Data integration with the Climate Science Modelling Language

Andrew Woolf[1], Bryan Lawrence[2], Roy Lowry[3], Kerstin Kleese van Dam[1], Ray Cramer[3], Marta Gutierrez[2], Siva Kondapalli[3], Susan Latham[2], Dominic Lowe[2], Kevin O'Neill[1], Ag Stephens[2]

[1]CCLRC e-Science Centre
[2]British Atmospheric Data Centre
[3]British Oceanographic Data Centre

Outline

Background

Standards – a framework for interoperability

Climate Science Modelling Language (CSML)

Perspectives

Background

Data integration requirements:

- scalability across providers
- warehousing not an option
- enhance access and use, 'outwards-facing' (e.g. impacts community, policymakers)
- storage heterogeneity

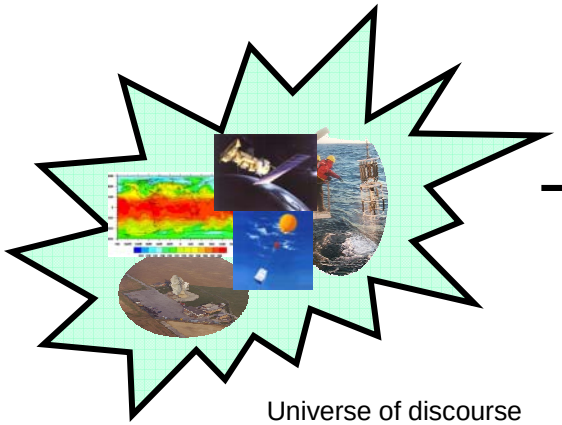
→ Semantics as integration 'key'

- **fundamentally, an information community is *defined* by shared semantics**
- common language across providers (and users)
- supports wrapper/mediator architecture

Standards

Emerging ISO standards

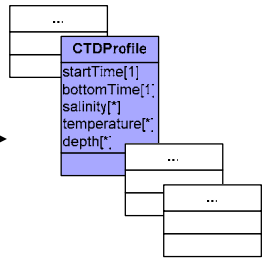
- TC211 – around 40 standards for geographic information
- Cover activity spectrum: discovery → access → use
- Provide a *framework* for data integration



Universe of discourse

ISO 19103

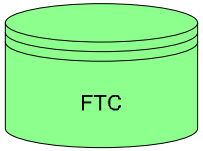
Feature types



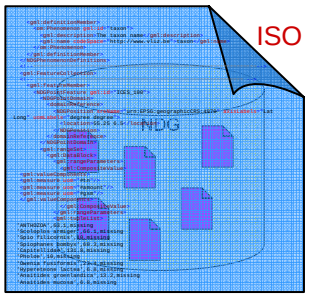
ISO 19109

Application schema

ISO 19110



ISO 19118



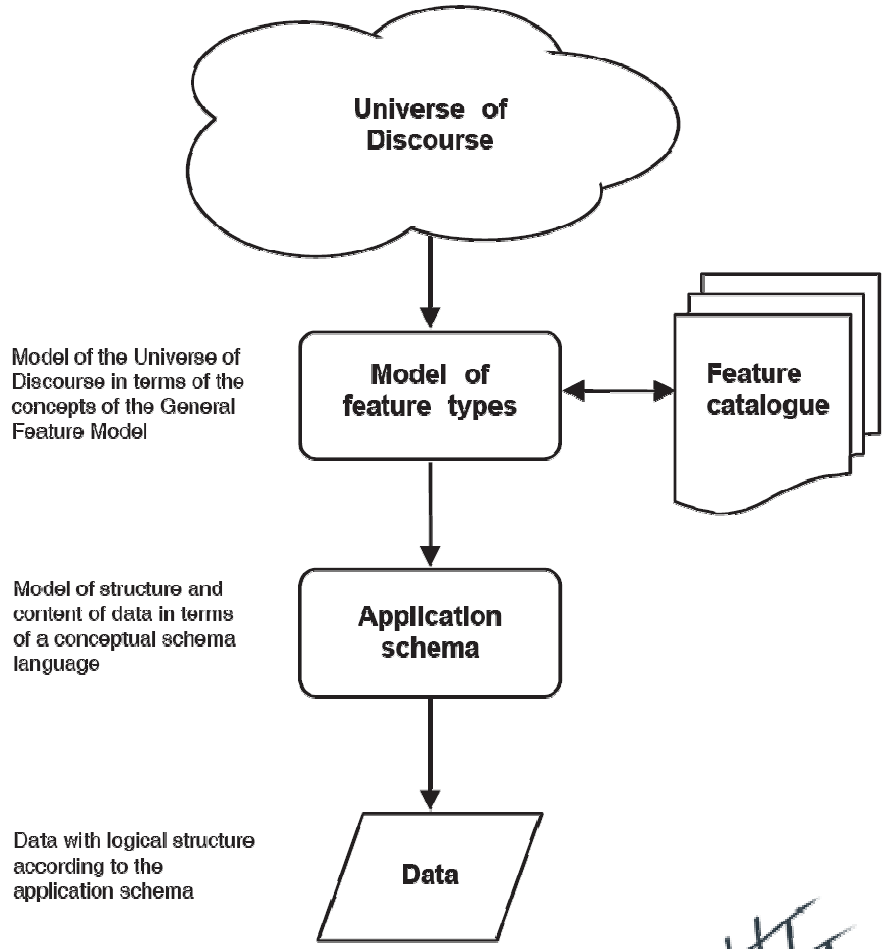
Standards

Geographic 'features'

- "abstraction of real world phenomena" [ISO 19101]
- Type or instance
- Encapsulate important semantics in universe of discourse

Application schema

- Defines semantic content and logical structure of datasets
- ISO standards provide toolkit:
  - spatial/temporal referencing
  - geometry (1-, 2-, 3-D)
  - topology
  - dictionaries (phenomena, units, etc.)
- GML – canonical encoding



[from ISO 19109 "Geographic information – Rules for Application Schema"]

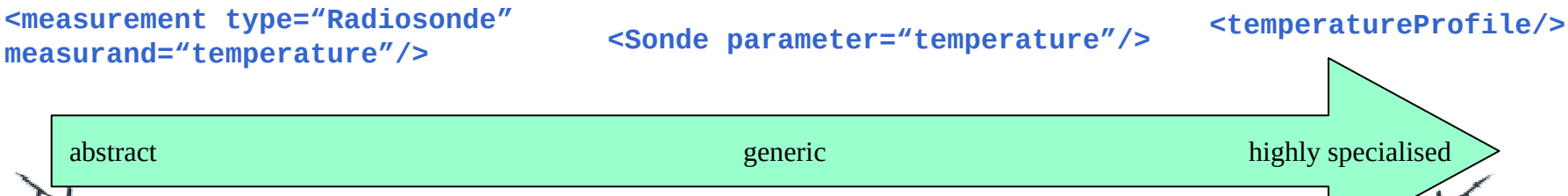
Standards

The importance of governance

- Information community defined by shared semantics
- Need community process to manage those semantics (definitions, models, vocabularies, taxonomies, etc.)
- e.g. CF conventions for netCDF files
- Role of *Feature Type Catalogues* [ISO 19110] and *registers* [ISO 19135]

Governance as driver for granularity

- Remit / interest determines appropriate granularity
- ref. IOC, IHO, WMO



feature types spectrum

Climate Science Modelling Language

Aims:

- provide semantic integration mechanism for NDG data
- explore new standards-based interoperability framework
- emphasise *content*, not *container*

Design principles:

- offload semantics onto *parameter type* ('phenomenon', observable, measurand)
  - *e.g. wind-profiler, balloon temperature sounding*
- offload semantics onto CRS
  - *e.g. scanning radar, sounding radar*
- 'sensible plotting' as discriminant
  - *'in-principle' unsupervised portrayal*
- explicitly aim for small number of *weakly-typed* features (in accordance with governance principle and NDG remit)

Climate Science Modelling Language

CSML feature types

- defined on basis of geometric and topologic structure

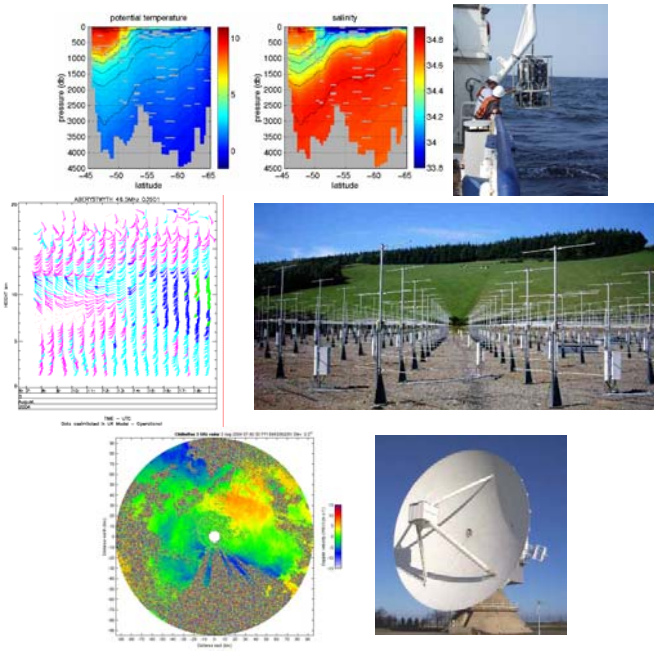
CSML feature type	***Description***	***Examples***
TrajectoryFeature	Discrete path in time and space of a platform or instrument.	ship's cruise track, aircraft's flight path
PointFeature	Single point measurement.	raingauge measurement
ProfileFeature	Single 'profile' of some parameter along a directed line in space.	wind sounding, XBT, CTD, radiosonde
GridFeature	Single time-snapshot of a gridded field.	gridded analysis field
PointSeriesFeature	Series of single datum measurements.	tidegauge, rainfall timeseries
ProfileSeriesFeature	Series of profile-type measurements.	vertical or scanning radar, shipborne ADCP, thermistor chain timeseries
GridSeriesFeature	Timeseries of gridded parameter fields.	numerical weather prediction model, ocean general circulation model

Climate Science Modelling Language

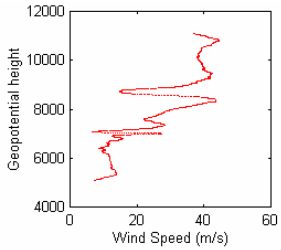
CSML feature types

- examples...

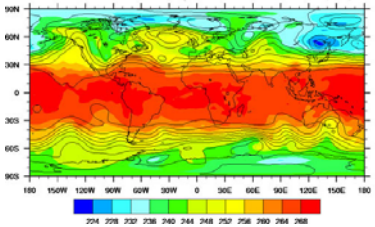
ProfileSeriesFeature



ProfileFeature



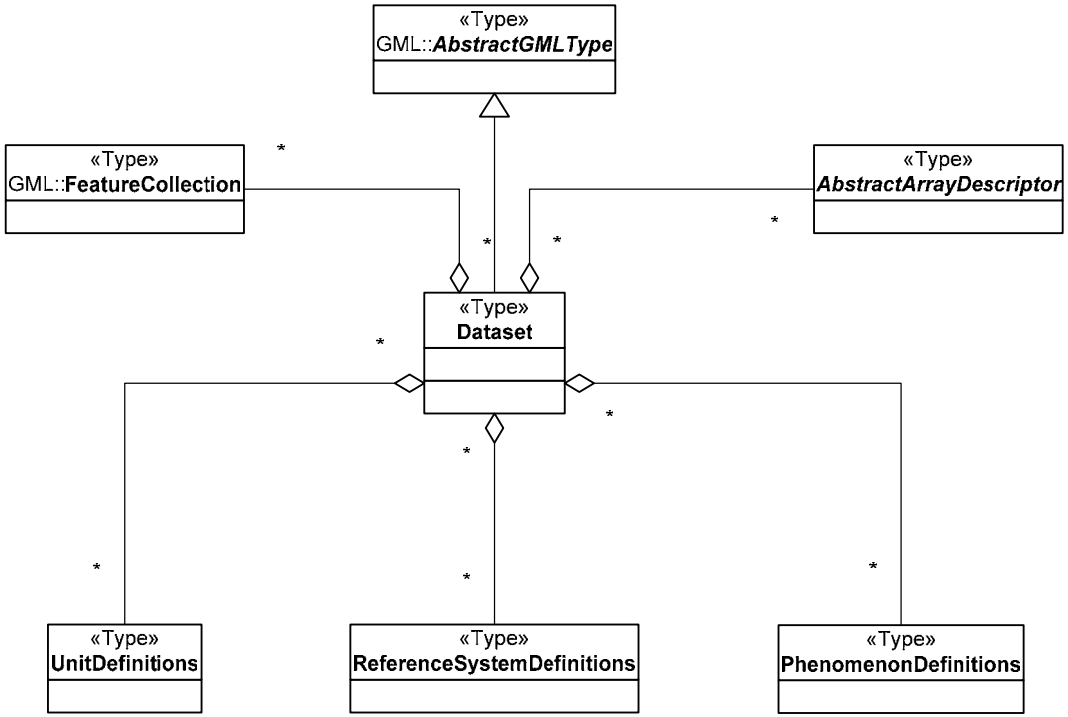
GridFeature



Climate Science Modelling Language

Application schema

- logical structure and semantic content of NDG 'Dataset'
- Based on GML 3.1



Climate Science Modelling Language

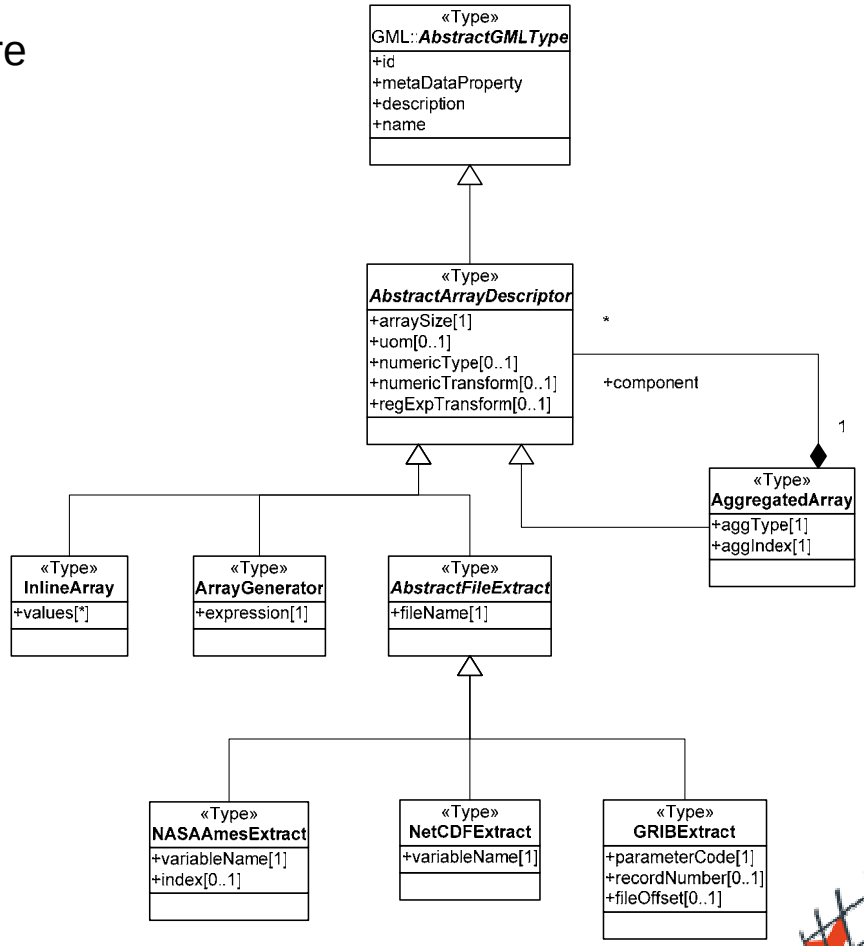
Numerical array descriptors

- provides ‘wrapper’ architecture for legacy data files
- Connected’ to data model numerical content through ‘xlink:href’

Subtypes:

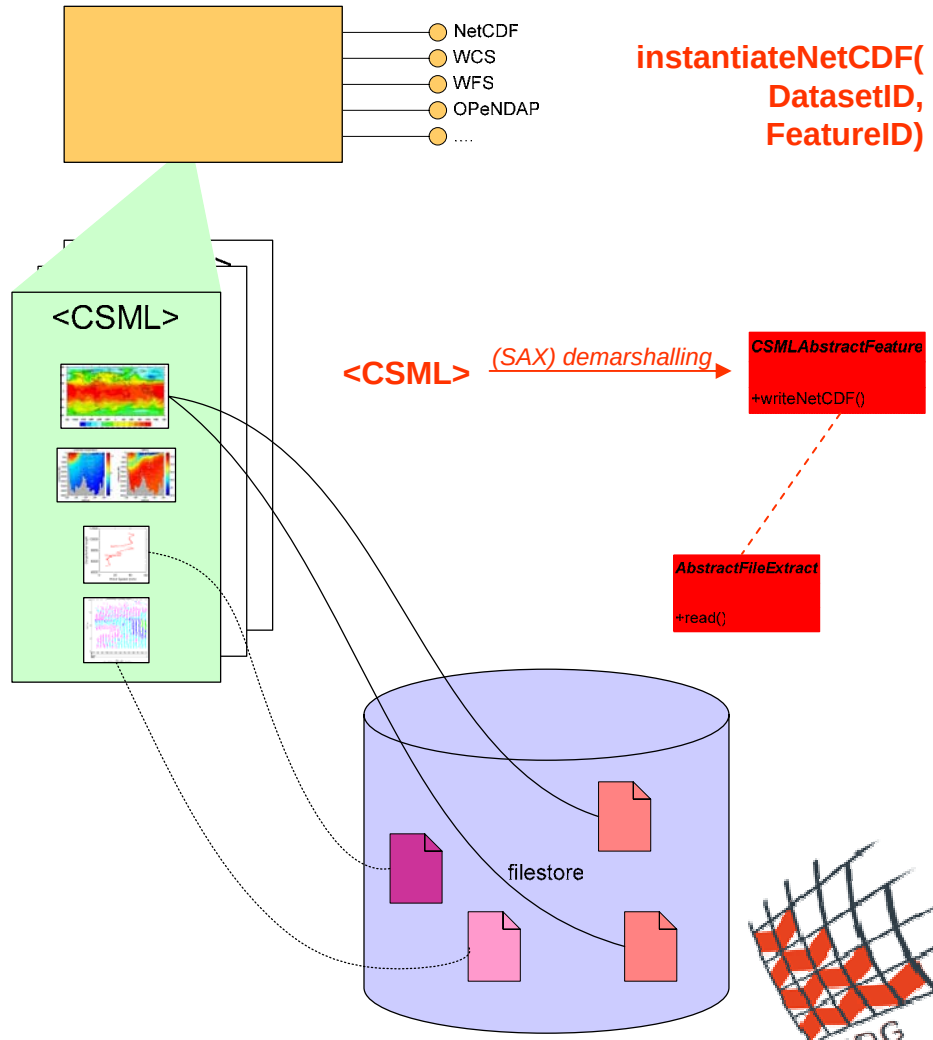
- InlineArray
- ArrayGenerator
- FileExtract (NASAAmes, NetCDF, GRIB)

Composite design pattern for aggregation



Climate Science Modelling Language

Provides semantic abstraction layer



Perspectives

Status:

- Initial feature types defined
- First draft application schema complete
- Trial software tooling being coded (parser, netCDF instantiation)
- Initial deployment trial across BODC, BADC datasets

Future:

- Separate out wrapper implementation (array descriptors)
- Disallow ‘internal’ dictionaries
- More strongly-typed features?
- Follow (and pursue!) GML evolution, enhance compliance
- Expand tooling

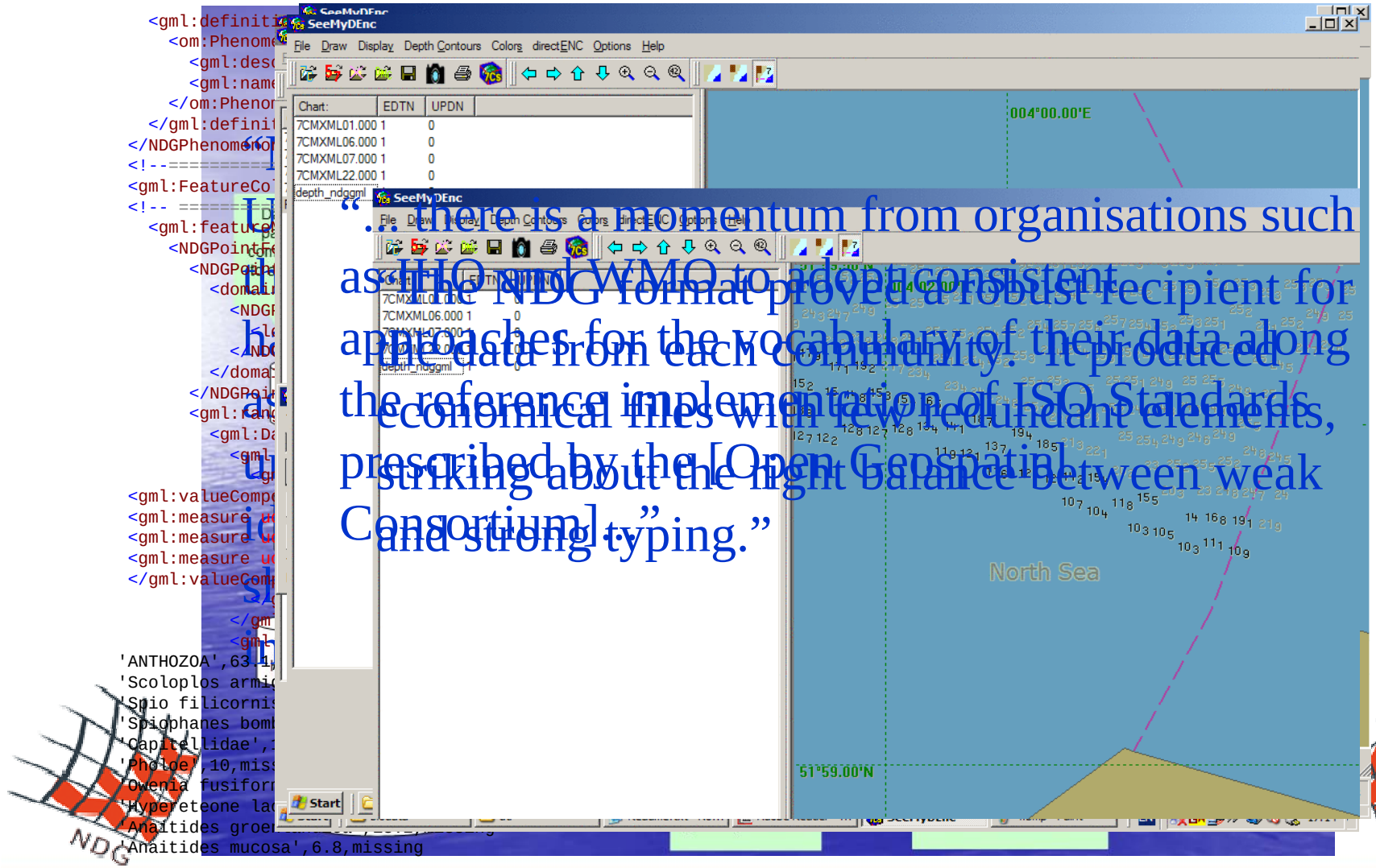
Related work

- WMO, IOC, IHO
- MarineXML
- MOTIIVE (INSPIRE)

Perspectives

EU project – MarineXML

http://www.marinexml.net

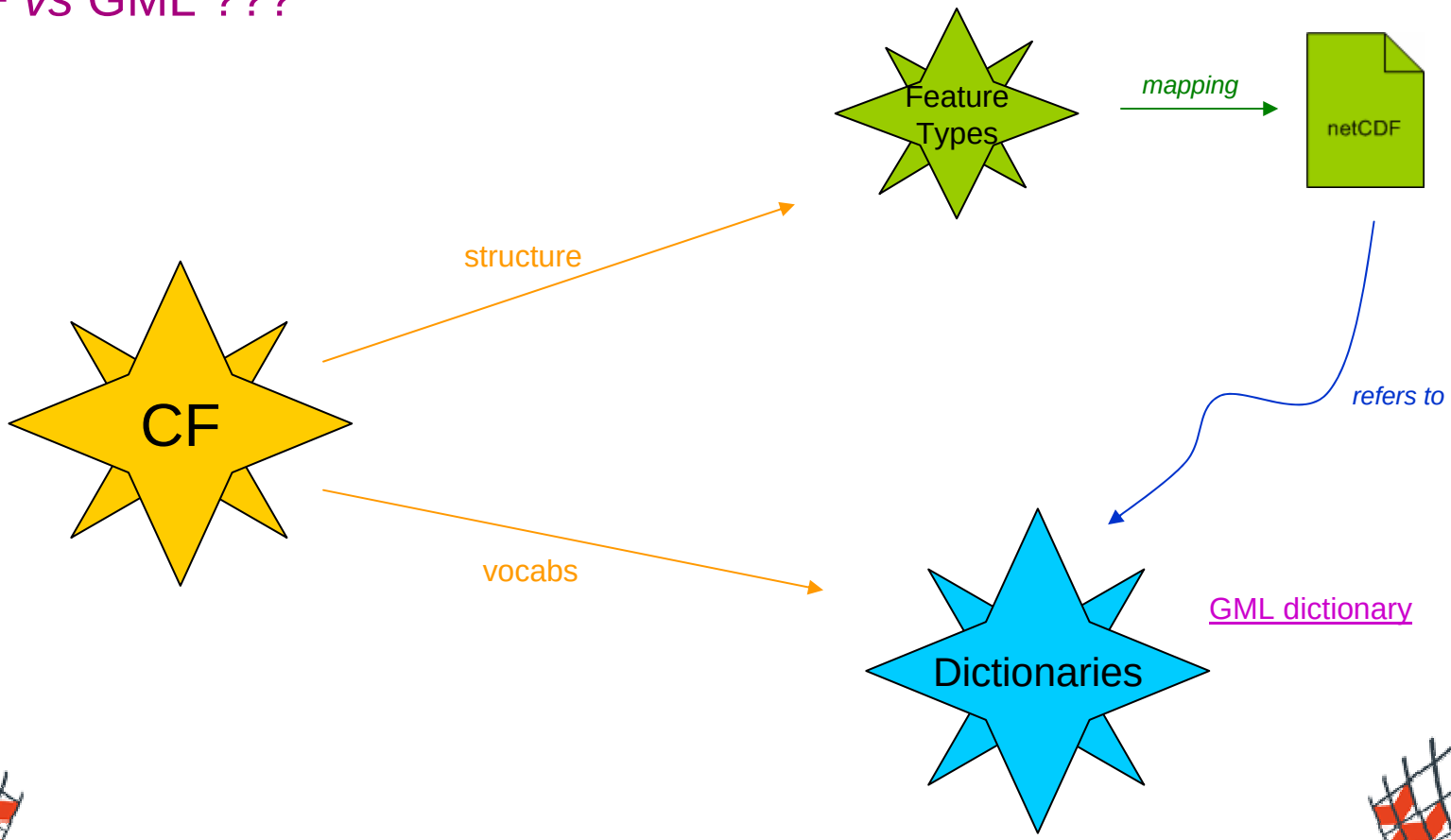


"… there is a momentum from organisations such as IHO and WMO to adopt consistent approaches for the vocabulary of their data along the reference implementation of ISO Standards prescribed by the [Open Geospatial Consortium] …"

"The NDG format proved a robust recipient for the data from each community. It produced economical files with few redundant elements, striking about the right balance between weak and strong typing."

Food for GO-ESSP thought

CF *vs* GML ???



Food for GO-ESSP thought

Dictionaries we need:

- units (udunits, POSC)
- 'phenomena' (CF, BODC)
- CRS (EPSG)

Governance roadmap:

- ISO 19110 "Feature cataloguing methodology"
- ISO 19126 "Profile FACC Data Dictionary"
- ISO 19135 "Procedures for registration of geographical information items"
- IOC 19xxx registries:

IOC (Recommendation IODE-XVIII.7, May 2005)

Recommends the establishment of a MarineXML Steering Group with the following terms of reference:

(i) establish a Pilot Project to set up an ISO 19100 series of standards compliant standards register, with possible collaboration with IHO, to be hosted by the IODE Project Office;

(ii) monitor and assist with XML development activities in other IODE/JCOMM groups, such as ETDMP, GEBICH and SGMEDI.

Food for GO-ESSP thought

GALEON

Use case	GALEON	NDG
1	THREDDS catalog → WCS 'Capabilities XML'	MOLES → WCS 'Capabilities XML'
2	ncML-G → WCS describeCoverage()	CSML → WCS describeCoverage()
3	{netCDF,OPeNDAP} → getCoverage() → GeoTIFF	✓
4	*{netCDF,OPeNDAP} → getCoverage() → ncML-GML*	*{netCDF,OPeNDAP} → getCoverage() → CSML*
5	{netCDF,OPeNDAP} → getCoverage() → ncML-GML + netCDF	{netCDF,OPeNDAP} → getCoverage() → CSML + netCDF
6	clients	✓ (e.g. DODS library for WCS)
7	rasdaman/PostgreSQL → getCoverage()	

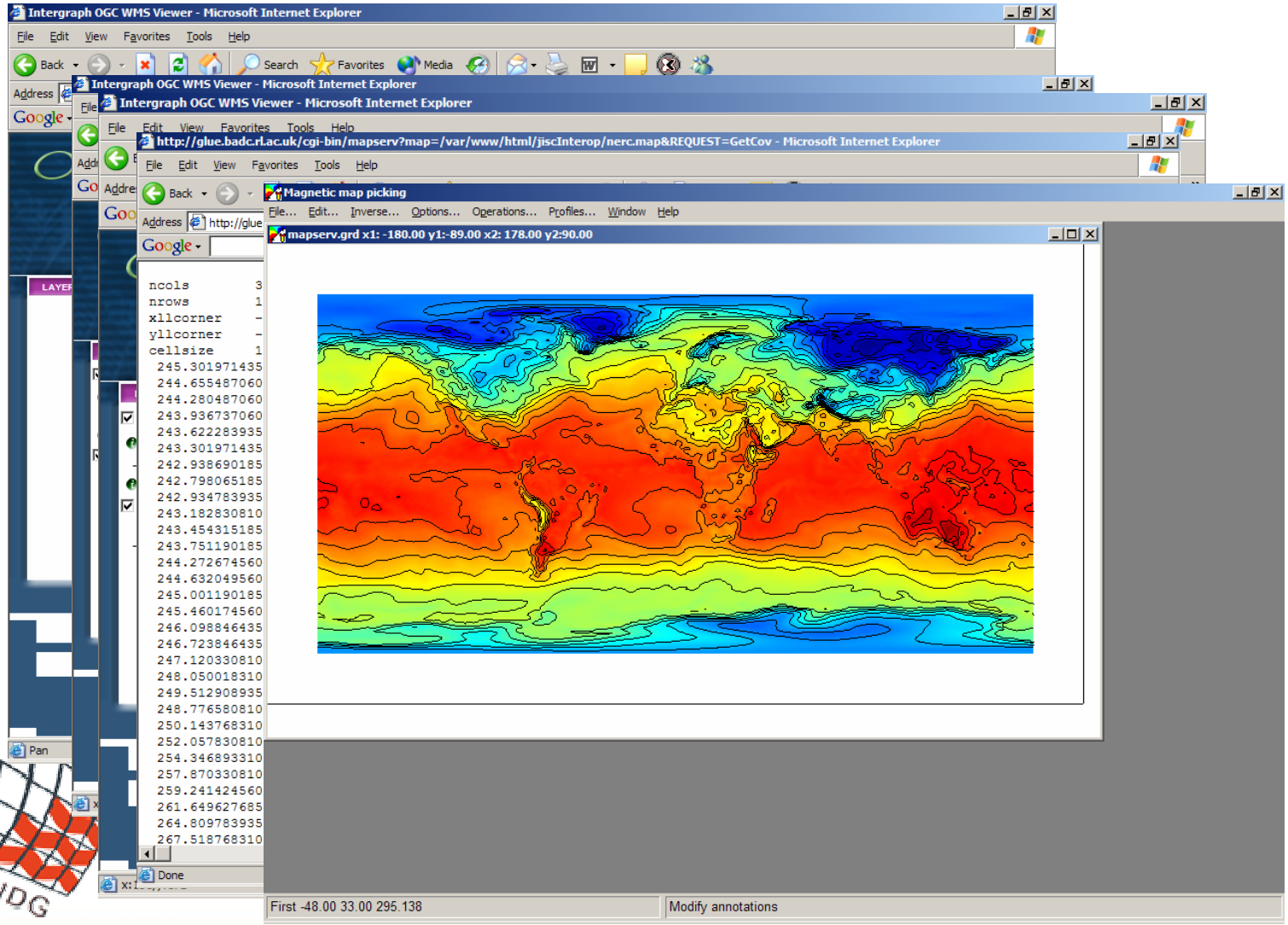
Thoughts:

- WCS requires support for ***at least*** {GeoTIFF | **HDF-EOS** | DTED | NITF | **GML**}
- ...however, CF-netCDF should ideally be one of these
- WCS spec shouldn't be modified for anyone's pet format

NDG & OGC

e.g.: netCDF data through W{M,C}S

- http://glue.badc.rl.ac.uk/cgi-bin/mapserv?map=/var/www/html/jiscInterop/nerc.map



Food for GO-ESSP thought

```
> foreach i (1 2 3 4)
    grep 'iso|ogc|gis|gml' GO-ESSP/* | wc
```

This year so far:

- Lawrence
- Middleton
- Hankin
- Tandy
- O'Brien & Hankin
- (~ 40%)

AUKEGGS

AUKEGGS
Australia-UK collaboration on
Exploitation of Grid and Geospatial Standards

Collaboration between NERC DataGrid (UK) and SEEGrid community (Australia)

- https://www.seegrid.csiro.au/twiki/bin/view/AUKEGGS/WebHome

Aims:

- NDG deployment at TPAC 'Digital Library for Oceans and Climate'
- Joint OGC demonstrator with NOO Oceans Portal
- 'Grid-enabling' OGC web services by profiling against WSRF
- Legacy data integration patterns (e.g. wrappers for relational/file)
- Sept workshop, Edinburgh: "Grid Middleware and Geospatial Standards for Earth System Science Data"