ECMWF workshop
Use of HPC in Meteorology
25-29 October, 2004

Integrating distributed climate data resources: NERC DataGrid

Andrew Woolf[3]

With: Bryan Lawrence[1], Roy Lowry[2], Kerstin Kleese van Dam[3], Ray Cramer[2], Marta Gutierrez[1], Siva Kondapalli[2], Sue Latham[1], Kevin O'Neill[3], Ag Stephens[1]

(1) British Atmospheric Data Centre

(2) British Oceanographic Data Centre

(3) CCLRC e-Science Centre

Outline

Introduction

Architecture

Metadata

Data model and standards

Security

Conclusions

Introduction

Perspective of *consumer* of HPC met products:

- **discovery**

  “I didn’t know Sam has a copy of that terabyte dataset, I needn’t have been keeping my own!”

  Metadata

- **access**

  “Speak to Anne when she’s back from holidays next week, she’ll know someone who will be able to give you a form to sign to request access”

  Services

- **use**

  netCDF (CF/COARDS/WOCE/QXF/...), PP, GRIB, NASA Ames, BUFR, HDF (4/5/EOS/...)

  Data

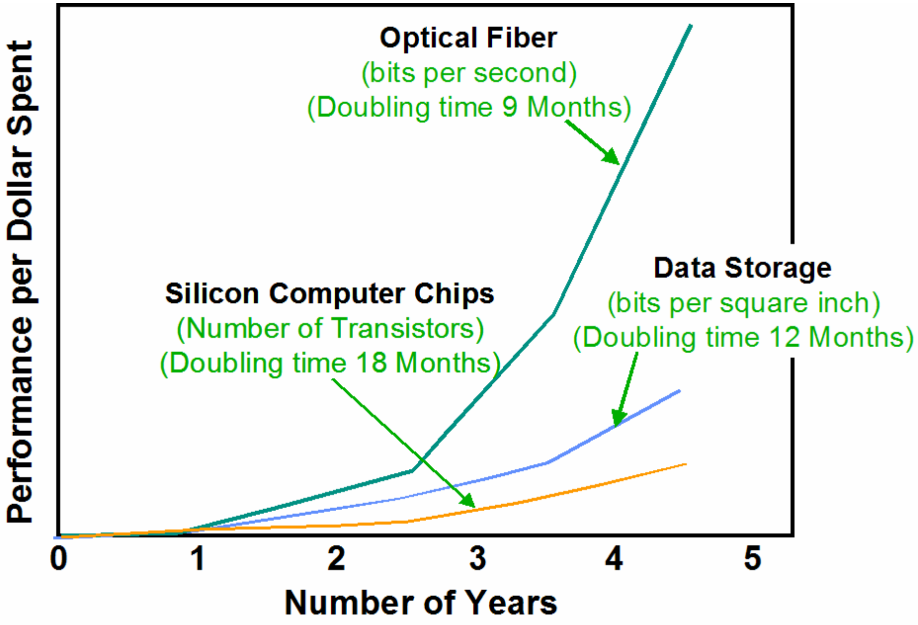
Introduction

Distributed data resources

- assimilation, verification: observations
- model intercomparison

Optical bandwidth as Grid driver

- distribution costs getting cheaper relative to cost of production



Scientific American, January 2001

Introduction

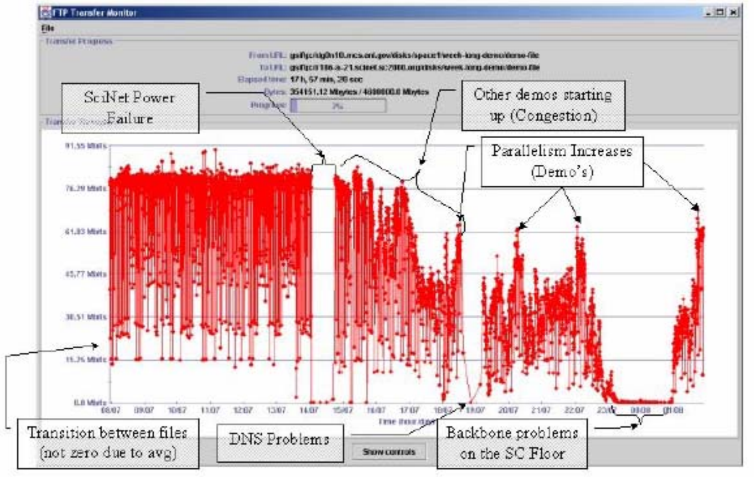
eg Earth System Grid, SC2000 data experiment Texas-California: *GridFTP*

- **2.5Gbs network (1.5 Gbs limit)**

Striped servers at source location	8
Striped servers at destination location	8
Maximum simultaneous TCP streams per server	4
Maximum simultaneous TCP streams overall	32
Peak transfer rate over 0.1 seconds	1.55 Gbits/sec
Peak transfer rate over 5 seconds	1.03 Gbits/sec
Sustained transfer rate over 1 hour	512.9 Mbits/sec
Total data transferred in 1 hour	230.8 Gbytes

multiple TCP streams, tuned TCP buffers

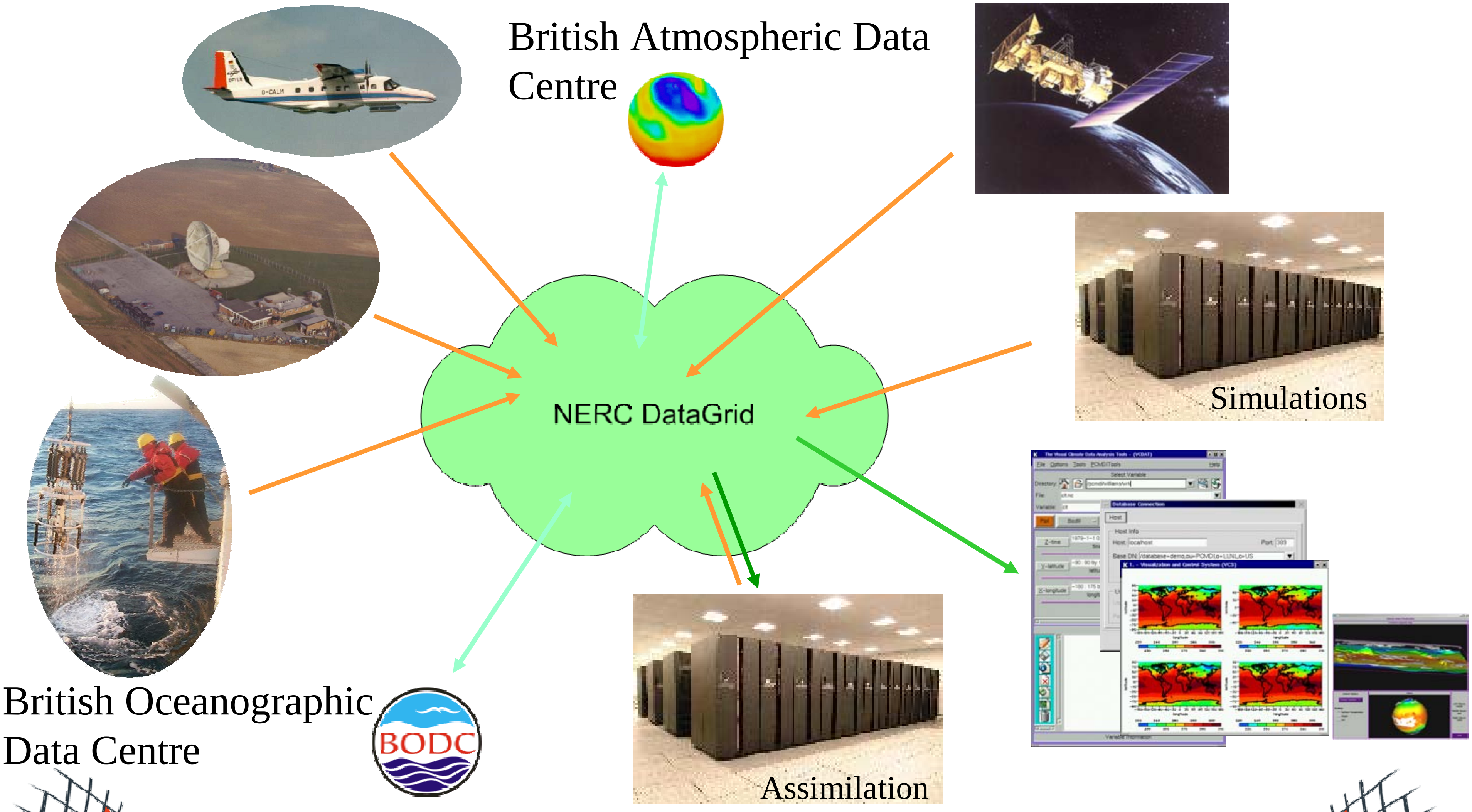
- **100Mbs NIC**



From URL: gsiftp://dg0n10.mcs.anl.gov/disks/space1/week-long-demo/demo-file
To URL: gsiftp://186-a-21.scinet.sc2000.org/disks/week-long-demo/demo-file
Elapsed time: 17 h, 57 min, 20 sec
Bytes: 354151.12 Mbytes / 4000000.0 Mbytes
Progress: 7%

High-Performance Remote Access to Climate Simulation Data: A Challenge Problem for Data Grid Technologies. B. Allcock, I. Foster, V. Nefedova, A. Chervenak, E. Deelman, C. Kesselman, J. Leigh, A. Sim, A. Shoshani, B. Drach, D. Williams. *SC 2001*, November 2001.

Introduction



British Atmospheric Data Centre
NERC DataGrid
Simulations
Assimilation
British Oceanographic Data Centre
BODC

Introduction

British Atmospheric Data Centre

British Oceanographic Data Centre

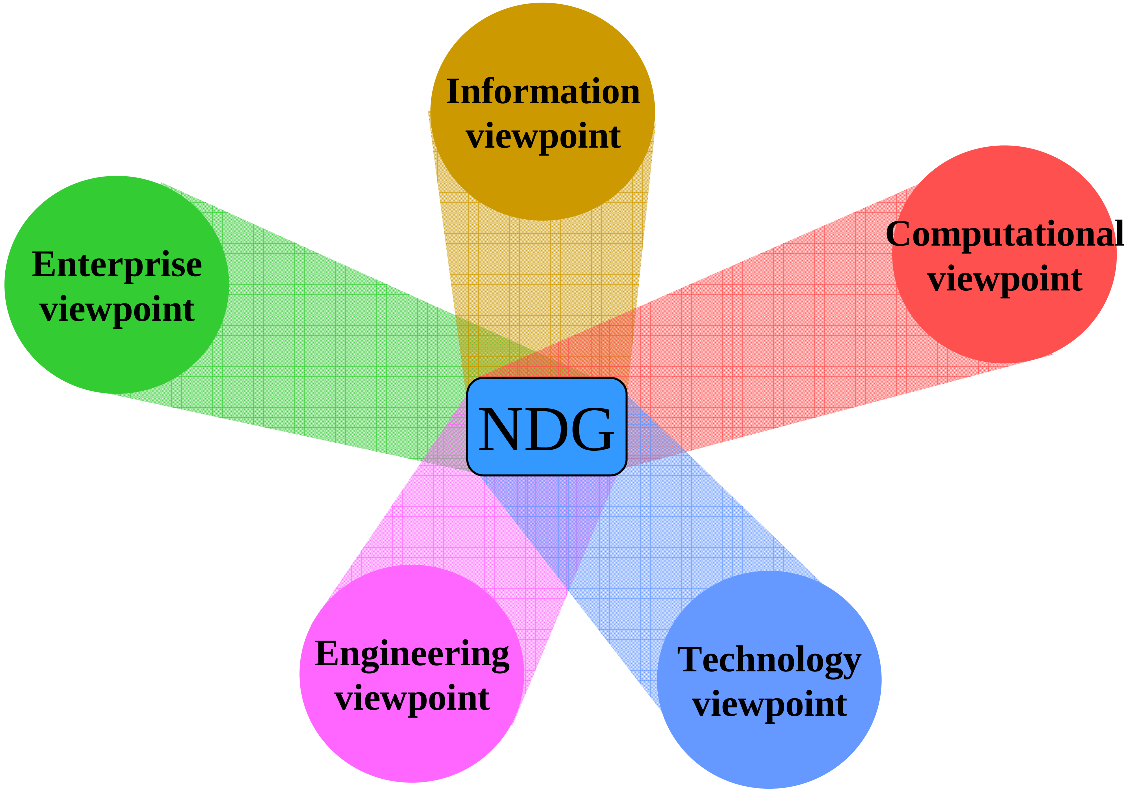
CCLRC e-Science Centre

Program for Climate Model Diagnosis and Intercomparison (LLNL), *EarthSystemGrid*

Architecture

Reference Model for Open Distributed Processing (RM-ODP)

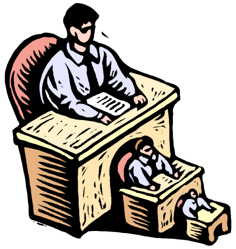
- ISO 10746-{1,2,3,4}
- Formal architecture methodology for distributed systems
- Viewpoints approach



Architecture

Enterprise viewpoint

- roles, activities, policies (incl. VO)

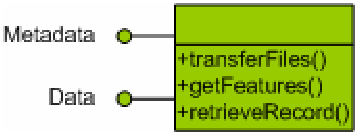


Information viewpoint

- semantics of information and information processing (static, invariant, dynamic schema)

Computational viewpoint

- interfaces and computational objects (cf. CORBA IDL, WSDL portTypes)

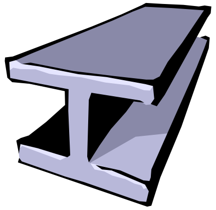


Engineering viewpoint

- distribution infrastructure (e.g. web services, WSRF vs OGSI)

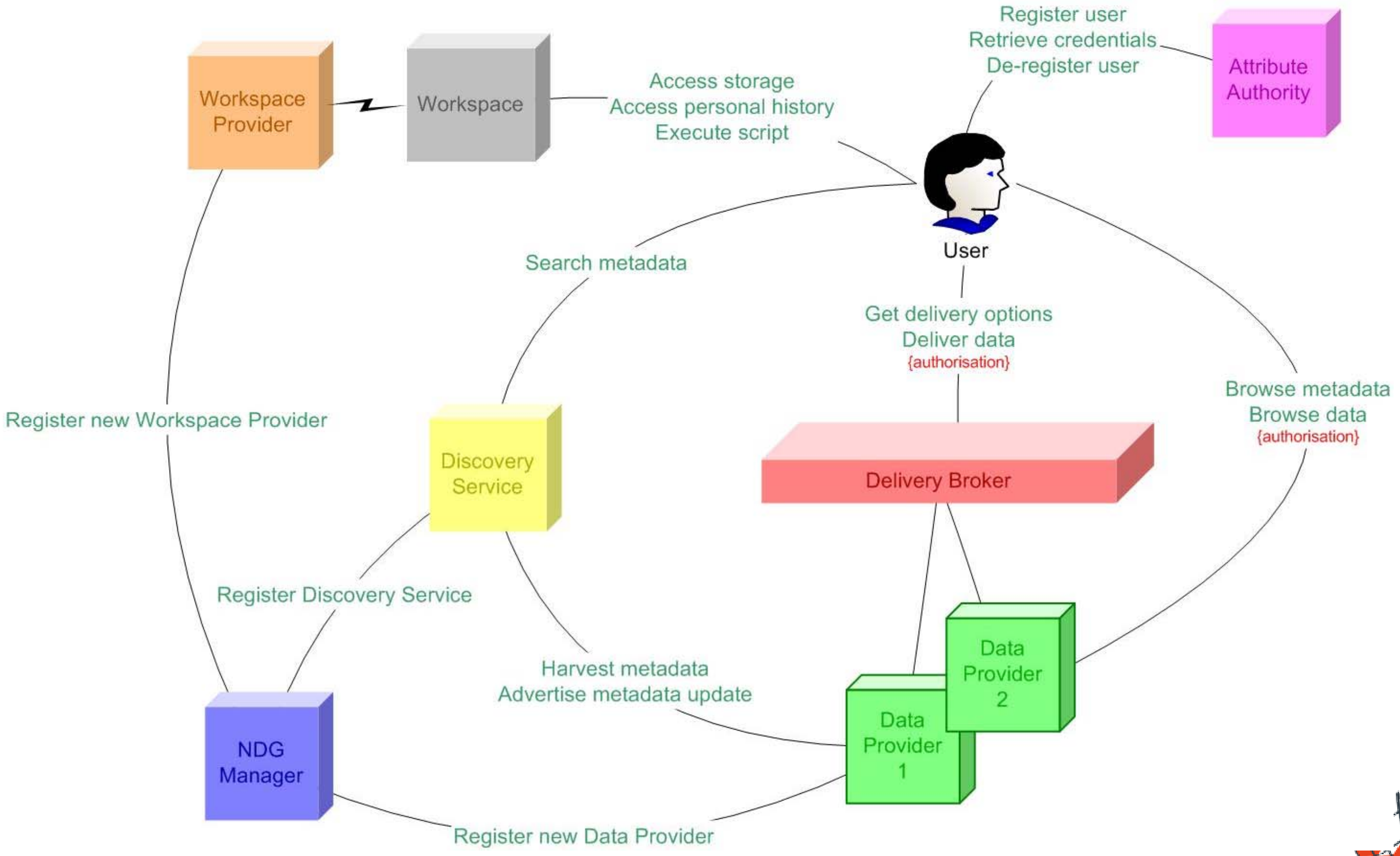
Technology viewpoint

- choices of technology (e.g. app servers, DBMS)



Architecture

Enterprise viewpoint

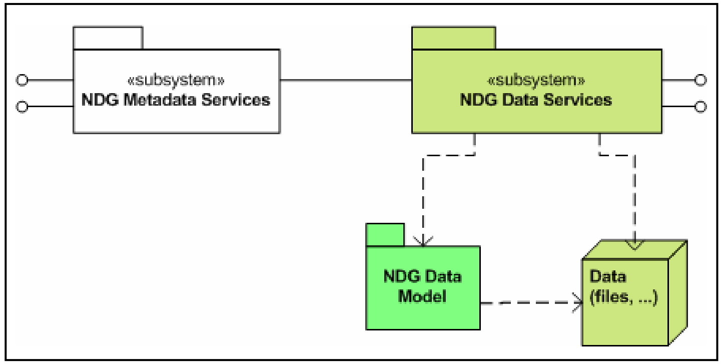


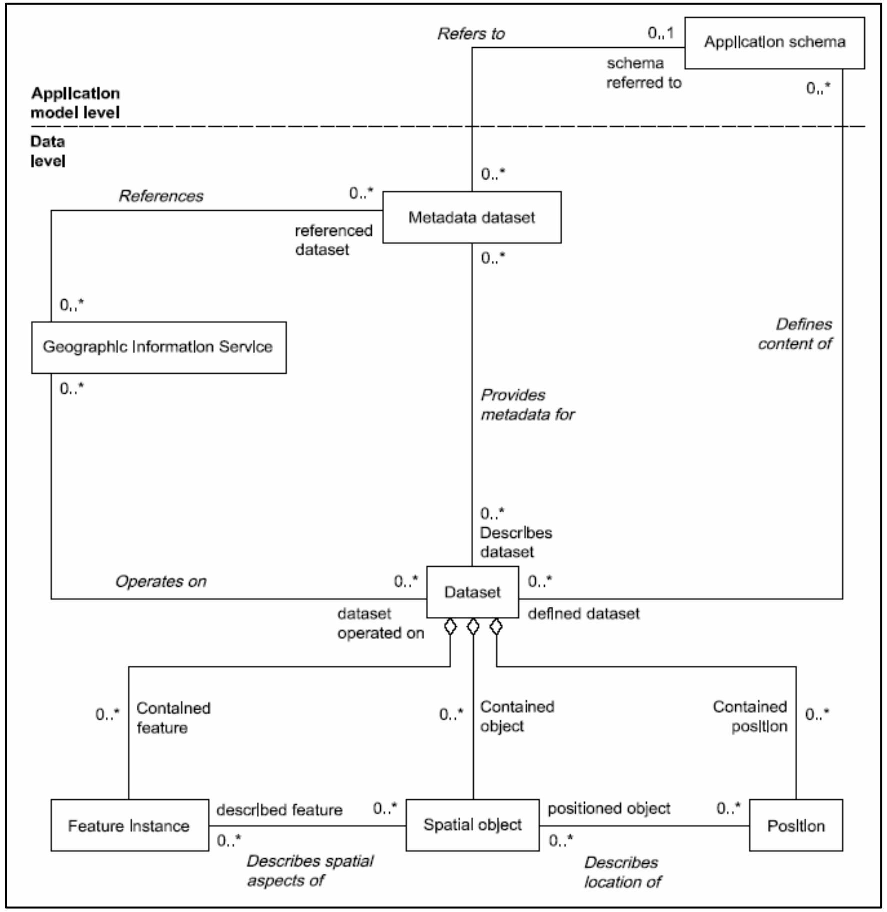
Architecture

Information viewpoint

- Conformant to ISO TC211 Domain Reference Model:

  "standardisation in the field of digital geographic information"

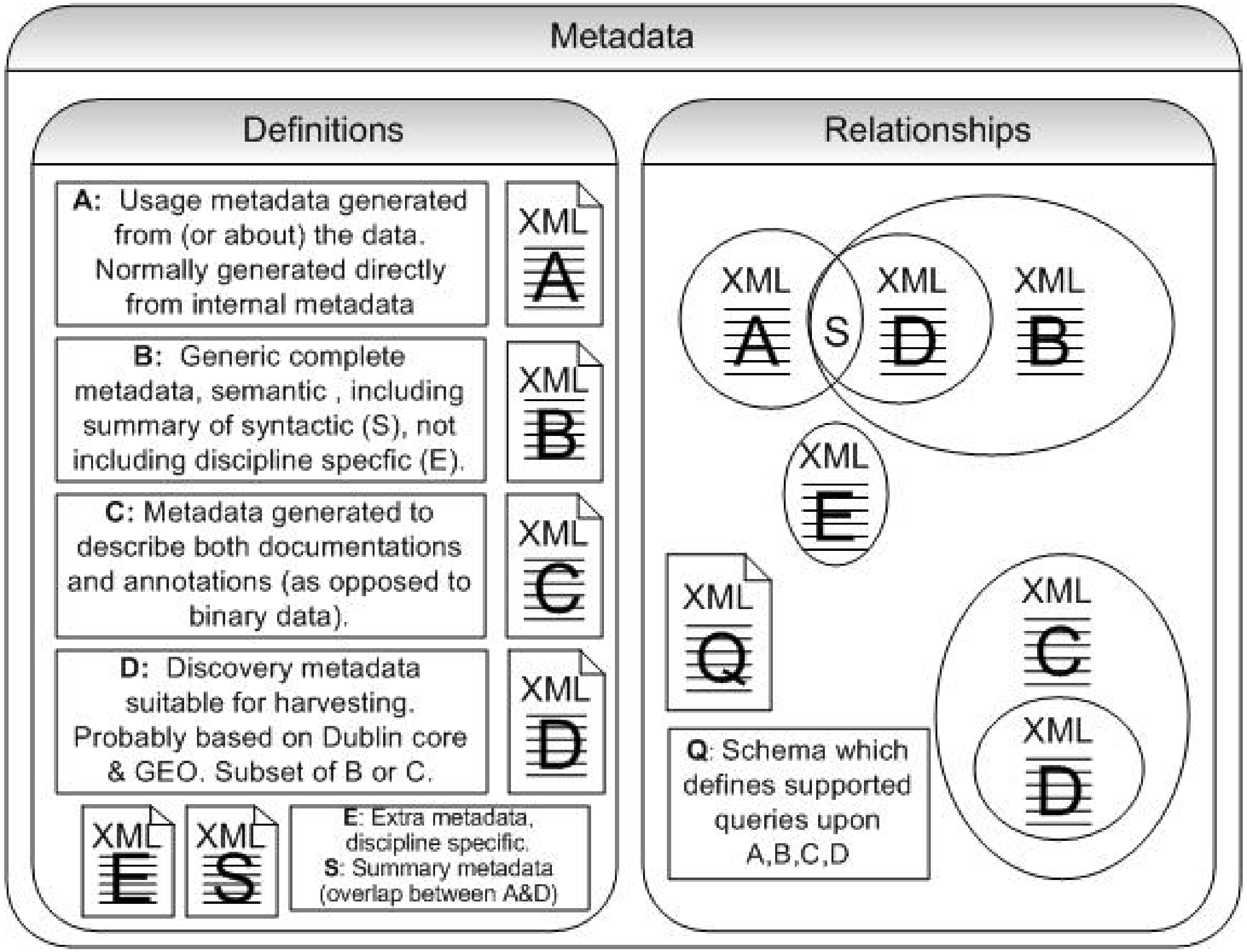




ISO 19101 Domain Reference Model

Metadata

NDG metadata taxonomy



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

B: Generic complete metadata, semantic , including summary of syntactic (S), not including discipline specfic (E).

C: Metadata generated to describe both documentations and annotations (as opposed to binary data).

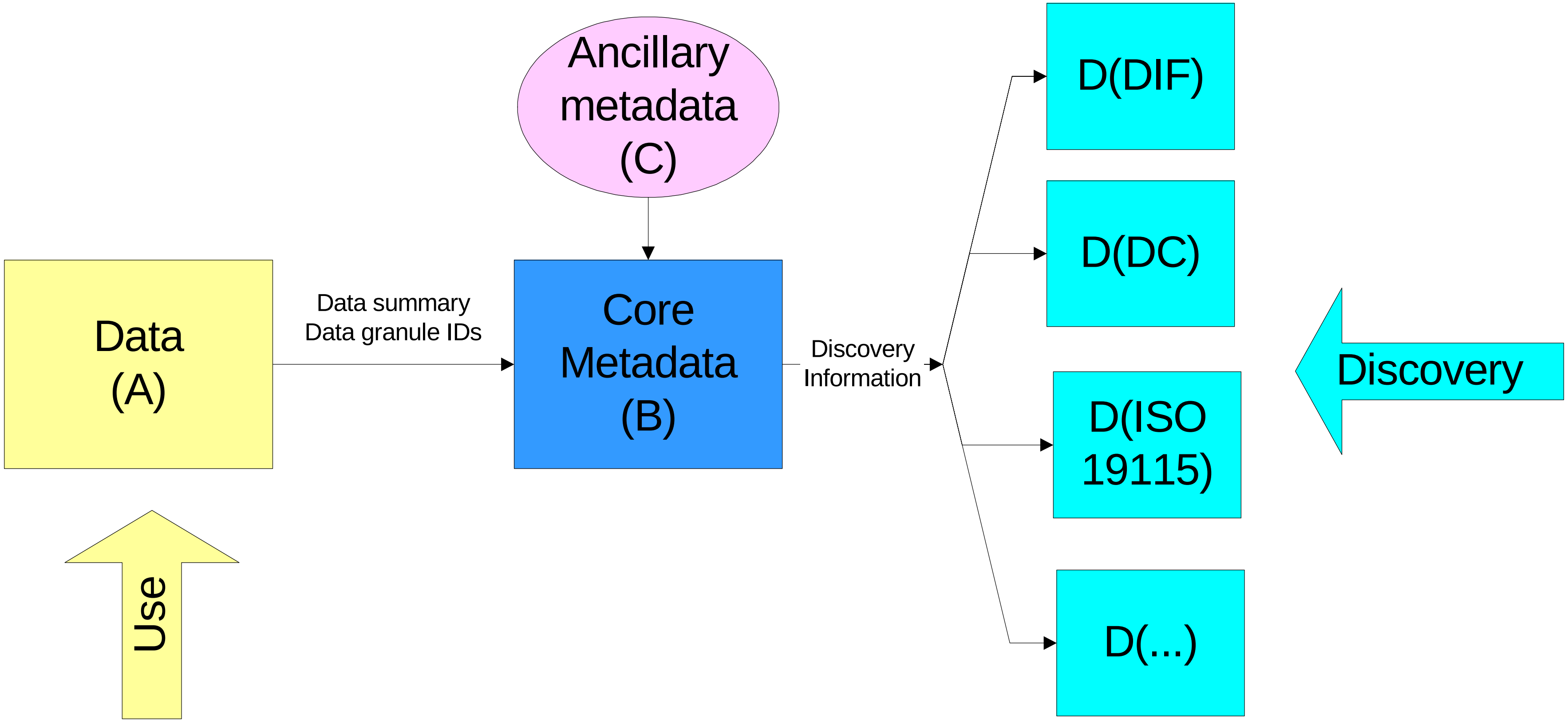
D: Discovery metadata suitable for harvesting. Probably based on Dublin core & GEO. Subset of B or C.

E: Extra metadata, discipline specific.
S: Summary metadata (overlap between A&D)

Q: Schema which defines supported queries upon A,B,C,D

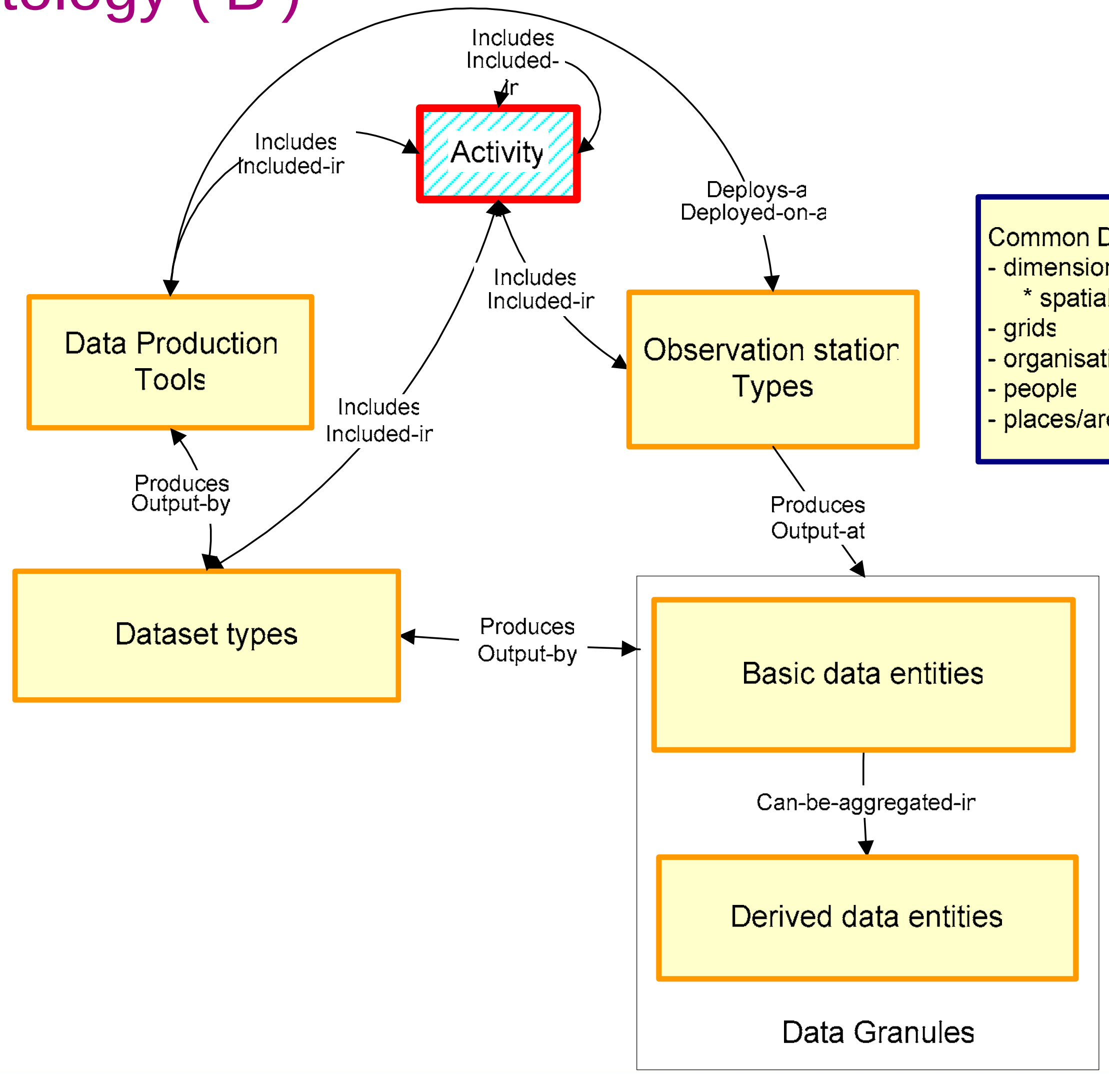
Metadata

NDG metadata taxonomy



Metadata

Domain ontology ('B')

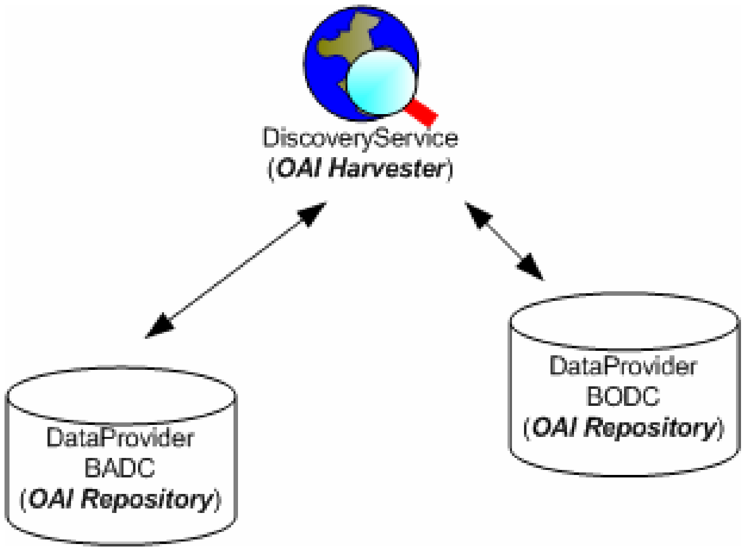


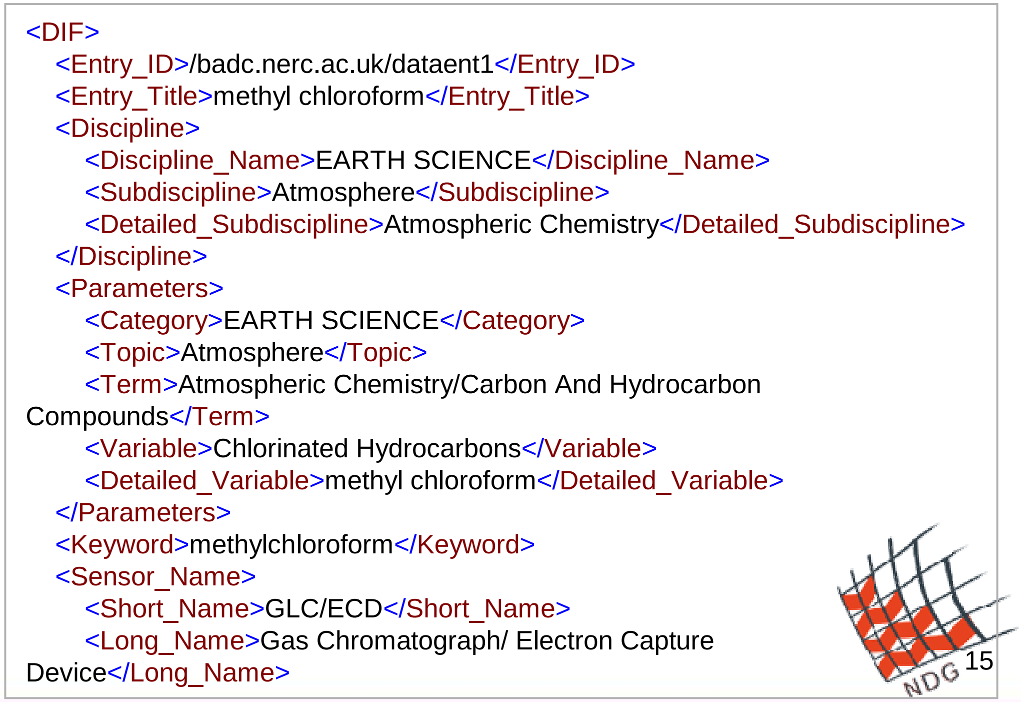
Common Data Entities
- dimensions,
  * spatial/tempora
- grids
- organisations
- people
- places/areas

Metadata

Metadata federation

- Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)
- http://www.openarchives.org




```
<DIF>
   <Entry_ID>/badc.nerc.ac.uk/dataent1</Entry_ID>
   <Entry_Title>methyl chloroform</Entry_Title>
   <Discipline>
      <Discipline_Name>EARTH SCIENCE</Discipline_Name>
      <Subdiscipline>Atmosphere</Subdiscipline>
      <Detailed_Subdiscipline>Atmospheric Chemistry</Detailed_Subdiscipline>
   </Discipline>
   <Parameters>
      <Category>EARTH SCIENCE</Category>
      <Topic>Atmosphere</Topic>
      <Term>Atmospheric Chemistry/Carbon And Hydrocarbon
Compounds</Term>
      <Variable>Chlorinated Hydrocarbons</Variable>
      <Detailed_Variable>methyl chloroform</Detailed_Variable>
   </Parameters>
   <Keyword>methylchloroform</Keyword>
   <Sensor_Name>
      <Short_Name>GLC/ECD</Short_Name>
      <Long_Name>Gas Chromatograph/ Electron Capture
Device</Long_Name>
```

Data model

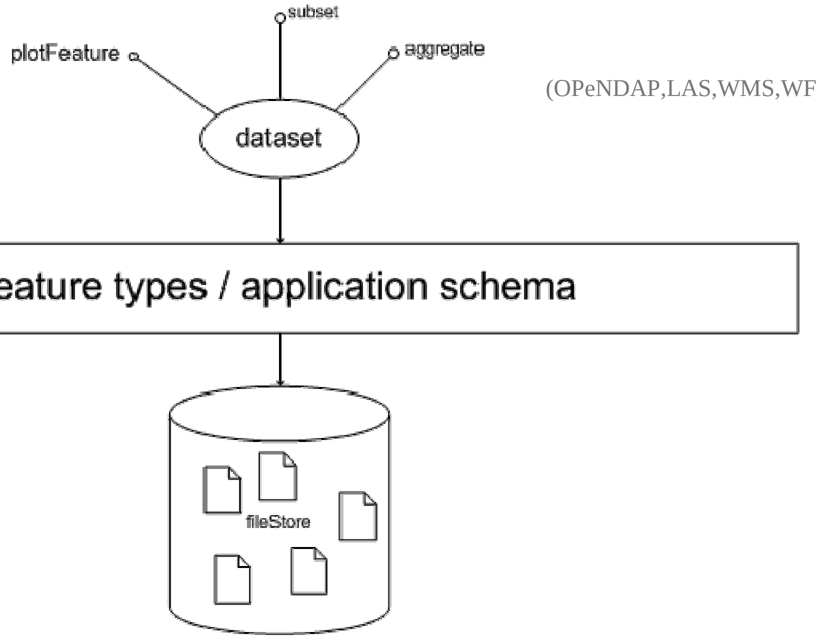
“Feature” (may be type or instance)

- *...abstraction of real world phenomenon...*

“Application schema”

- *...logical structure and semantic content of dataset...*

Offers semantically-rich abstraction layer



(OPeNDAP,LAS,WMS,WFS,WCS,...)

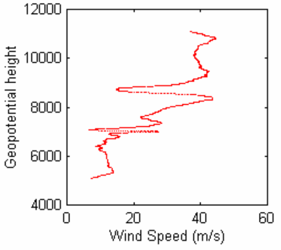
Data model

Feature type principles:

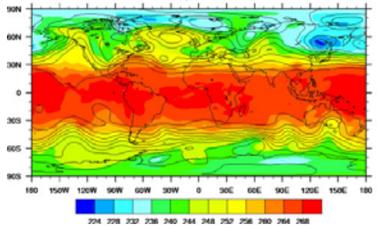
- offload semantics onto parameter type, CRS
- 'sensible plotting' as useful discriminant

Climate Science Modelling Language (CSML)

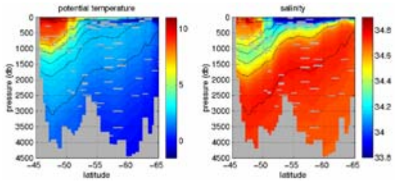
ProfileFeature

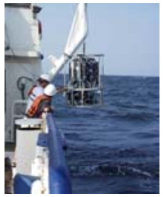


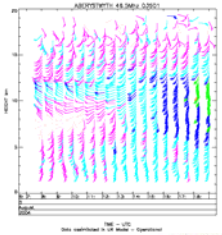
GridFeature

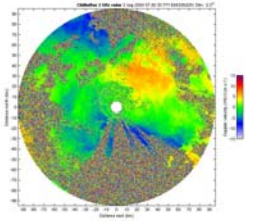


ProfileSeriesFeature



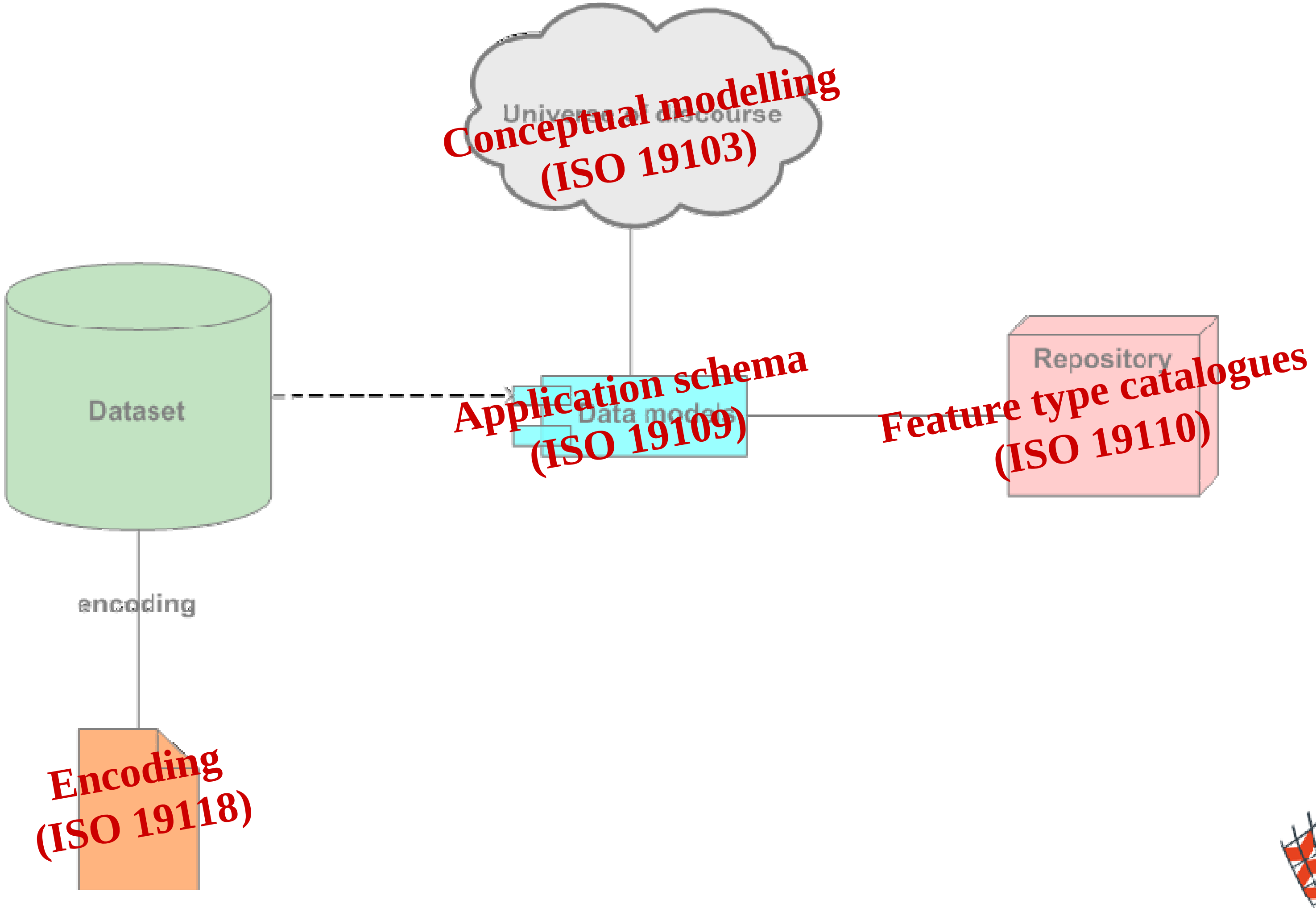






Standards

Data modelling for interoperability



Standards

ISO TC211 projects:

- **19101 (15046-1): Geographic information - Reference model**
- 19102 (15046-2): Geographic information - Overview (Project deleted, see resolution 192 - Adelaide)
- **19103 (15046-3): Geographic information - Conceptual schema language**
- **19104 (15046-4): Geographic information - Terminology**
- **19105 (15046-5): Geographic information - Conformance and testing**
- **19106 (15046-6): Geographic information - Profiles**
- **19107 (15046-7): Geographic information - Spatial schema**
- **19108 (15046-8): Geographic information - Temporal schema**
- **19109 (15046-9): Geographic information - Rules for application schema**
- **19110 (15046-10): Geographic information - Feature cataloguing methodology**
- **19111 (15046-11): Geographic information - Spatial referencing by coordinates**
- **19112 (15046-12): Geographic information - Spatial referencing by geographic identifiers**
- **19113 (15046-13): Geographic information - Quality principles**
- **19114 (15046-14): Geographic information - Quality evaluation procedures**
- **19115 (15046-15): Geographic information - Metadata**
- **19116 (15046-16): Geographic information - Positioning services**

Standards

Open Geospatial Consortium (OGC)

- International consortium of nearly 300 companies, government agencies and universities participating in a consensus process to develop publicly available geoprocessing specifications
- close liaison with ISO TC211

Specifications:

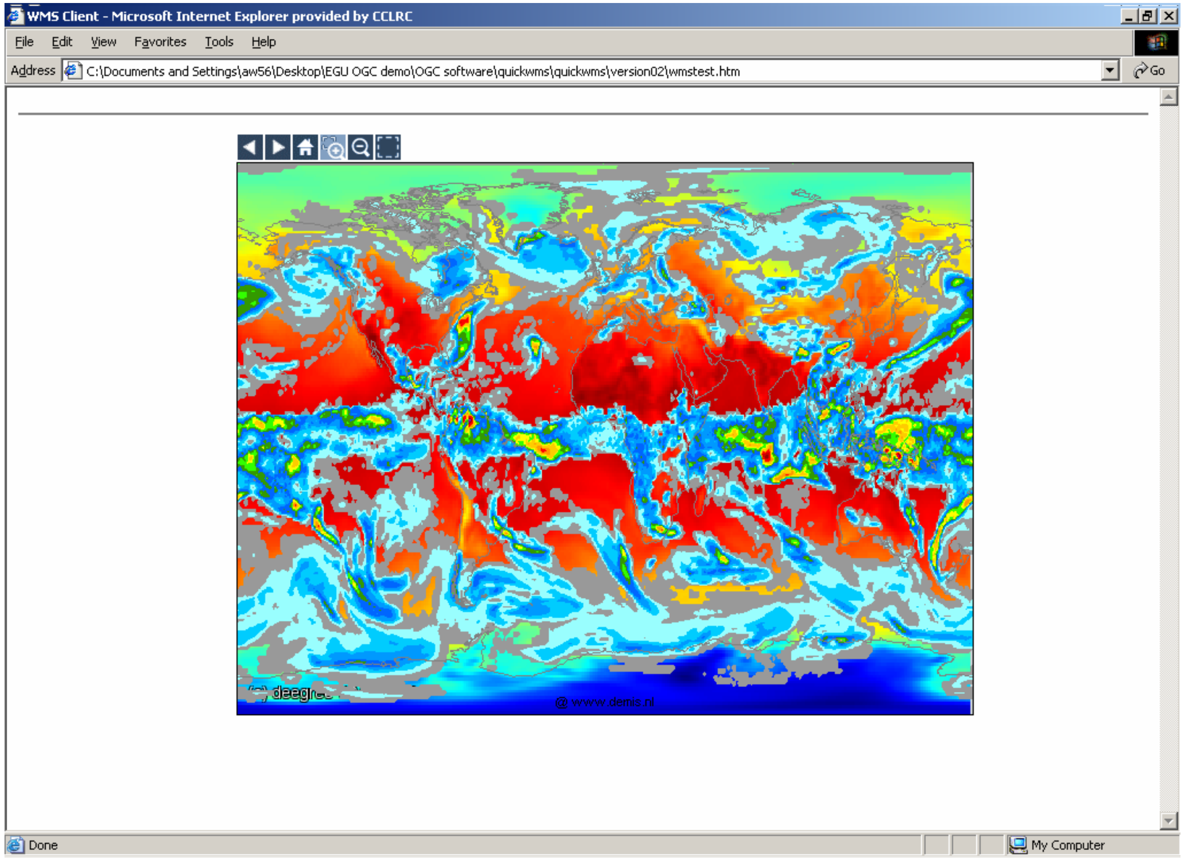
- Web Map Server (ISO 19128)
- Web Feature Server
- Web Coverage Server
- Geography Markup Language (ISO 19136)

Grid-WG, OWS 2/3

Standards

e.g.: ERA40 re-analysis surface air temperature, 2001-04-27

- deegree open-source WMS modified with netCDF connector


WMS Client - Microsoft Internet Explorer provided by CCLRC
File Edit View Favorites Tools Help
Address
C:\Documents and Settings\aw56\Desktop\EGU OGC demo\OGC software\quickwms\quickwms\version02\wmstest.htm
Go
deegree
© www.demis.nl
Done
My Computer

Overlaid with rainfall from globe.digitalearth.gov WMS server

Security

Authentication:

- x.509 PKI

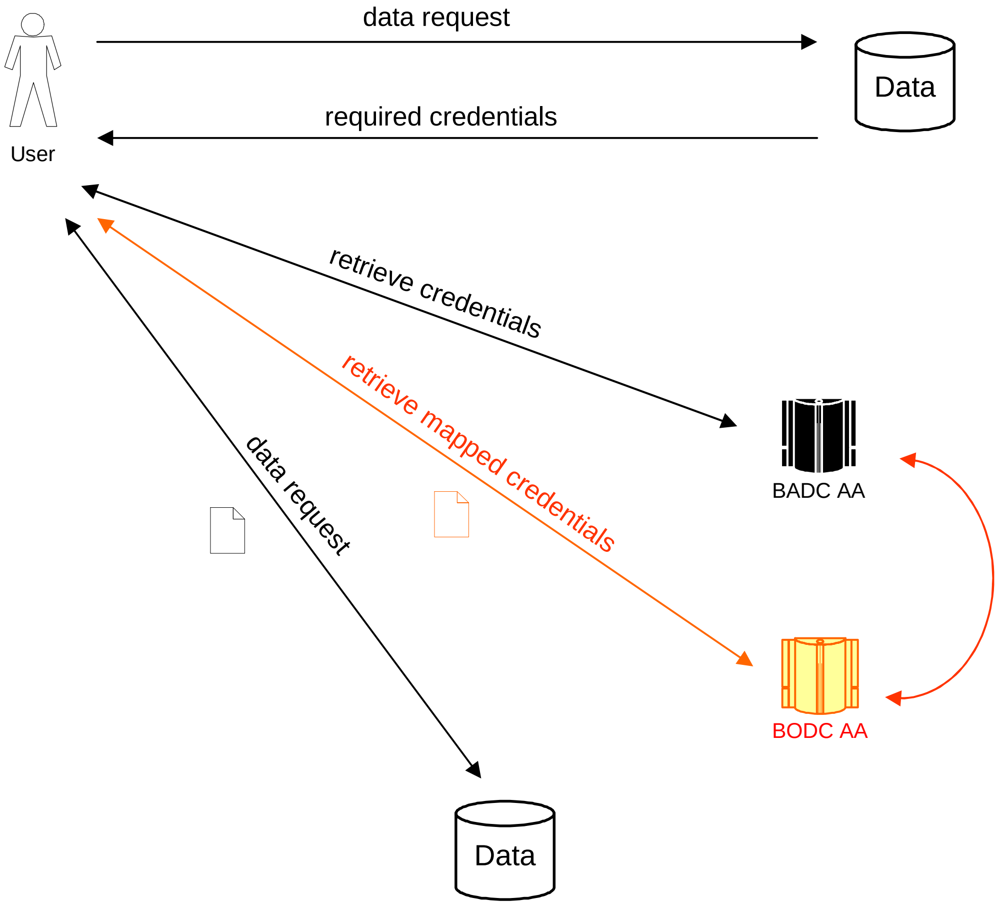
Accounting:

- server logs

Authorisation:

- role-based
- Attribute Certificates
- multiple Attribute Authorities

Security



BADC/MPIM

- investigating shared access to respective ERA-40 archives for authorised users

Conclusions

Integration of distributed climate data resources:

- discovery
- different interfaces
- diversity of formats
- security

standards

Future:

- integrated Grid infrastructure for modelling and data (inc. obs!)