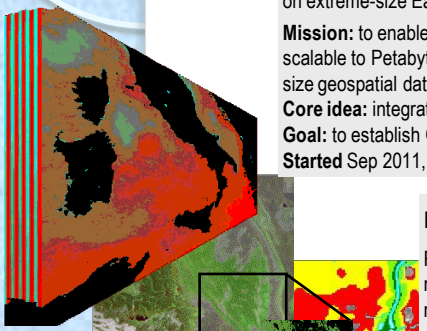
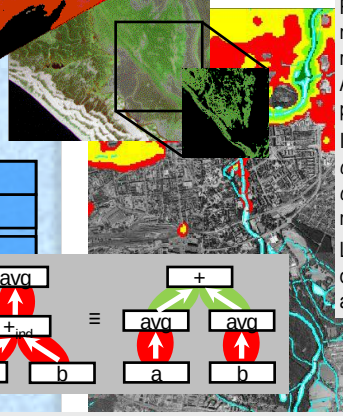
Agile Analytics with EarthServer

Large Scientific Information Systems
Research Group
Head: Peter Baumann
JACOBS UNIVERSITY

Project Summary: EarthServer is establishing open access and ad-hoc analytics on extreme-size Earth Science data, based on and extending the rasdaman Array DBMS.

Mission: to enable standards-based ad-hoc analytics on the Web for Earth science data: scalable to Petabyte/Exabyte volumes, with direct manipulation, analysis & remix of any-size geospatial data.
Core idea: integrated query language for all spatio-temporal coverage data
Goal: to establish OGC standards based client & server technology
Started Sep 2011, runtime 3 years, 5.38m EUR budget (7m US$)





rasdaman

rasdaman raster data manager

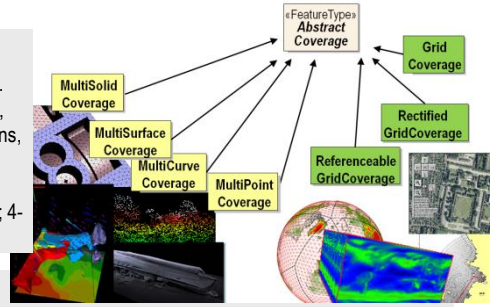
Rasdaman („raster data manager") is a **multi-dimensional Array DBMS** [Baumann, VLDBJ 1994] adding n-D arrays as a new attribute type to relations. The rasdaman query engine extends SQL with n-D array operators, based on Array Algebra which also guides server-side optimizations. Arrays of unlimited size are partitioned („tiled") and stored in a standard relational DBMS.

Incoming queries routinely undergo optimization, like: *tile streaming; adaptive compression; query rewriting (heuristics); physical operator clustering; just-in-time compilation; utilizing GPUs; dynamic preaggregation; tile caching; etc.* Benchmarks have shown that this achieves a speed-up typically by orders of magnitude.

Large-scale **sensor, image, model, & statistics** data can benefit from all classical database advantages, such as **information integration, flexible data retrieval**, and **scalability**.

```
select jpeg( img.green[x0:x1,y0:y1] > 130 )
from   LandsatArchive as img
where avg_cells( img.red+img.green+img.blue) > 120
```

Coverages: A coverage is a „**space-time varying phenomenon**". Cover-age data are gathered in huge, rapidly growing amounts and variations, such as: 1-D sensor time series; 2-D maps; 3-D *x/y/t* satellite image timeseries and *x/y/z* exploration data; 4-D *x/y/z/t* climate & ocean data.



WCS & WCPS

Coverages can be served via OGC **WCS** (Web Coverage Service). The WCS suite of standards establishes a modular set of extensions around a Core facilitating n-D spatio-temporal subsetting (trim & slice). One of these extensions is the OGC **WCPS** (Web Coverage Processing Service) Language, an n-D Array QL enriched with geo semantics and flavored along XQuery.

The following example [Baumann, Geoinformatica 2009] is "*From MODIS scenes M1, M2, and M3, the absolute of the difference between red & nir, in HDF-EOS - but only those where nir exceeds 127 somewhere inside region R*":

Rasdaman is WCS and WCPS reference implementation.

```
for $c in ( M1, M2, M3 ),
    $r in ( R )
where some( $c.nir > 127 and $r )
return encode( abs( $c.red - $c.nir ), "image/tiff" )
```

OGC

The Open GeoSpatial Consortium (OGC), in collaboration with ISO, developes standards for geo-spatial and location based services. Some coverage standards:

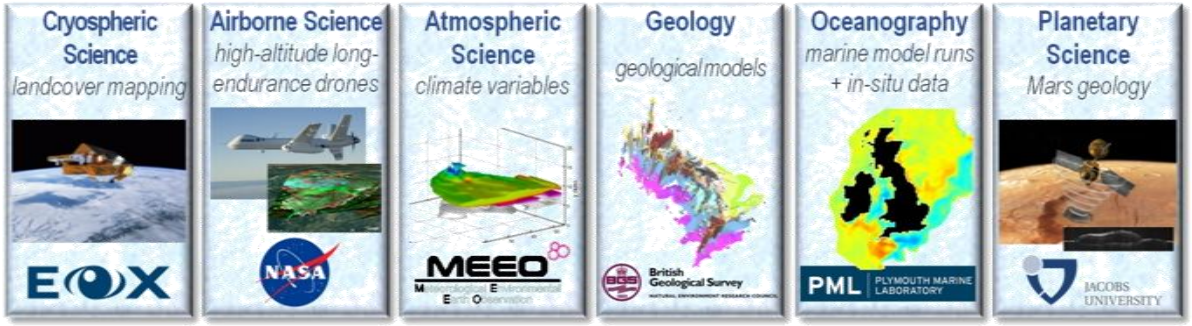
P. Baumann (ed.).: GML 3.2.1 Application Schema – Coverages, OGC 09-146r2

P. Baumann (ed.): WCS 2.0 Core, OGC 09-110r3

P. Baumann (ed.): WCPS Language, OGC 08-068r2

RTD Goals: integrated data/metadata queries +++ support all coverage types +++ databases & in-situ integration +++ Federated query processing +++ Cloud parallelization +++ GIS integration +++ nD visual clients – from mobile to immersive VR +++ Advance OGC standards

Lightouse Applications: Six services, together covering all Earth sciences, 100+ TB of in-situ & imported data



Further partners: rasdaman raster data manager; JACOBS UNIVERSITY; EOX; Consiglio Nazionale delle Ricerche; Fraunhofer