Science & Technology Facilities Council

British experience with building standards based networks for climate and environmental research

Bryan Lawrence

CEDA

Rutherford Appleton Laboratory

Outline

- Organisational Drivers
  - NERC (Natural Environment Research Council)
  - European & Discipline Drivers
- Technology Drivers
  - Some personal perspectives on alphabet soup (SOAP, REST, OGC services)
  - ISO standards & Semantic Web
- Our own endeavours
  - Information
  - Software
  - Access Control

This is talk 1 of 3 linked talks:
Talk 2 "*rethinking metadata*" addresses metadata specific things, and
Talk 3 "*provenance, metadata and information to support climate*" puts all these things together in an extended discussion of how we are supporting climate science in general and climate modelling specifically.
All three will be online (eventually) at http://home.badc.rl.ac.uk/lawrence/talks

Drivers in General

Requirements
Solution Purchasers:
"you need to find
and utilise material
from multiple organisations
to solve this problem"

Legislative
Governments:
"you must be able
to work together,
and you're not,
so we'll make you"

Technology
Industry:
"There is a market
in working together,
particularly
if you use my tools "

Future Proofing
Aesthetics:
"There is an elegant
and maintainable
solution to your problem
Use it!

Multiple requirements
Multiple technologies
Multiple governments
Often
pulling apart?

Image from http://www.flickr.com/photos/sokabs/

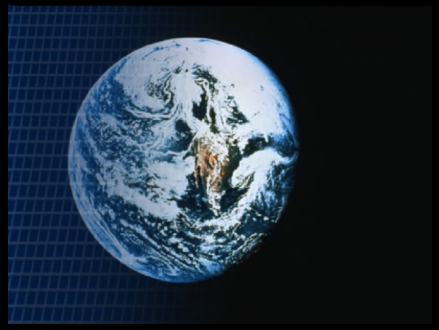
Science & Technology Facilities Council

NERC Data Centres

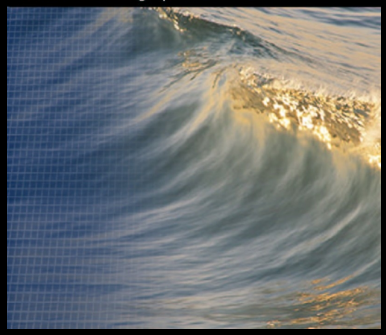
Hydrology:
National Water Archive

Atmosphere:
British Atmospheric Data Centre

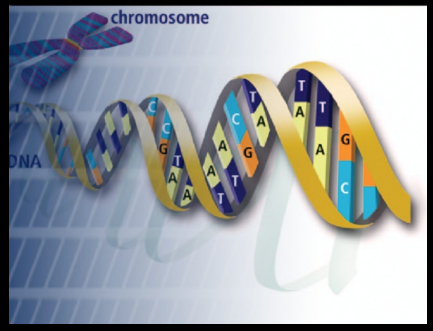
Earth observation:
NERC Earth Observation Data Centre



Ocean & marine:
British Oceanographic Data Centre



Bioinformatics:
NERC Environmental Bioinformatics Centre



Earth:
National Geoscience Data Centre

Terrestrial & freshwater:
Environmental Information Centre

Polar:
Antarctic Environmental Data Centre

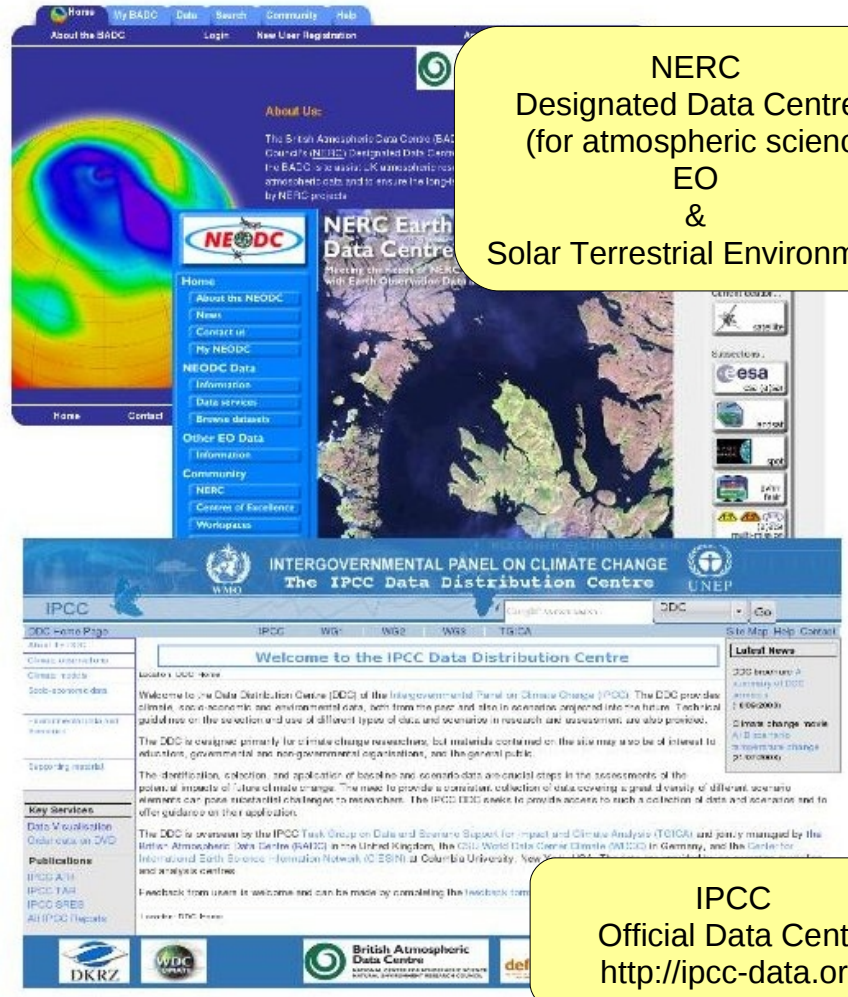
British Atmospheric Data Centre
NATIONAL CENTRE FOR ATMOSPHERIC SCIENCE
NATURAL ENVIRONMENT RESEARCH COUNCIL

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

National Centre for Earth Observation
NATURAL ENVIRONMENT RESEARCH COUNCIL

CEDA
Centre for Environmental Data Archival

- British Atmospheric Data Centre
- NERC Earth Observation Data Centre
- IPCC Data Distribution Centre
- UK Projects
  - NDG (NERC Data Grid)
  - Defra Climate Impacts LINK
  - UKClimate Projections 09
- EU Projects
  - Metafor (Climate Model Documentation)
  - EUFAR (Distributed Archive for European Aircraft)
  - IS-ENES (Distribute Archive for European Climate Model Data)
- Major role in Standards on the European and Global Scale
- Petascale Archive



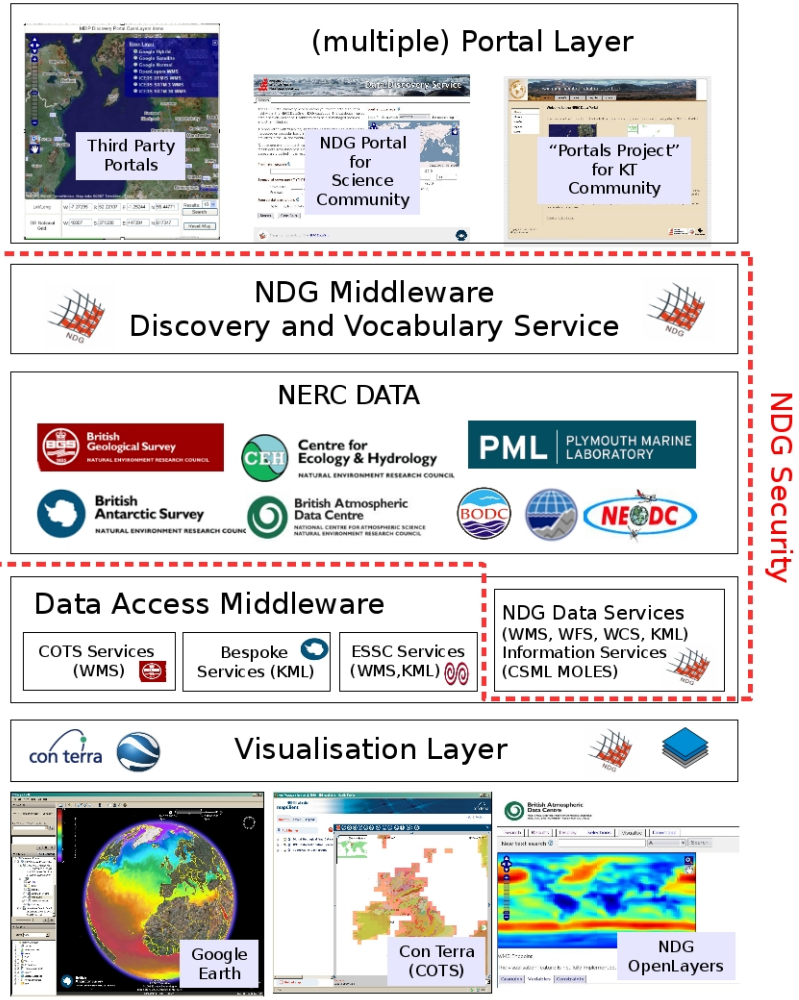
NERC
Designated Data Centres
(for atmospheric science
EO
&
Solar Terrestrial Environment)

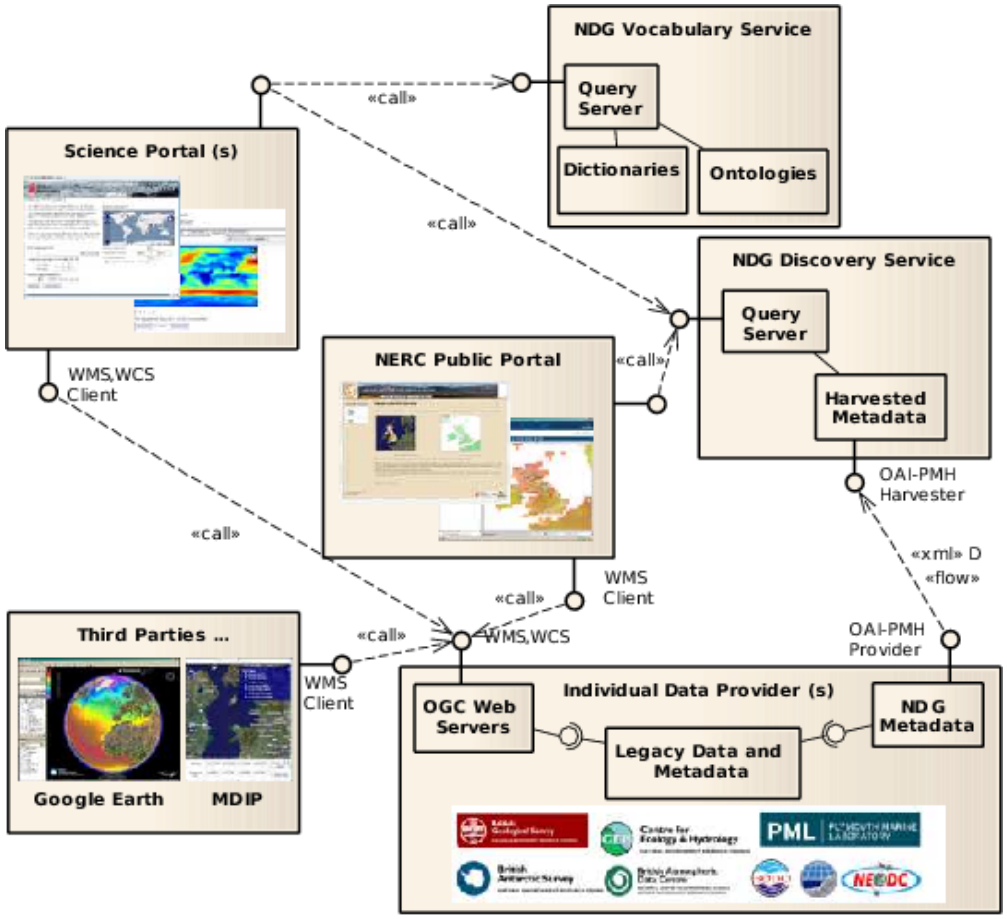
IPCC
Official Data Centre
http://ipcc-data.org

… and just for the BADC:

- Dataset: A collection of files sharing some administrative and/or project heritage.
- BADC has approximately 150 real datasets (and thousands of virtual datasets).
- BADC has approx 200 million files containing thousands of measured or simulated parameters.
- BADC tries to deploy information systems that describe those data, parameters, projects and files, along with services that allow one to manipulate them …
- Calendar year 2008: 1800 active users (of 12000 registered), downloaded 30 TB data in 13 million files from 134 datasets.
- ***Less than half of the BADC data consumers are "atmospheric science" users!***

Some prior art:NERC DataGrid



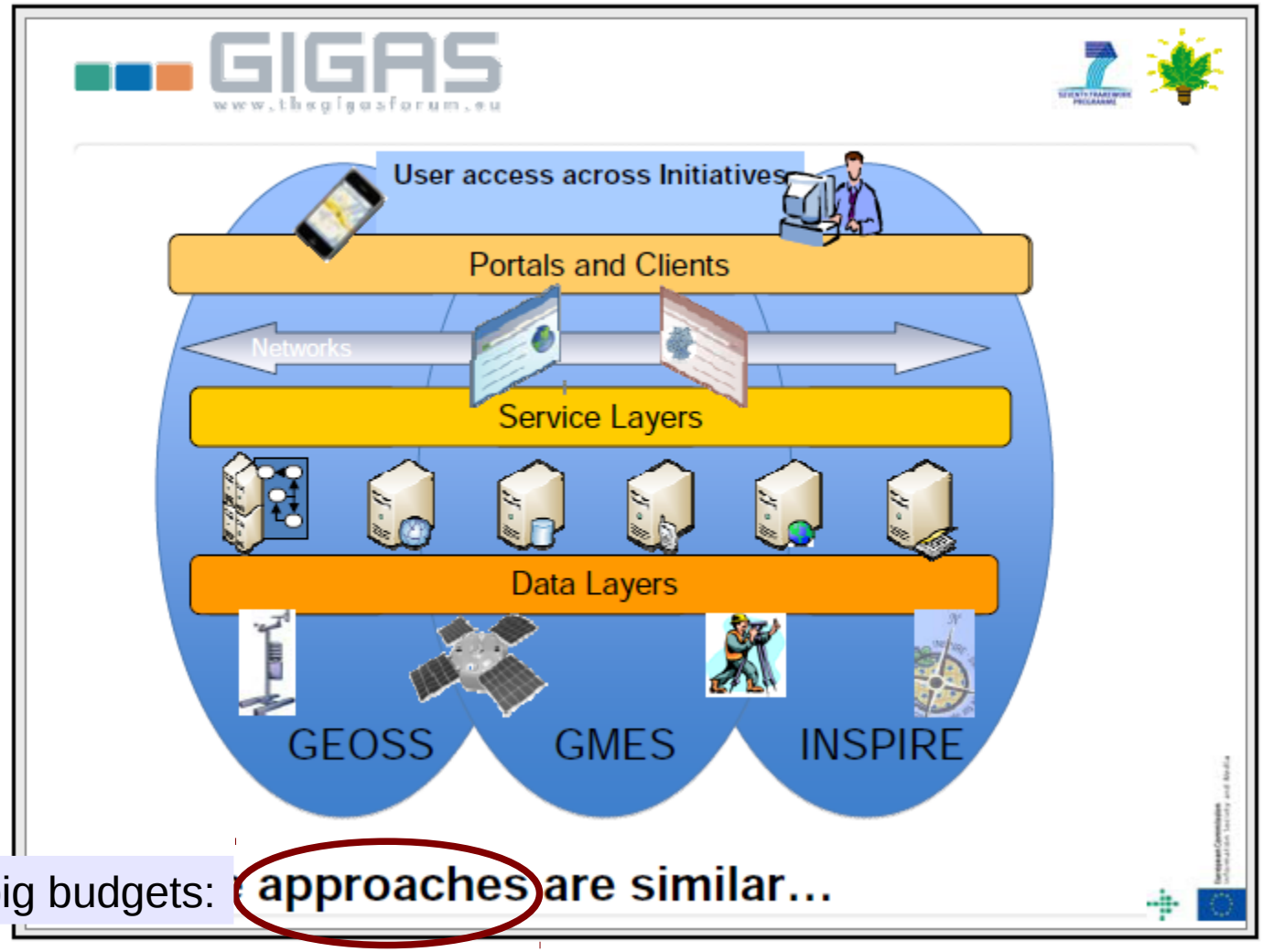


European and Global Drivers

GEOSS:
Global Earth Observation System of Systems.

GMES:
Global Monitoring for Environment and Security

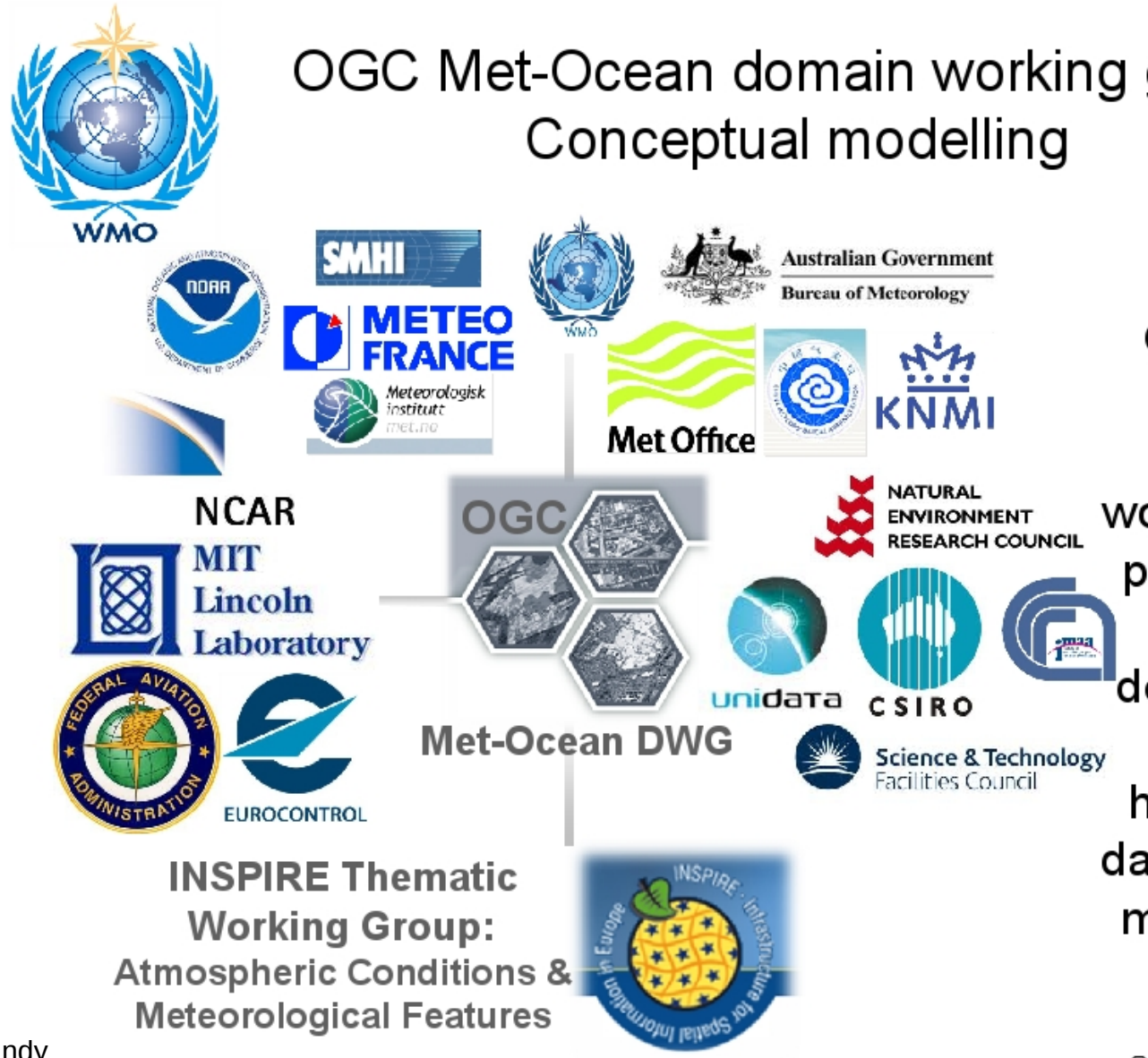
INSPIRE:
European Spatial Data Initiative Directive.



Big Projects; big budgets:

E. Klein,2010, http://www.thegigasforum.eu/cgi-bin/download.pl?f=461.pdf (Accessed, Oct, 2010)

OGC Met-Ocean domain working group: Conceptual modelling

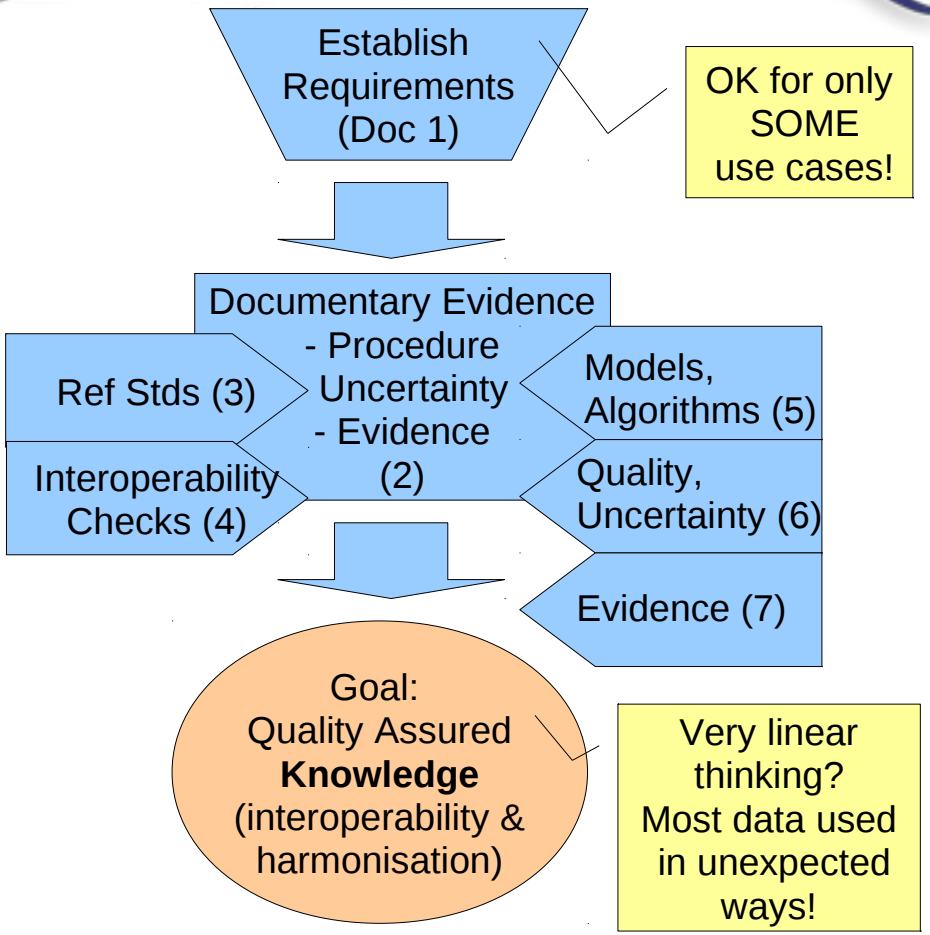


OGC Met-Ocean domain working group provides the forum for development of a harmonized data model for meteorology

Courtesy of J.Tandy (UKMO)

QA4EO

A QUALITY ASSURANCE FRAMEWORK FOR EARTH OBSERVATION



Quality Indicators: information providing a product user with sufficient information to assess its suitability for a particular application.

… based on quantative assessment of its **traceability** to an agreed **reference** or **measurement** standard. Numeric or text.

Traceability = Provenance (unbroken chain of calibrations each contributing to measurement uncertainty)

Reference/Measurement Std: defined, itself with uncertainty. Individual or community defined.

Uncertainty: characteristic of value range *dispersion*, preferably based on experimental evaluation, but if necessary on documented subjective assessments such as experience.

… but this is not enough, necessary but not sufficient: should not be considered to be a complete solution for quality attribution ...

CMIP5

CMIP5: Fifth Coupled Model Intercomparison Project

- Global community activity under the auspices of the World Meteorological Organisation (WMO) via the World Climate Research Programme (WCRP)
- Aim:
  - to address outstanding scientific questions that arose as part of the AR4 process,
  - improve understanding of climate, and
  - to provide estimates of future climate change that will be useful to those considering its possible consequences.

Ten(s) of Petabytes of data, globally distributed, several petascale "cache archives"

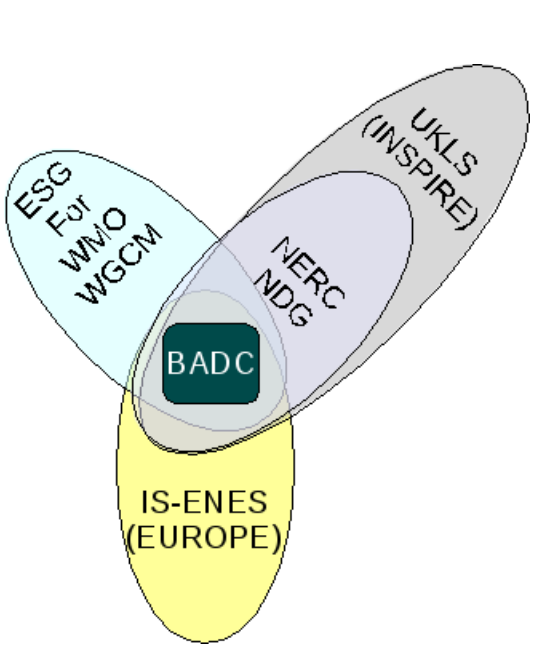
- one here in Canberra.

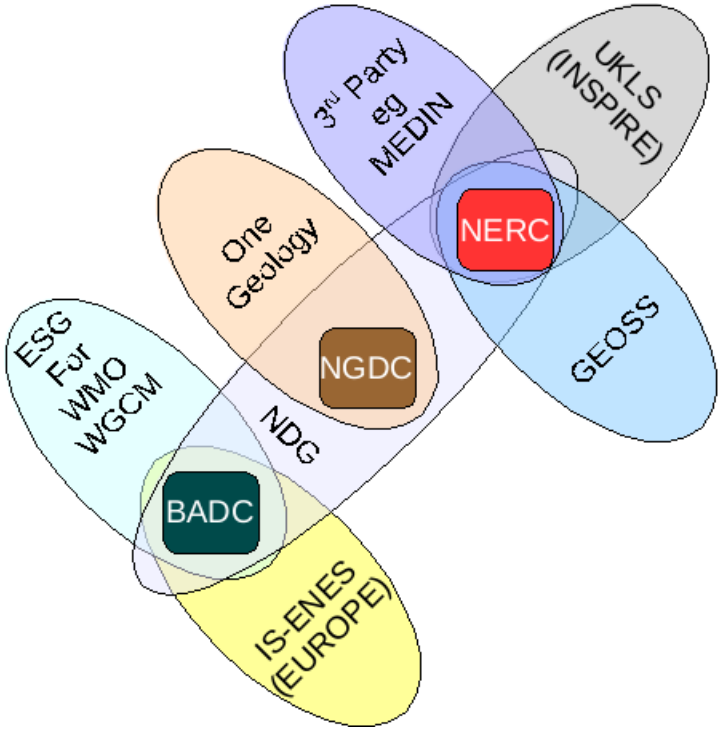
Major international projects:

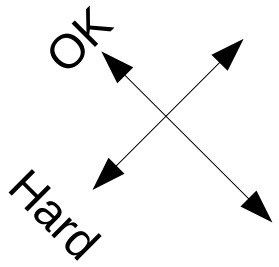
- Earth System Grid
- Earth System Grid Curator
- Metafor
- IS-ENES (InfraStructure for a European Network for Earth Simulation)
- and more in the pipline

CMIP5 SCALE, COMPLEXITY and SOCIOLOGY, DOMINATES THINKING!

Compliance: the organisational Reality







Generally muti-axis problems

Discline specific activiites exist in a world-wide domain: no national scale activity (or even, European) scale activity can exist in isolation.

Some integration efforts are doomed to failure (as currently constituted)!

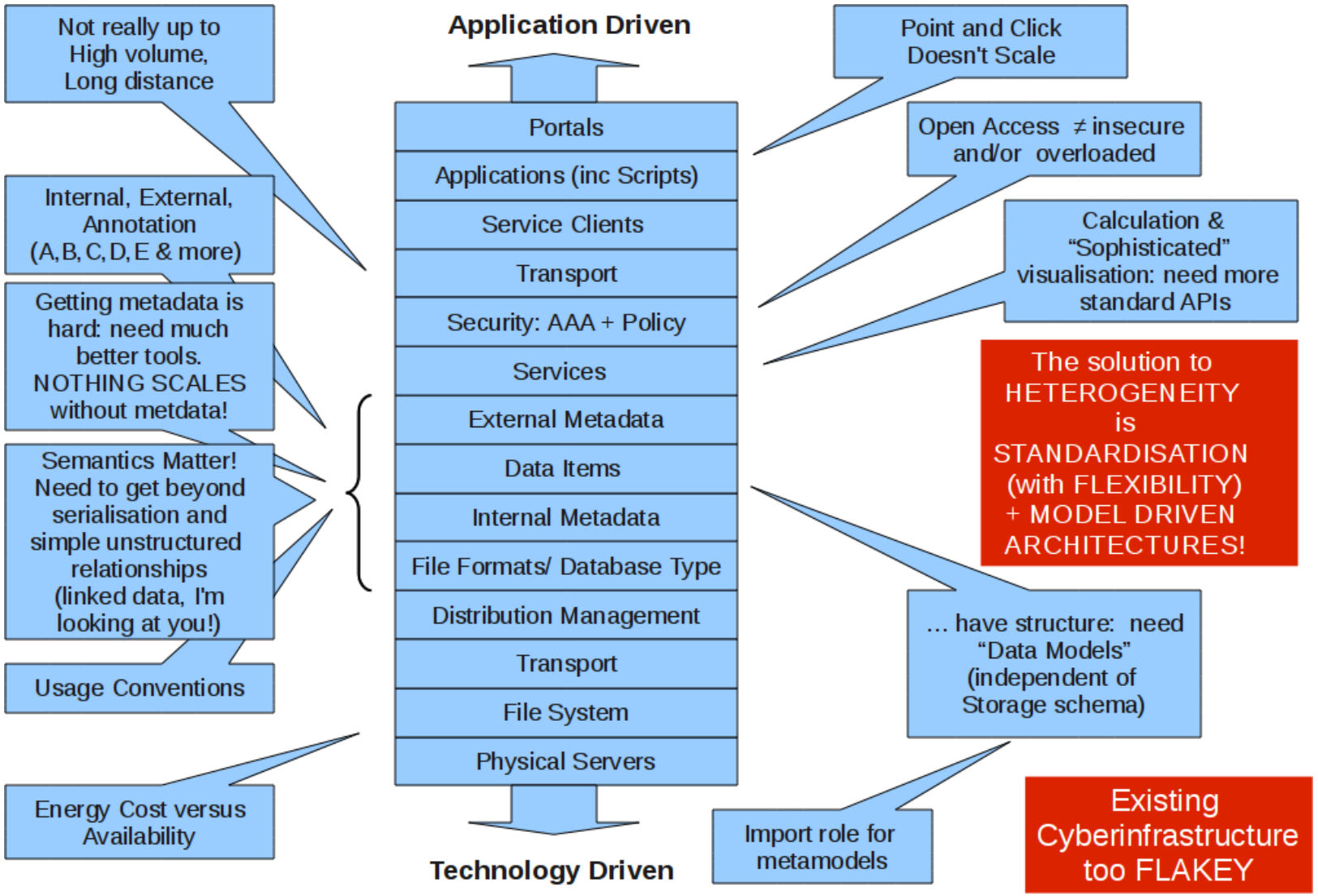
Technology Drivers

The So-Called Data Interoperabilty Components (25 or 26!)

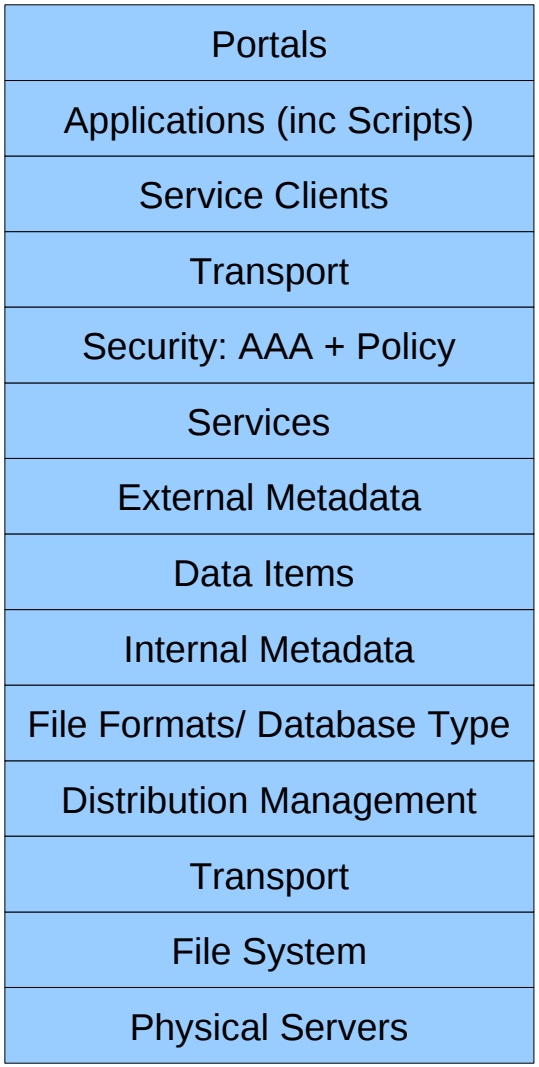
- Terminology
- Reference model
- Architectural support for data interoperability
- Rules for application schemas and feature catalogues
- Application schemas
- Ontologies
- Spatial and temporal aspects
- Coordinate referencing and units of measurement model
- Identifier management
- Object referencing
- Multi-lingual text and cultural adaptability
- Data transformation model / guidelines
- Portrayal
- Maintenance and Publication of information about geographic data (2)
- Metadata for discovery, evaluation and use
- Maintenance of data
- Data & information quality
- Delivery of data
- Consistency between data
- Multiple representations
- Data capturing rules
- Conformance
- Governance
- Extension points

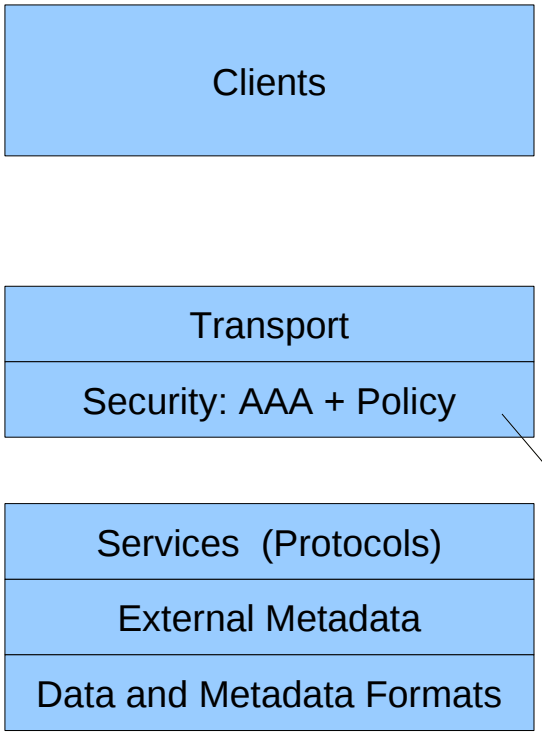
http://inspire-forum.jrc.ec.europa.eu/pg/pages/view/9782/
(Cox 2009, accessed Oct 2010)

Technology Drivers



Realistic Intervention Points





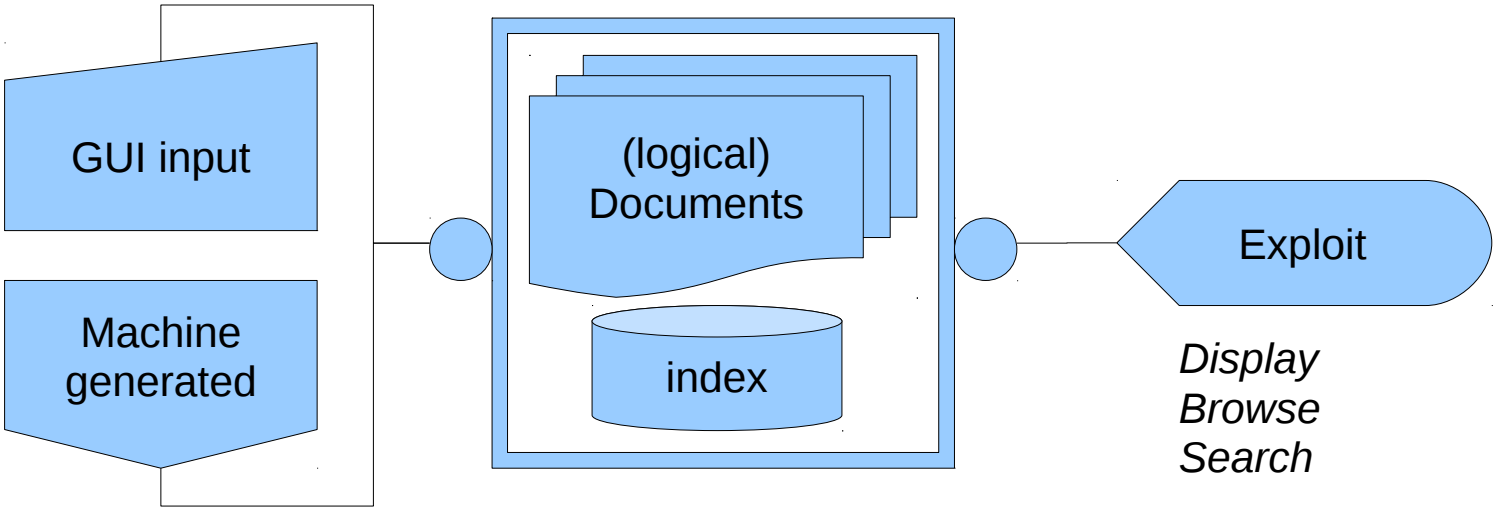
Most folk have got stuck on data description, (and that at such a high level it's almost useless to application developers and users)!

Catalogs (Data Description; Service Descriptions; Service Binding)

Nearly all interesting applications founder on this … eventually

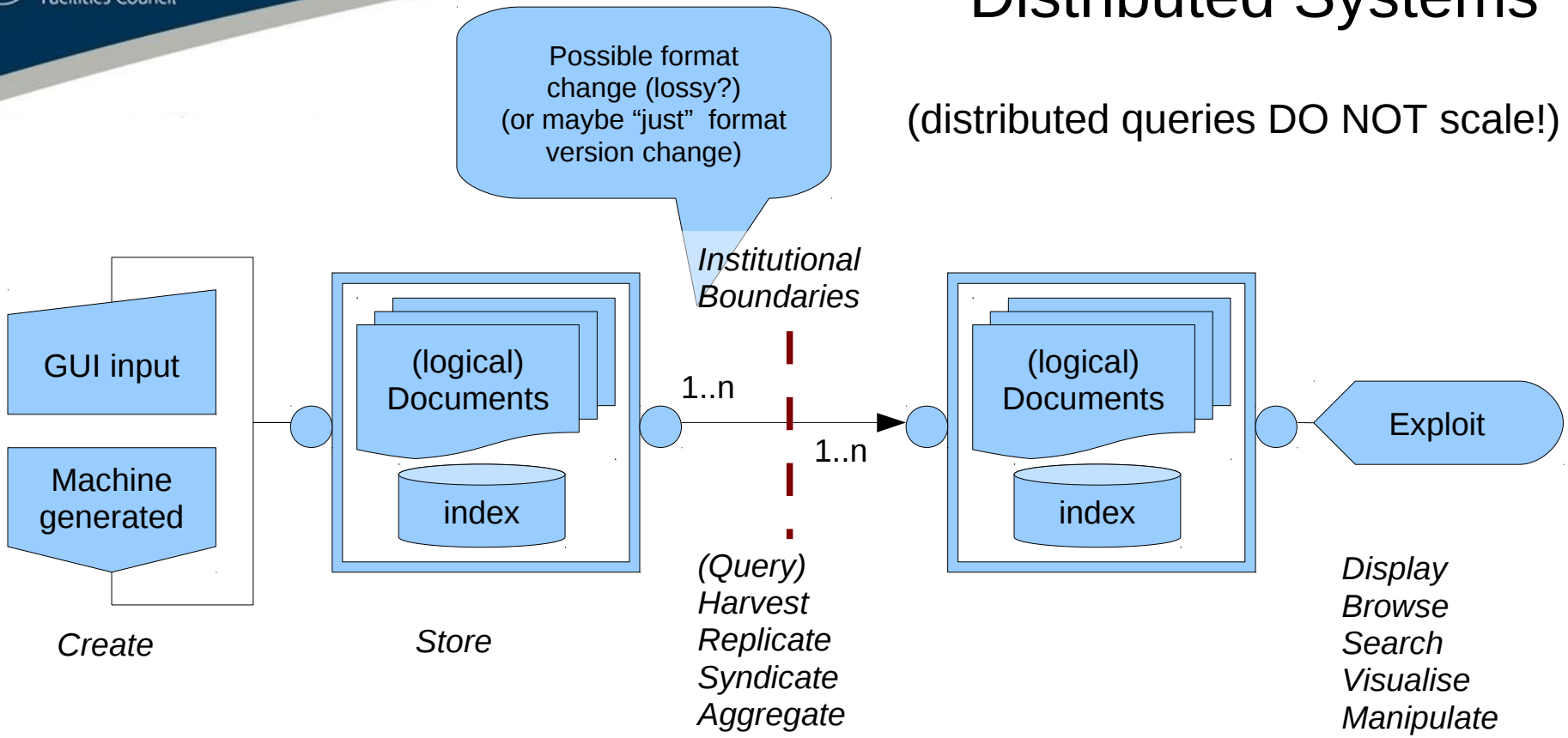
Crucial!

The beginners guide to Information Systems



Distributed Systems

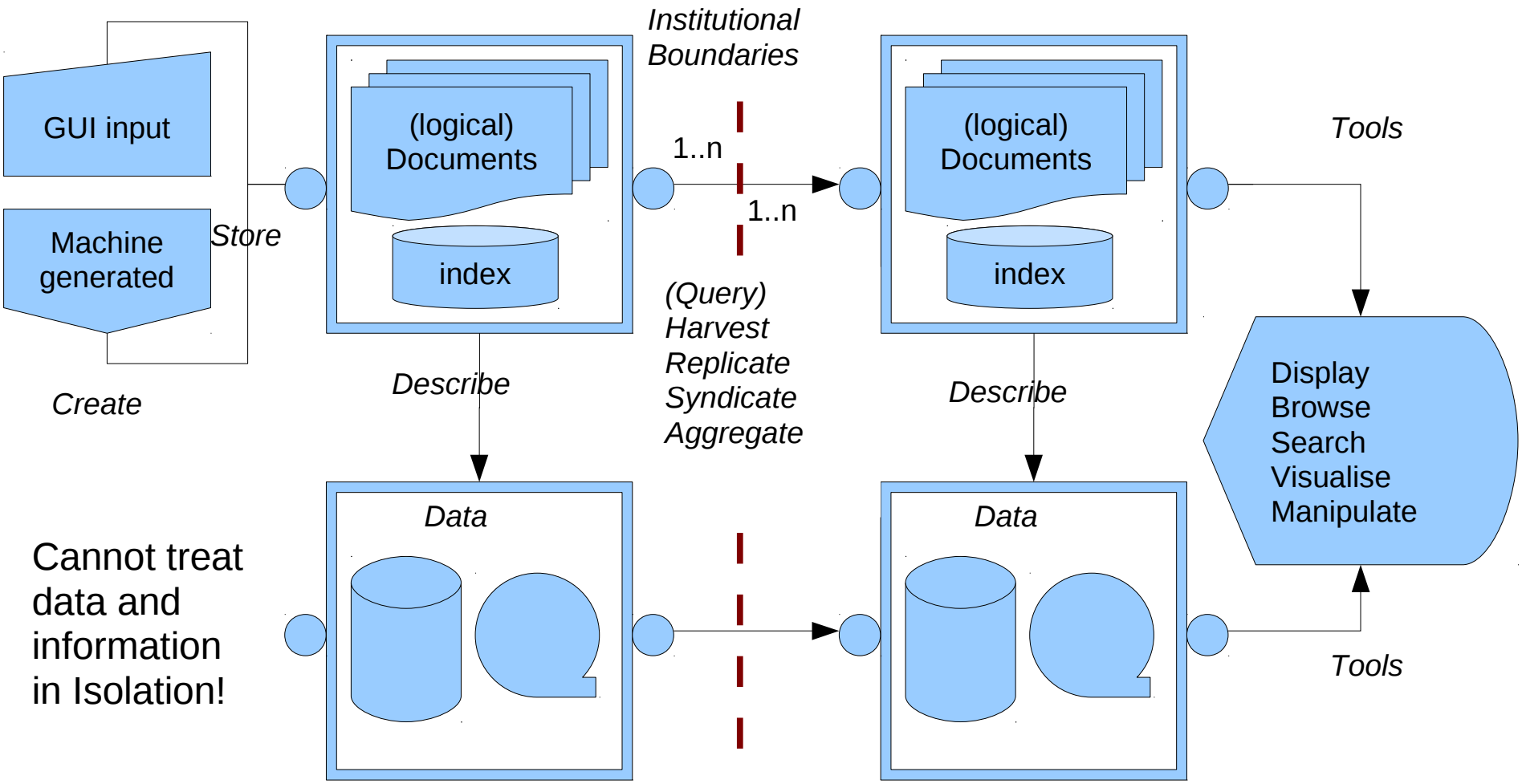
(distributed queries DO NOT scale!)



Different versions of "documents" describing the same real world objects can (and will) co-exist in this ecosystem.
These versions may differ in one or more of format, format version, and content (whether there is intended semantic format conversion or not).

Cannot ignore life cycle!

Federated Repositories



Cannot treat data and information in Isolation!

SOAP, REST and OGC

Some personal perspectives:

- Invest in SOAP, and you're investing in “big business” not in delivering interoperability.
  - European INSPIRE is making a bad mistake
  - (We have built systems with SOAP, and we don't want to, any more)
- Invest in REST, and you're building systems which can scale and expand.
  - BUT: Recognise that REST is only a very low-level framework, interoperability comes from the assumptions about vocabularies you build upon it. Still need service binding definitions!
- Current OGC services are very fragile.
  - The concepts are excellent, the implementations less so. We're all going to be engineering around these issues for some years to come (interoperability is not going to get DONE for a while yet).

Linked data and the ISO TC211 approach.

Linked Data

- "Anyone can say Anything about Any Topic"
  - Good, but "RDF cannot prevent anyone from making nonsensical or inconsistent assertions, and applications that build upon RDF must find ways to deal with conflicting sources of information"
- Much easier to build services which CONSUME linked data, if linked data built on top of controlled, versioned ontologies ...

ISO TC211

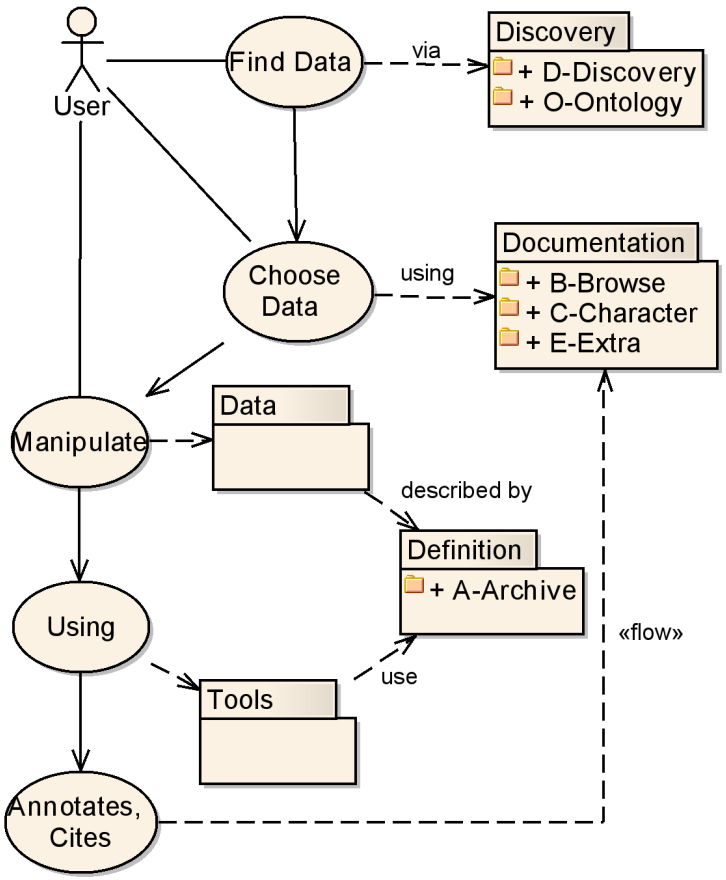
- Domain modelling formalism based on ISO19101 and friends provides controlled versioned view of the real world.
  - (May have some issues distinguishing between what is the real world and what is a description of it), but
  - Proivdes a methodology for distributed communities building consistent ontologies.

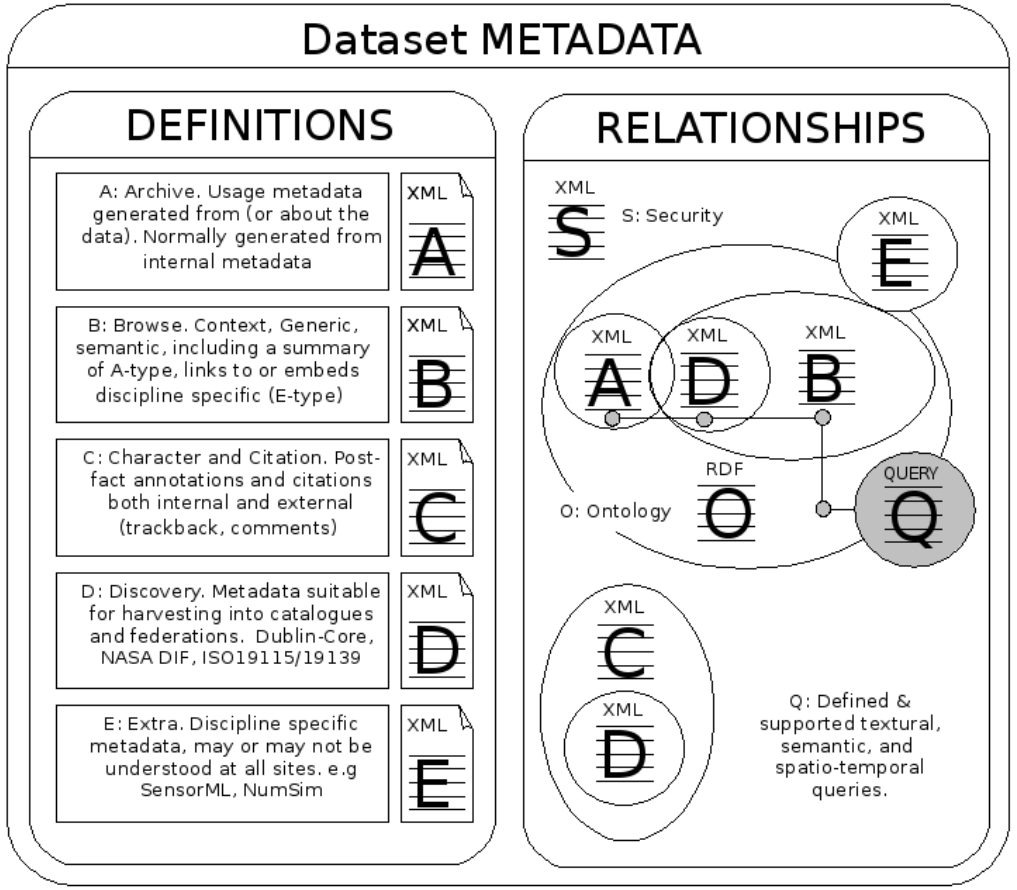
These two worlds are not incompatible. ISO TC211 information can be serialised in multiple ways, and supports EFFECTIVE use of linked data: the "re-use" mantra of the linked data world depends on the existence of controlled ontologies such as that provided by the TC211 ISO philosophy.

Some CEDA activities:

- Information Modelling,
- COWS, and
- Access Control

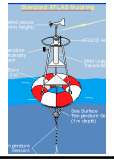
Discovery, Documentation, Definition





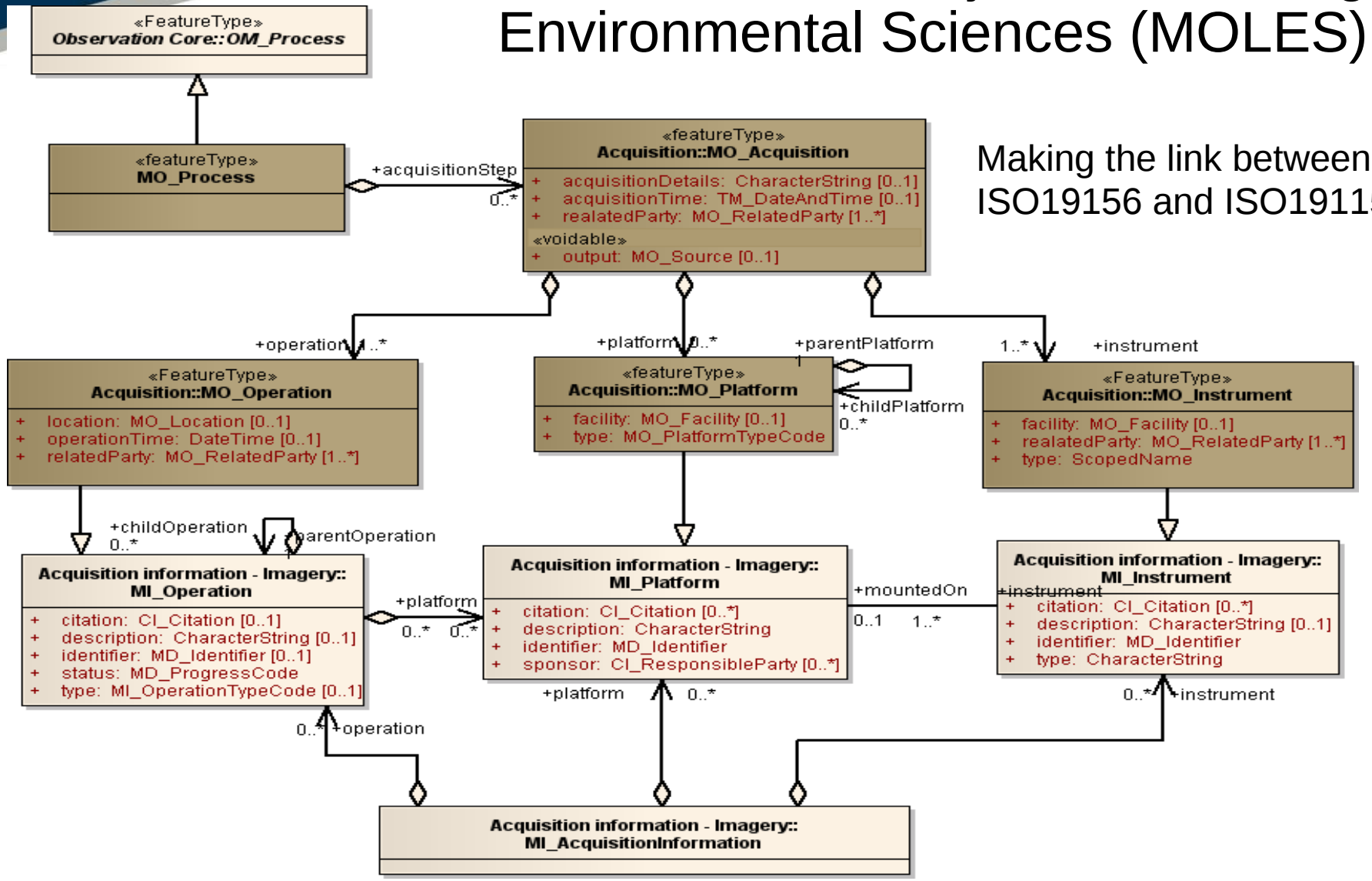
A: CSML Feature Types

(Slide courtesy of Jon Blower)

PointSeriesFeature (*timeseries at a point*)	 
ProfileFeature (*vertical profile at a point*)	  
GridSeriesFeature (*series of multidimensional grids*)	 
SwathFeature (*single satellite sweep*)	
SectionFeature (*vertical section*)	 

Classified by their geometry(not an exclusive list)
Climate Sciences Modelling Language:
influence on NetCDF CF and on the Sampling Features part of O&M

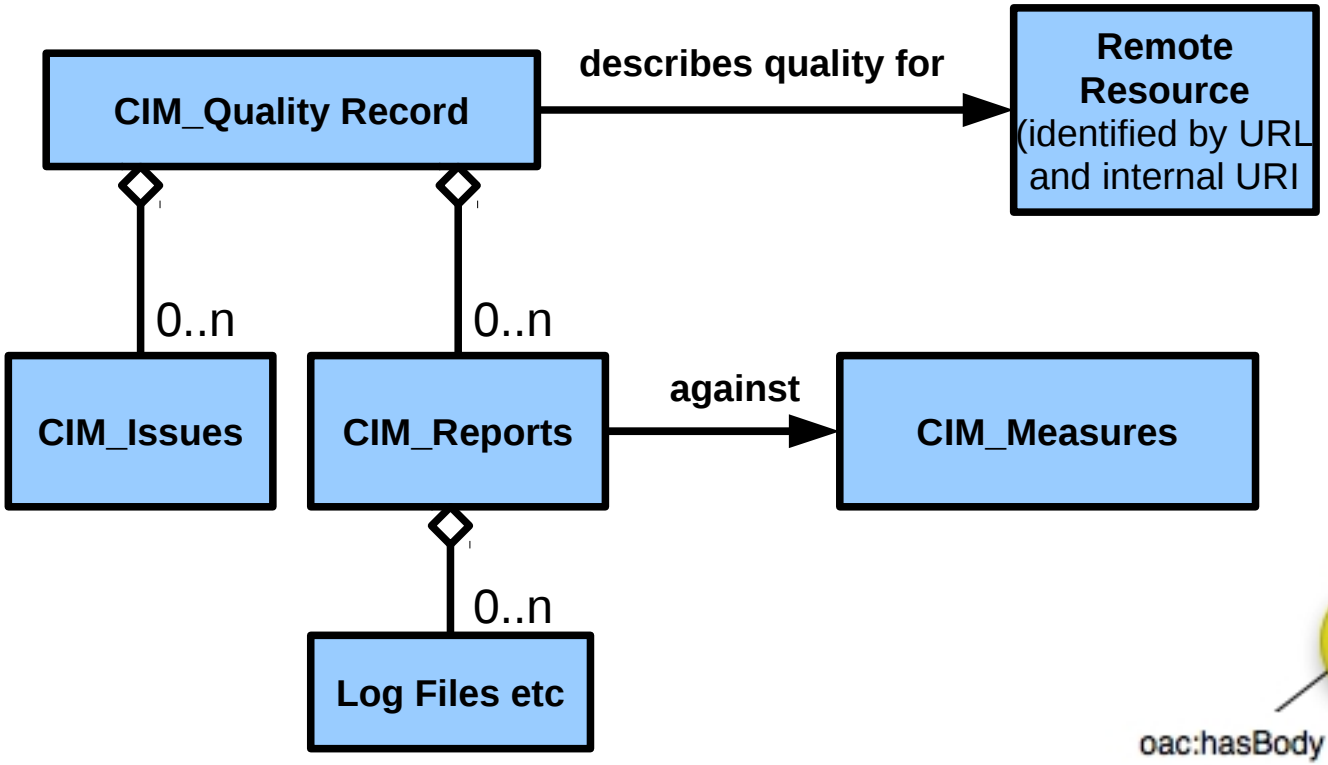
B: Metadata Objects for Linking Environmental Sciences (MOLES)



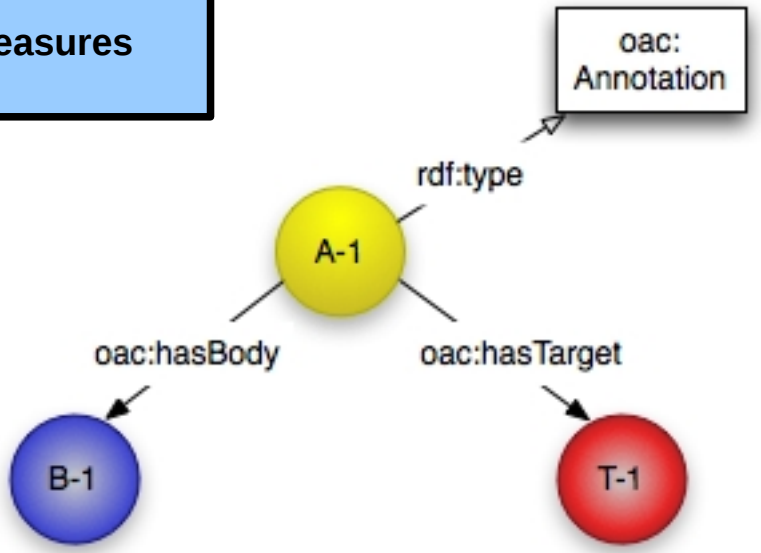
Making the link between ISO19156 and ISO19115

C: Metafor Quality Package

Specialises ISO19115 DQ package



Extrinsic!
Not intrinsic!



Maps nicely onto the upcoming Open Annotation Model (http://annotation.lanl.gov)

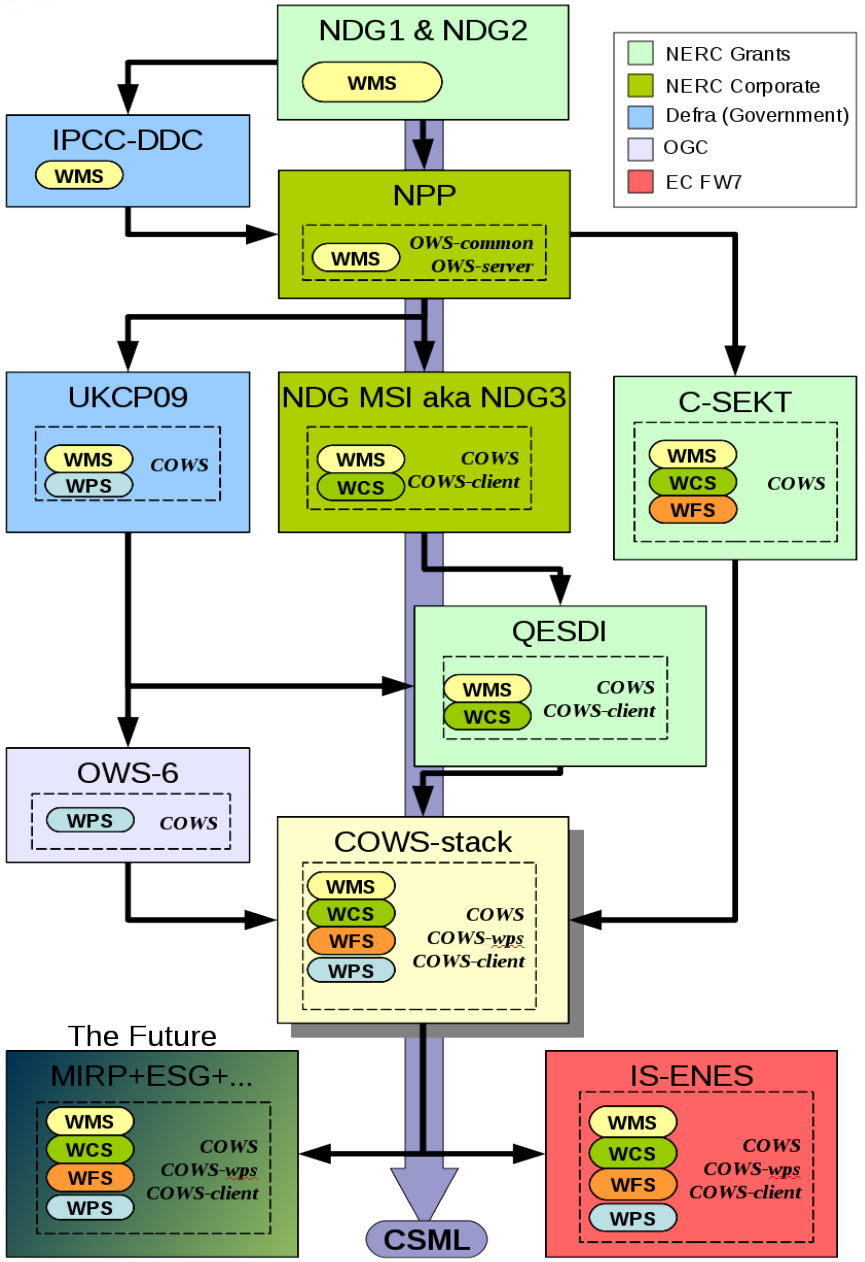
COWS

CEDA OGC Web Services

"Complete" solution for OGC webservices for 4D atmos/ocean data.

(but by making it suitable for atmos/ocean, heavily customised)

(Will include DAP stack too!)



Access Control

(Access control for COWS)

AAA: Authentication,
Authorisation,
Accounting

(Open does not mean insecure and/or overloaded)

Federation Issue:

- Which federation?
- Most organisations part of multiple federations!

Web Service Issue:

- How to wrap “calls” in a secure manner?
- How to proxy credentials through a chain of web services?

Important not to mandate solutions that ***preclude*** organisational flexibility *even as they* ***provide*** security and appropriate service levels.

(=> multiple security solutions in middleware, design from security layers from the off!)

What we've learnt about interoperability

Successful Interoperability

- Identify where (in the stack) interoperability is required by the application(s), not where it is easiest to do …
- … and limit the scope of those with whom interoperability is required:
  - The larger the group, the less there is in common!
  - (Dublin Core and common profiles of ISO19115 approach being useless, except in toys – ***on their own***.)
  - Accept that any given organisation may need to interoperate with multiple groups in different ways.
  - Internal structure need to support those links, but not mimic them.

… is mostly a social activity, not a technical activity.

- Incredibly important role of shared governance and objectives …
- … and the state of the art is still that partners are likely to have to showcase their own technological contributions
  - Nothing available off the shelf right now (and unlikely to be in my domain)

→ *ISO (and even OGC) process is **incredibly** hamstrung by actual and/or perceived secrecy!*