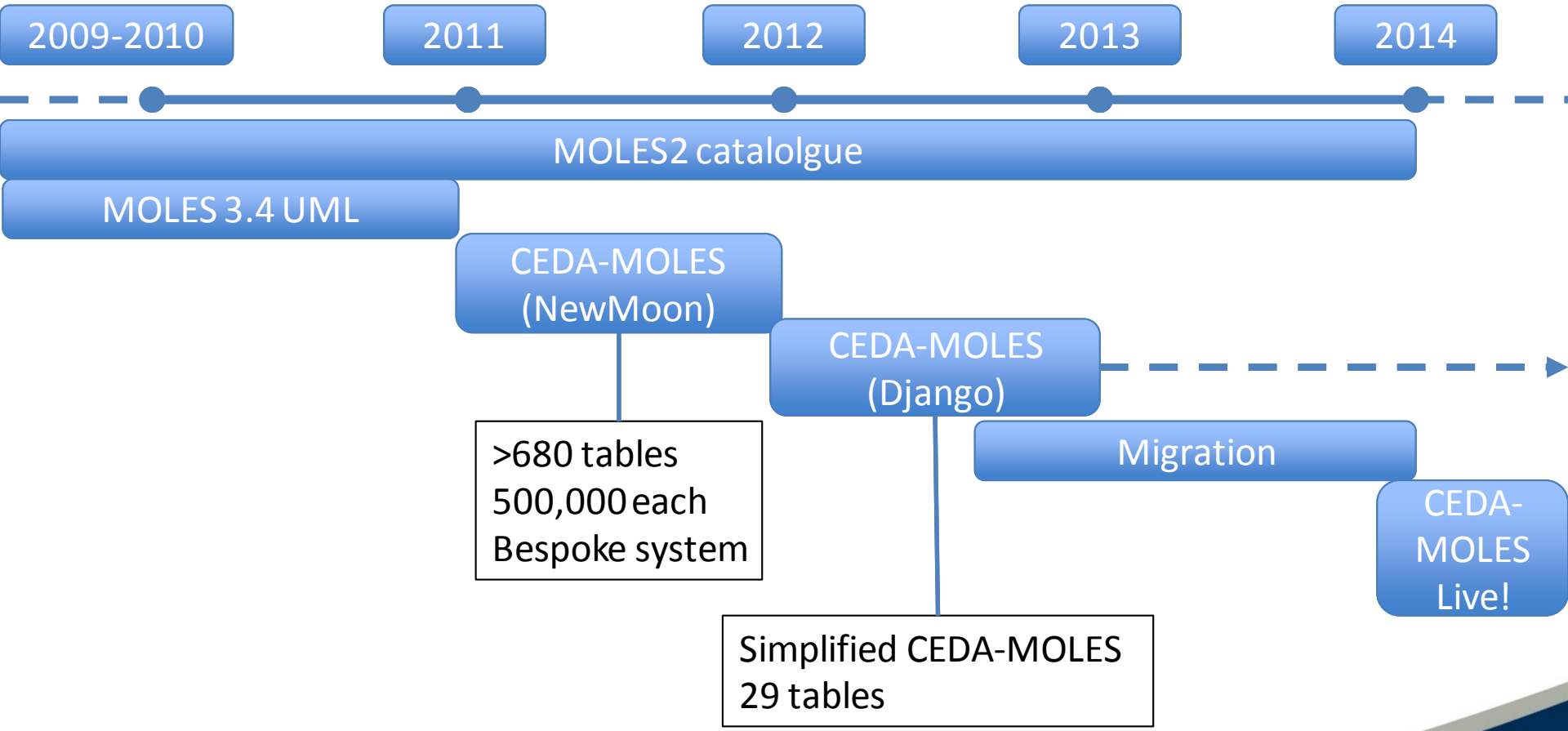

D
B
Data Entity
Where, When, Who
What
Archive
Data
A
B
Observation Collection
Project
Why
Deployment
Why
Activity
How
DPT
OBS
What
Observation
Where, When, Who
How
Process
Acquisition
Composite
Computation
Instrument
Platform
Operation

Implementing MOLES3.4

MOLES3.4 + extra= CEDA-MOLES



Implementing CEDA-MOLES



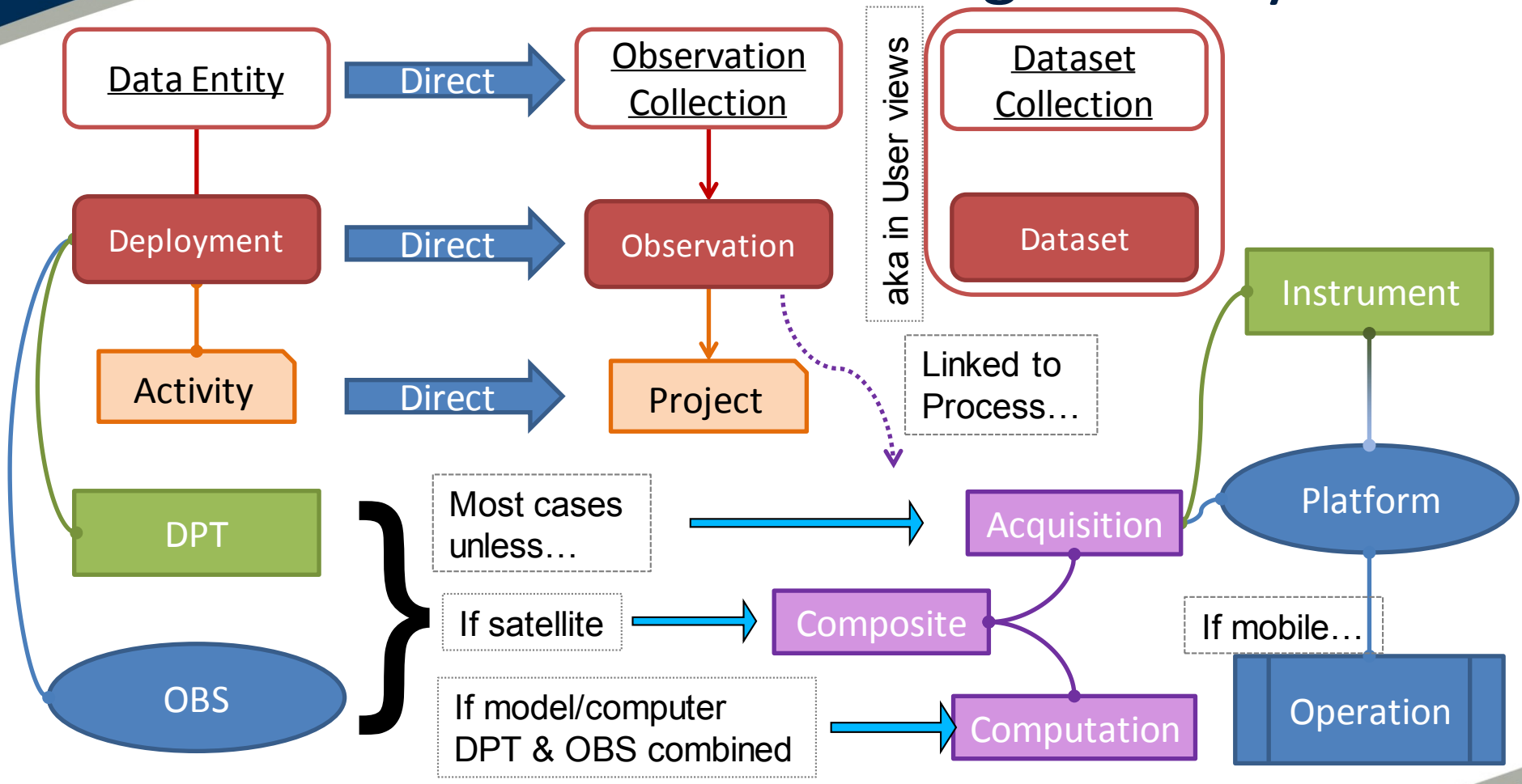
Populating the Django database

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 Data Production Tools	553 1012	Platform Instrument Computation	507 865 337
Total MOLES2	**5815**	**Total MOLES3**	**5990**
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)

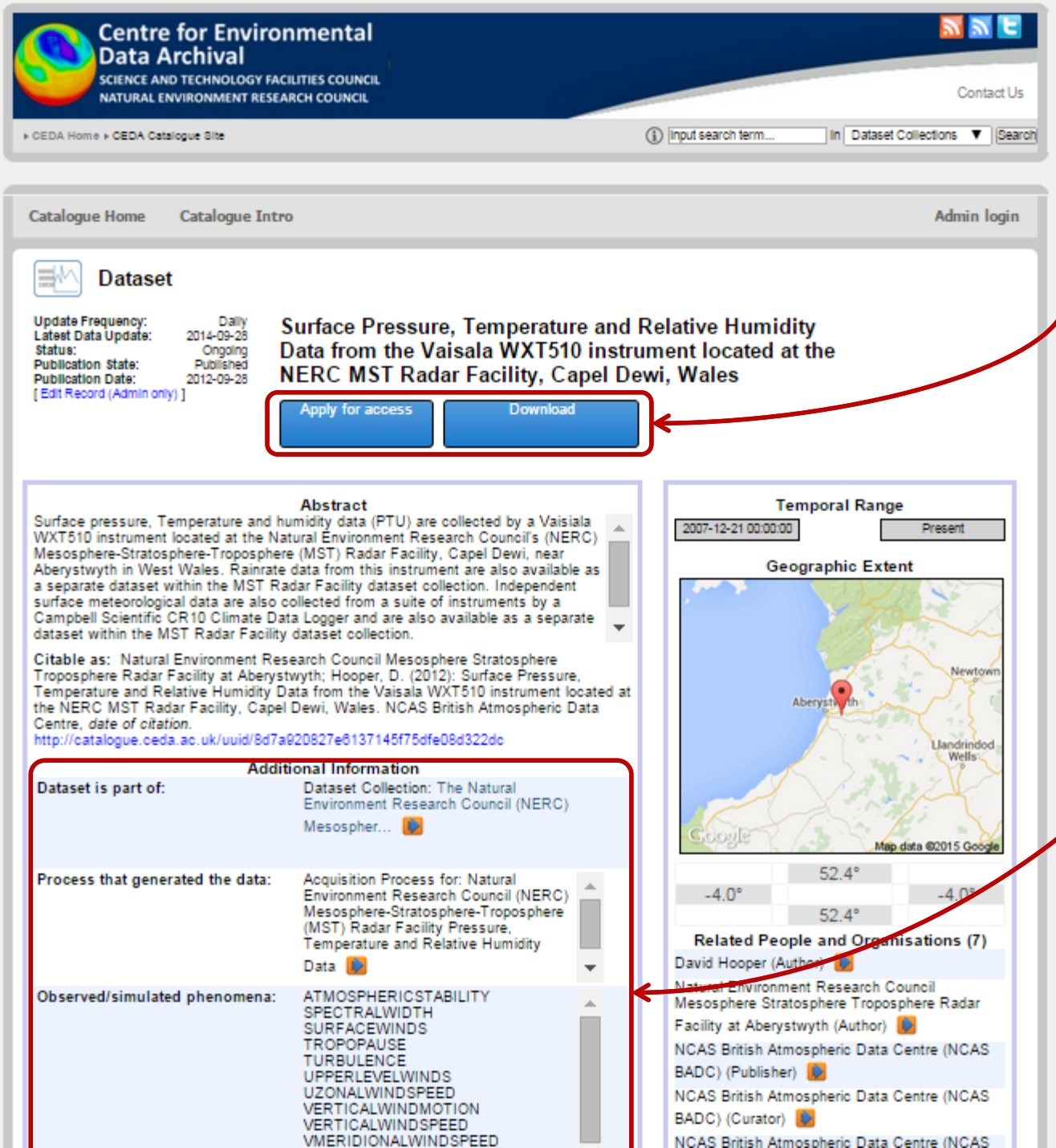
3 Authors

http://catalogue.ceda.ac.uk

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

2 – temporal + geographic ranged superset of underlying datasets

Centre for Environmental Data Archival
SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL

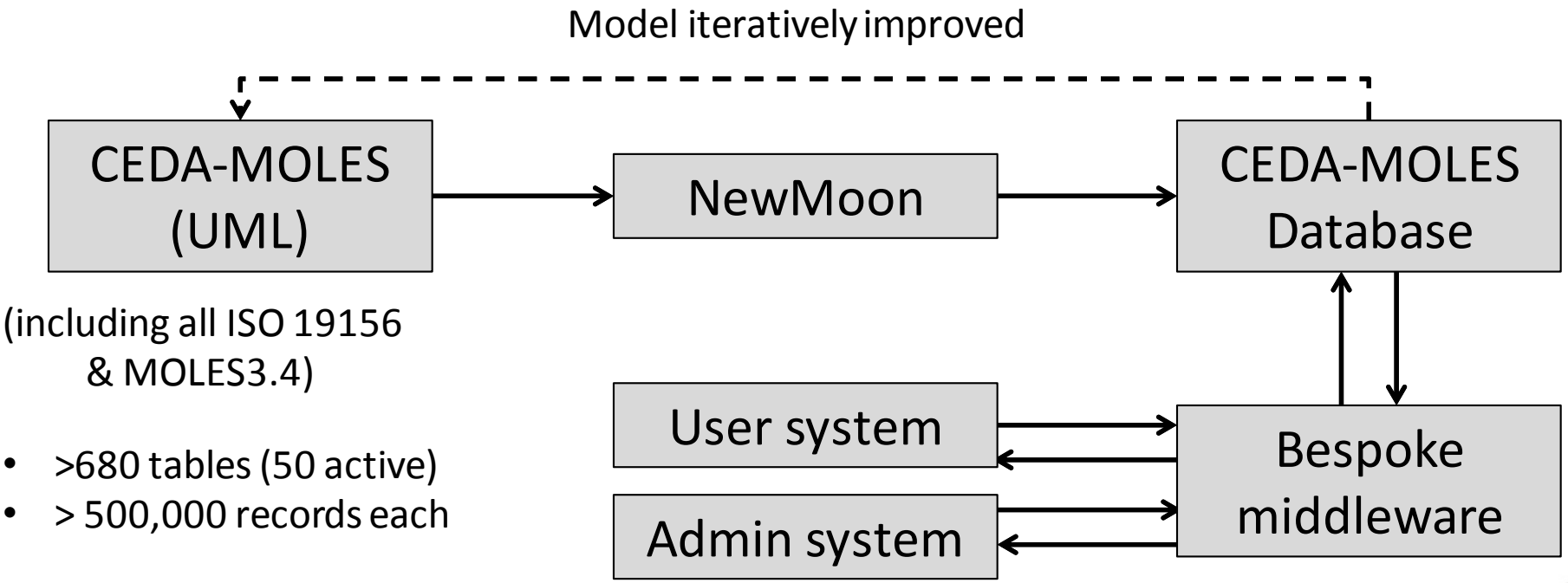


1- Download link

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

Centre for Environmental Data Archival
SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL

Implementing CEDA-MOLES
The "NewMoon" approach



(including all ISO 19156 & MOLES3.4)

- >680 tables (50 active)
- > 500,000 records each

- High performance and maintenance costs

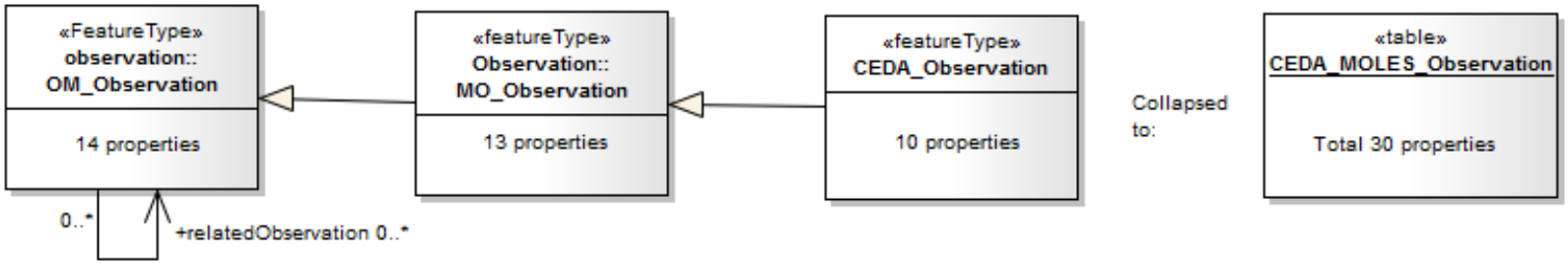
Implementing CEDA-MOLES
The Django approach

- 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!)