Supported by EU FP7 eInfrastructure, contract 283610 EarthServer

Data, Metadata – Who Cares?

GSA 2013 -- T90. Digital Geosciences:
A Framework for Data-Intensive, Multi-Disciplinary Research and Education
Denver, USA, 2013-oct-30

Peter Baumann[1], Marina Sahakyan,
Giorgios Kakaletris, Panagiota Koltsidas, Thanasis Perperis

Jacobs University, rasdaman GmbH, Athena Research, CNR

[1]p.baumann@jacobs-university.de

[comics © Goscinny & Uderzo]

Data vs Metadata

Data:

- Large in volume – „Big Data“
- semantic-poor
- difficult to interpret,
  at best statistics
- only for clumsy download
- Search only through previously extracted metadata

Metadata:

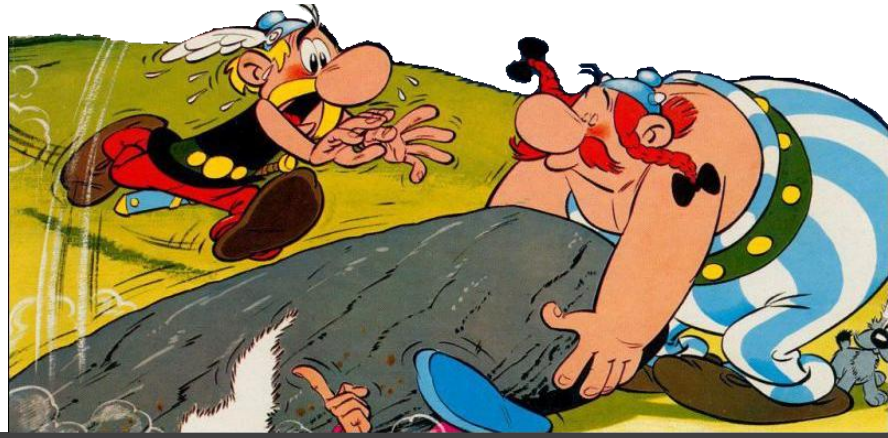
- Small in volume
- rich in content & semantics
- Manifold techniques for interpretation, such as ontologies
- flexibly queryable
- rich body of search technologies
  - SQL, OpenSearch, …

Metadata Often XML → XPath

- "Identifiers of all coverages offered"
- "All formats supported by this server"

```
/CoverageOfferings/ServiceIdentification/ServiceMetadata/
   formatSupported/text()
```

- "spatial extent of coverage X"
- "...and its coordinate system"
- "...and its pixel values"



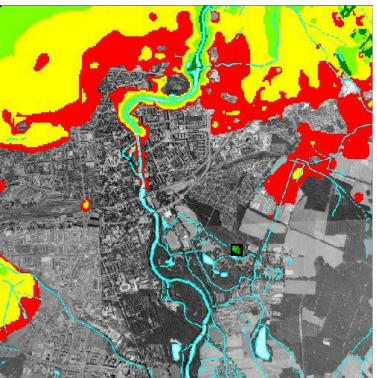
OGC Web Coverage Processing Service (WCPS)

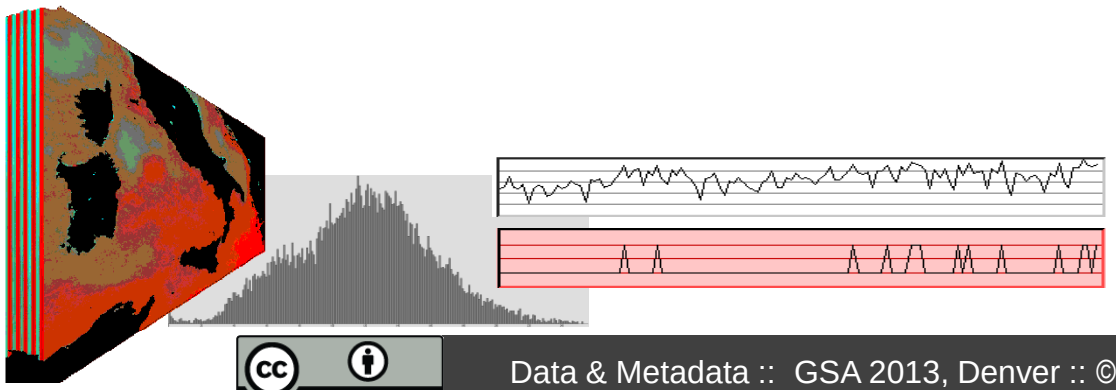
- WCPS = query language
  for ad-hoc navigation, extraction, aggregation, analytics
  on spatio-temporal gridded coverages
  - adopted 2010 [OGC 08-068r2]

- Ex:

```
for $c in ( M1, M2, M3 )
return
    encode( $c.red - $c.nir,
            "image/hdf" )
```

→ (hdf_1, hdf_2, hdf_3)





...and now: Integration

- Merging WCPS with XQuery FLWOR → WCPS 2.0

```
for $c in doc("http://acme.com")//coverage
where
    some( $c.nir > 127 ) and metadata/@region = "Austria"
return
    encode( $c.red - $c.nir, "image/tiff" )
```

```
for $c in doc("WCPS")//coverage/[ some( $c.nir > $c.red )]
return
    <id> { $c/@id } </id>
    <area> { $c/boundedBy } </area>
```

- Implementation: federation of eXist + rasdaman
  - Jacobs University & Athena Research Lab

Array SQL

- „table LandsatScenes, with attributes id, acquisition time, and image“

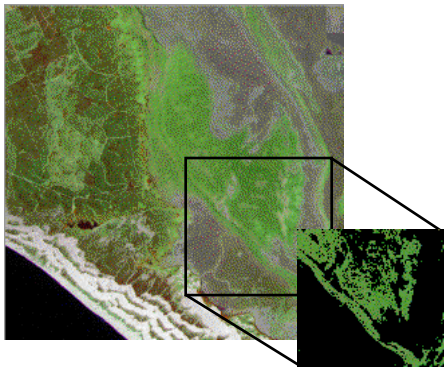
```
create table LandsatScenes(
    id: integer not null,
    acquired: date,
    scene: array( red: integer, ..., blue: integer ) [ 0:4999,0:4999]
)
```

- „difference between red and near-infrared band, in CSV, of Landsat scenes“

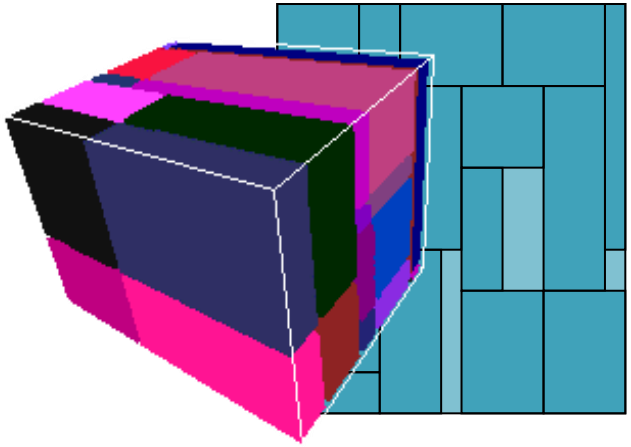
```
select encode( scene.red – scene.nir, „image/tiff“ )
from LandsatScenes
where acquired < „1990-12-24“
```

on the table of ISO SQL WG

rasdaman: Scalable Array Analytics

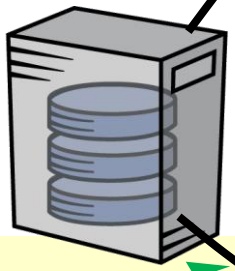


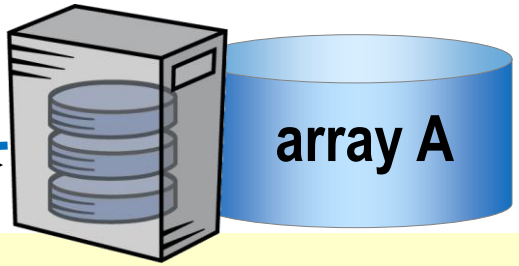
- „raster data manager“: SQL + n-D arrays
- Scalable parallel “tile streaming” architecture
- Storage: [database | preexisting file archive]
- OGC WMS, WCS, WCPS, WPS
  geo service standards
  - WCPS, WCS Core reference implementation

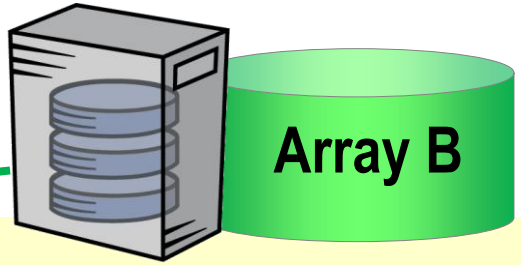


www.rasdaman.org

Semantics-Based Scalability







select
max((A.nir - A.red) / (A.nir +A.red))
from A

select
max((A.nir - A.red) / (A.nir + A.red))
- max((B.nir - B.red) / (B.nir + B.red))
from A, B

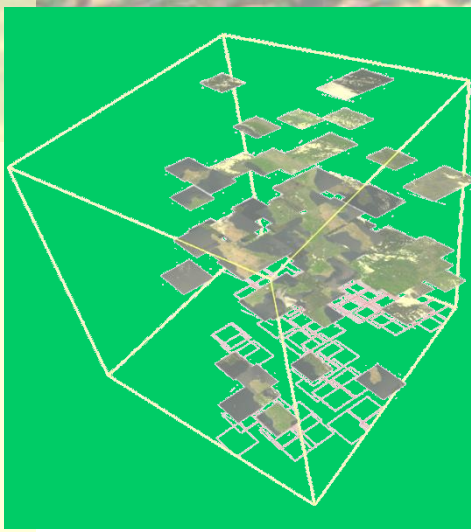
select
max((B.nir - B.red) / (B.nir + B.red))
from B

[Owonibi 2012]

3D Database Visualization

[data courtesy BGS, ESA]

```
select
  encode(
    struct {
      red:   (char) s.b7[x0:x1,x0:x1],
      green: (char) s.b5[x0:x1,x0:x1],
      blue:  (char) s.b0[x0:x1,x0:x1],
      alpha: (char) scale( d, 20 )
    },
    "png"
  )
from SatImage as s, DEM as d
```

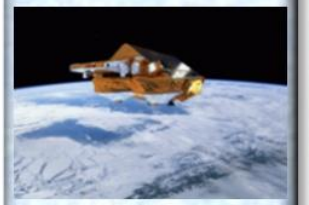


[JacobsU, Fraunhofer 2012]

EarthServer: *Big Earth Data Analytics*

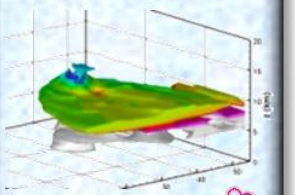
- Scalable On-Demand Processing
  - EU funded, 3 years, 7m US$
  - Platform: rasdaman, Array Analytics server
  - Distributed query processing, integrated data/metadata search, 3D clients
- 100+ TB databases for Earth & Planetary science

Cryospheric Science
landcover mapping



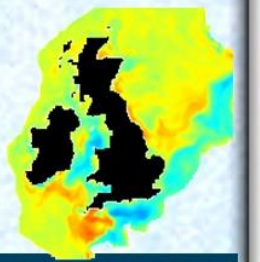
Airborne Science
high-altitude long-endurance drones

Atmospheric Science
climate variables

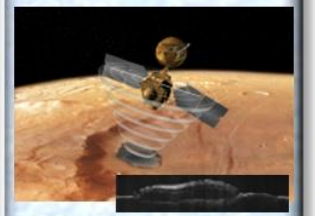


Geology
geological models

Oceanography
marine model runs + in-situ data



Planetary Science
Mars geology



Take Home Messages

- Coverages = sensor, image, simulation, & statistics data
  = a main source of *Big Data*
- high-level query languages overcome data / metadata divide
  - Semantic interoperability on *data*
  - Visual clients can hide QL, servers can parallelize
- rasdaman adds arrays to SQL & XQuery