Discussion on Big Data & Processing in MetOcean

5th Workshop on the use of GIS/OGC standards in meteorology
2014-oct-28, DWD, Offenbach, Germany

Peter Baumann & all the good people discussing with him

Jacobs University | rasdaman GmbH
baumann@rasdaman.com

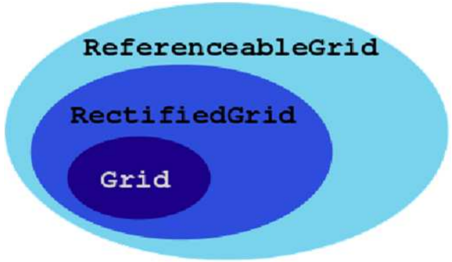
Interesting Facets

- multi-dimensional data (12)
- Performance for realtime access through Web services (11)
  - „Calculations are free, moving data costs“ [HPC]
- Metadata / data integration (11)
- Distributed storage & processing (10)
- Service discovery (9)
- c/s interfaces with enough power, but retaining flexibility & scalability (8)
- Access to heterogeneous (legacy) data (6)
- Quality (incl versioning) (5)
- Processing (paradigms, etc.) (4)
- Subscription vs ad-hoc requests (3)
- Using non-meteorological data for meteorological purposes -> cross-domain integration & fusion (3)
- Security (2)
- Data upload (2)
- Archiving, long-term data preservation (1)
- Persistent IDs (1)
- Predictive analytics & modelling(1)

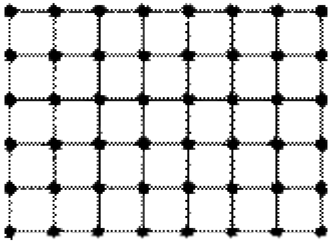
Multi-dimensional data

- Categories of dimensions (aka CRS with datum, offset, UoM, ...)
  - Lat/long
  - Time: SI unit of seconds; calendars
  - Elevation / height / depth (hybrid levels):
    - *Height: nonlinear, height can be obtained only from involving further data; location dependent*
  - Reference time (model run#)
  - N.n. -> Ensemble member#
- NB: OGC coverage definition mandates: 1 CRS per coverage
- Different grid types: regular vs irregular
- Impact of n-D data on implementations
  - Ex: lat/long correlated -> image pyramids simple; not so with more axes

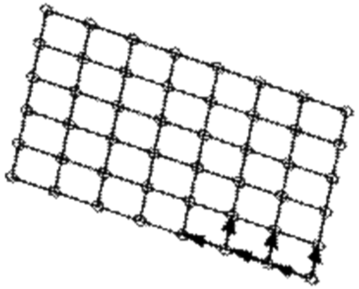
Gridded Coverage Types



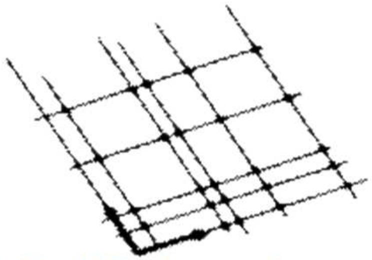
- Not georeferenced, „just pixels“
  - GMLCOV::GridCoverage



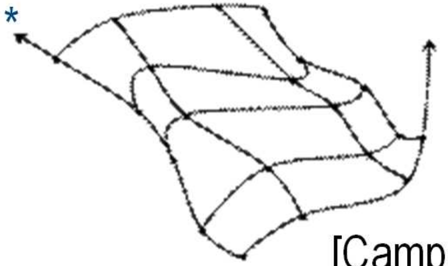
- Georeferenced, regular
  - GMLCOV::RectifiedGridCoverage



- Georeferenced, 1+ irregular axes
  - All axes irregular: GML 3.3 ReferenceableGridByVectors *
  - GMLCOV::ReferenceableGridCoverage



- Georeferenced, 1+ axes warped
  - All axes warped: GML 3.3 *ReferenceableGridByArray* *
  - GMLCOV::ReferenceableGridCoverage



Mix, eg, with sat image timeseries

[Campalani 2013]

*) CR to GML planned

CRS Name Types [OGC 11-135]

- WGS84, RESTful:
  - http://www.opengis.net/def/crs/EPSG/0/4326
- WGS84, KVP:
  - http://www.opengis.net/def/crs?authority=EPSG&version=0&code=4326
- Parametrized („AUTO“) CRSs:
  - *http://www.opengis.net/def/crs?authority=OGC&version=1.3 & code=AUTO42003 & UoM=m & CenterLongitude=-100 & CenterLatitude=45*
- Ad-hoc combination of CRSs:
  - *http://www.opengis.net/def/crs-compound? 1=http://www.opengis.net/def/crs/EPSG/0/4326 & 2=http://www.opengis.net/def/crs/ISO/2004/8601*
- Proprietary CRS definition:
  - *http://www.acme.com/def/this-is-EPSG-4326*
- Inline CRS definition:
  - *srsName="#crsdef"*

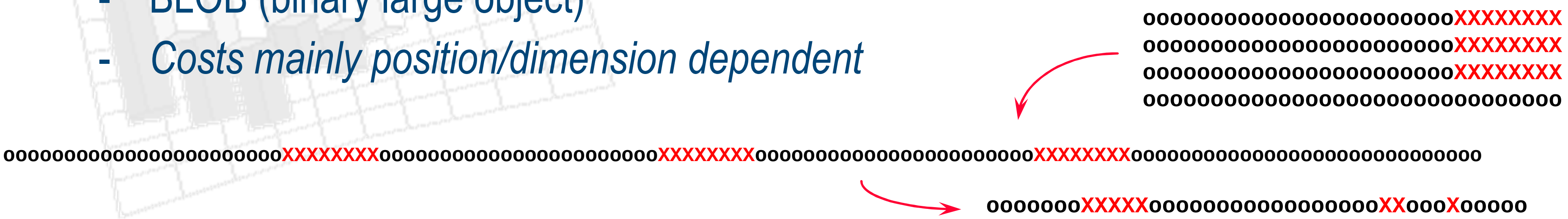
OGC resolver implementation
provided by Jacobs U:
www.earthlook.org/demos/secore

Performance for realtime access through Web services

- „Calculations are free, moving data costs“ [HPC]
- Axis order during sequentialization determines access performance
    - Traditionally, meteo archives store multiple copies for different access patterns
- WMS: tile caching, works well for 2D
- Compression: sometimes can perform evaluation w/o decompressing
- Distributed storage: impact of distribution vs access pattern
- Access behavior patterns?
    - We know a priori how data are structured
    - Caching policies; (in)validation issues
- Pre-materialized products / derivations

Storage Mapping: Variants

- Coordinate-free sequence
  - BLOB (binary large object)
  - *Costs mainly position/dimension dependent*

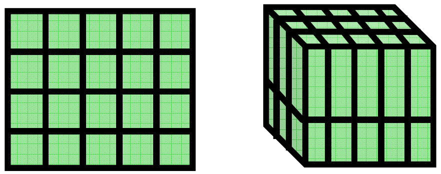


ooooooooooooooooooooooooXXXXXXXX
ooooooooooooooooooooooooXXXXXXXX
ooooooooooooooooooooooooXXXXXXXX
ooooooooooooooooooooooooooooooooo

ooooooooooooooooooooooooXXXXXXXXooooooooooooooooooooooooXXXXXXXXooooooooooooooooooooooooXXXXXXXXoooooooooooooooooooooooooooooo

oooooooXXXXXoooooooooooooooooXXoooXooooo

- Sequence independent, coordinates explicit
  - ROLAP
  - *Costs not position correlated, but high*
- Imaging, multidimensional OLAP
  - Partitioning, sequence within partition
  - *Costs low for bulk access, usually not location correlated*



Metadata / data integration

- Metadata may be derived from something else than „their“ data
  - How to maintain connectivity?
- Metadata may get changed by the process of retrieval
  - Big Data is peculiar in that you typically subset it
  - May lead to such incoherence