Preserving Geo-Scientific Data Assets Through Service Interoperability

Peter Baumann[1], Andrei Aiordachioaie[1]

[1] ***Jacobs University Bremen***
Campus Ring 12, 28759 Bremen, Germany
EMail: p.baumann@jacobs-university.de

ABSTRACT

The OGC Web Coverage Processing Service (WCPS) provides a flexible, high-level language for server-side raster data ("coverages"), processing, bridging Web Coverage Service (WCS) and Web Processing Service (WPS). We present core language concepts and their efficient implementation and argue WCPS's usefulness for flexible ground segment and on-board services.

Keywords: interoperability, geo standards, OGC, WCPS

INTRODUCTION

Sensor data heavily contribute to today's geo data mix. Often remote and in-situ measurements can be represented as multi-dimensional raster data, such as 1-D timeseries, 2-D imagery, 3-D image time series or geophysical data, 4-D climate/ocean data, and n-D statistics data with "abstract", non-spatio-temporal axes. While today's efforts still emphasize mere data navigation and download on raster data, the next generation of services foreseeably will include on-demand analysis capabilities.

In Fall 2008, the Web Coverage Processing Standard (WCPS) Language Interface Standard [3] has been issued by the Open GeoSpatial Consortium (OGC, www.opengeospatial.org). WCPS is part of the Web Coverage Service (WCS) suite which is OGC's raster service standard. WCS is organized into a service core which provides simple access functionality (subsetting, scaling, reprojection, and format encoding) and an open-ended list of extensions which add bespoke functionality. In WCS 1.1, two extensions are available: the processing service, WCPS, and a so-called "transactional" service, WCS-T, for updating a server's coverage offerings. Figure 1 puts WCPS in perspective relative to some fundamental OGC standards, WMS, WFS, WCS, and CS-W together with some of their extensions.

In this contribution we focus on WCPS. Based on the coverage model of ISO 19123 [8] / OGC Abstract Specification Topic 6 [9] and WCS, the WCPS extension adds a request language for multi-dimensional raster data, suitable for specifying navigation, extraction, and analysis of sensor, image, and statistics data. By nesting expressions, tasks of unlimited complexity can be formulated. WCPS, therefore, has been dubbed "SQL for coverages".

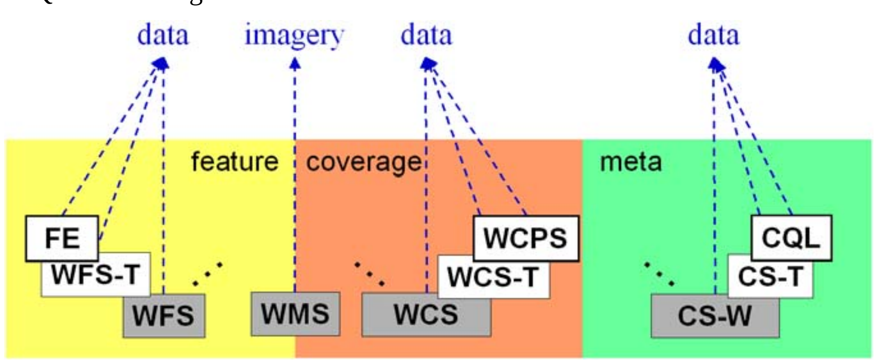

Figure 1: Role of WCPS in a synoptic view of some core OGC standards – WCS, WCS, CS-W, and WMS – together with some of their extensions.

A WEB COVERAGE PROCESSING LANGUAGE

We introduce the WCPS language by means of an example; for details, the specification document [3], a background paper [5], and an online tutorial at [10]. Assume the server offers a set of three coverage objects *ModisTS1*, *ModisTS2*, and *ModisTS3*, each one representing a 3-D MODIS time series data cubes. Further, assume a classification coverage containing a 2-D binary mask with a pixel value of 1 indicating land and 0 indicating water. Task is to pick those coverage objects which contain at least one pixel where, over land, the near-infrared component (*nir*) exceeds 250; of these matches, the NDVI of a particular time slice needs to be delivered as HDF-EOS file. In WCPS Abstract Syntax[1], the corresponding request looks as follows:

```
for $c in ( ModisTS1, ModisTS2, ModisTS3  ),
   $m in ( LandSurfaceMask )
where
   count( $c.nir > 250 and $m ) > 0
return
   encode(
        ( ($c.nir-$c.red) / ($c.nir+$c.red) ) )  [ t( "Sun Mar 22 13:33:29 CET 2009" ) ],
      "HDF-EOS"
   )
```

The result of this request will consist of a set of EOS-HDF files, each one containing a 2-D NDVI image. Depending on where the predicate fires, the result set may contain between zero and three result files.

WCPS only defines the abstract service language and is protocol agnostic. Concrete bindings are available for both WCS (as an optional request type *ProcessCoverages*) and WPS (as an application profile of the *Execute* request type). As far as OGC Sensor Web Enablement (SWE) addresses coverages, WCPS also adds a retrieval, extraction, and analysis service component there as well.

Providing a formal semantics definition, as is the case with WCPS, makes such requests machine understandable. This feature has several beneficial effects: machines can communicate without human intervention; reasoning about requests is possible, which is useful for complexity analysis ("*How long will it take for answering this request? How much shall it cost?*"), for server-side optimization ("*Is there any equivalent rephrasing which can be evaluated faster?*"), and for dynamic request orchestration ("*How can this request be split efficiently over a set of nodes available?*"). Commonly, the label "Semantic Web" is attributed to services exhibiting such qualities.

A WCPS demo site is available at [10]; among others, it allows users to write their own WCPS requests against a sample data set.

IMPLEMENTATION AND OPTIMIZATION

Petascope is the reference implementation of WCPS [8]. In addition, petascope includes a WCS, WCS-T, and soon WPS interface for access to the server's offerings. The implementation stack consists of a Java servlet which performs protocol handling and request analysis. Incoming requests from all interfaces are translated into queries of the rasdaman ("raster data manager") database management system [1][2] which is available under a free licence at [13]. Rasdaman offers SQL-style retrieval and manipulation of multi-dimensional raster data of unlimited size. Query language and server architecture have been designed taking into account the state of the art in databases, high-performance computing, distributed computing, and image processing.

Emphasis in rasdaman is on intelligent algorithms allowing for admin control, but not requiring human intervention. Several effective run-time optimizations have been designed, including:

1 Additionally, an XML encoding of the request language is specified in the standard.

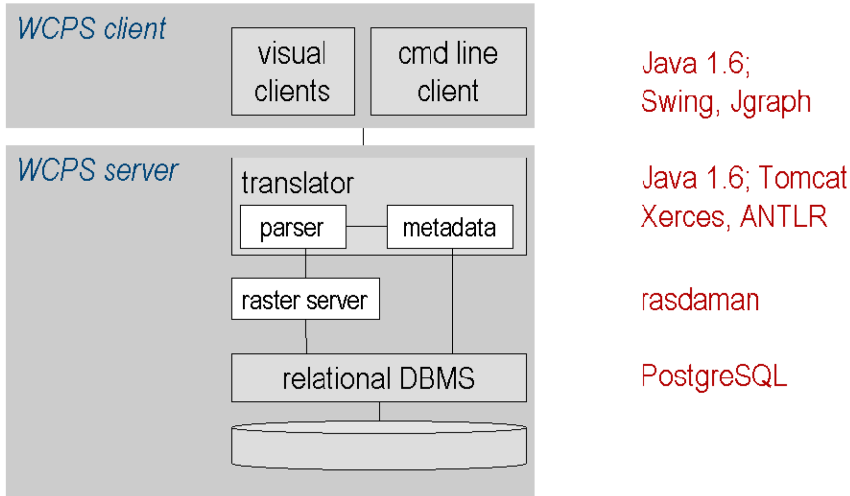

Figure 2: Petascope system architecture

- *algebraic rewriting:* incoming queries are compared against more than 150 rules to find ways of rephrasing the query to a more efficient variant delivering the same result.
- *storage optimization:* the tile storage architecture allows to lay out tile sizes and shapes to achieve the best possible response time for a given application profile. Additionally, tiles can be compressed using one of currently twelve different techniques. Compression is available for storing tiles and, independently, for transmission of query results.
- *dynamic pre-aggregation:* By adopting OLAP techniques, image pyramids are extended to the n-dimensional case and non-correlated axes. This allows efficient support, e.g., of *x/y/t* image time series and *x/y/z/t* climate and ocean data.
- *parallel and distributed execution:* parallel server instances allow for concurrent, independent access to a data set; additionally, complex and/or data intensive requests can be distributed in a network; automatic service orchestration is under development. This not only improves performance, but adds significantly to the system's failover capabilities; automatic recovery is built in.
- *just-in-time compilation (JIT)* of incoming queries: maximal sub-queries are collected, compiled, and linked dynamically; code can be generated for either CPU or GPU. Compiled queries can be reused even if parameter values change, hence achieving almost the speed of handcrafted code while keeping the flexibility of interpreted languages.

Let us inspect the last technique in more detail. Figure 3 shows a performance evaluation where a variable number of pixel-wise multiplications has been applied to a 512x512 double-precision floating-point image. The corresponding rasql query is

*select x * … * x*
from doubleImage as x

In Figure 3, the horizontal axis shows the number of multiplications per pixels, i.e., the number of occurrences of “*” in the above query. The “ORIGINAL” graph measures retrieval with all server optimizations switched off. It is clearly visible that the number of operations executed has heavy impact on response time. “COLD” refers to JIT compilation where C code is generated, compiled, linked into the server, and executed whenever the query arrives. In the “HOT” case, the query is already pre-compiled and just needs to be executed. For comparison, a hand-programmed version of this particular query’s code has been implemented, shown as the “TAILORED” variant. Obviously, the HOT evaluation comes close to the TAILORED code’s speed for moderate numbers of multiplications

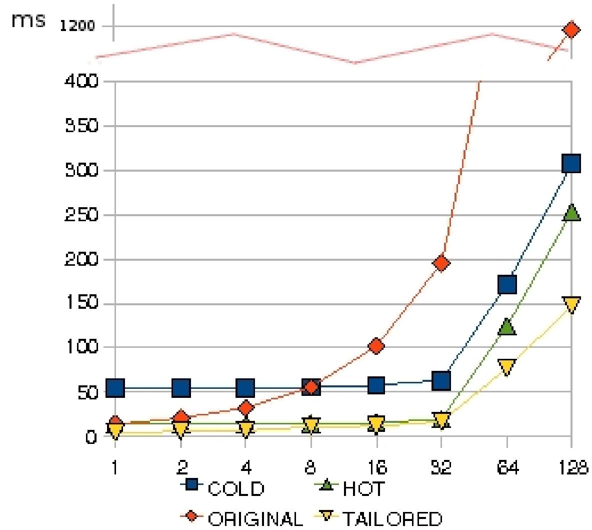

Figure 1: Sample WMS query response time of rasdaman (result image size 512x512).
COLD = cold cache of dynamically compiled library; HOT = hot ynamic library cache;
ORIGINAL = unoptimized query; TAILORED = hand-optimized code.

On off-the-shelf PC hardware, complex requests have shown to be answered by rasdaman far below 100ms. The rasdaman system is in operational use since many years; one installation is at the French National Geographic Institute in Paris where it serves a ~13 TB airborne image through a WMS and rasql queries.

Altogether, rasdaman has been dubbed "the most comprehensive implementation of an array DBMS" in [5].

CONCLUSION

The OGC WCS suite, currently encompassing WCPS and WCS-T, seem to offer suitable interfaces for establishing next-generation Earth observation management systems. We have presented WCPS, OGC’s raster processing language, giving a glance at WCPS core concepts, implementation, and optimization. WCPS closes several gaps currently existing:

- Feature and meta data standards already have their query capabilities, namely OGC Common Query Language (CQL) / Filter Encoding (FE). With WCPS, retrieval capabilities are now available on all geo data categories.
- The Sensor Web Enablement (SWE) deals with coverages over the full width of its processing chain; however, versatile retrieval on the sensor data assets generated is not specified (for good reasons, such as modularity of specifications).
- The Web Processing Service (WPS) provides a powerful service paradigm. However, it lacks interoperability and semantics sharing: the operation semantics is only defined in the human-readable title and abstract, only the function signature is machine-readable and discoverable. Further, WPS functionality is static – any change in functionality requires programming. WCPS as a WPS profile [4] adds interoperable ad-hoc coverage processing to WPS.

For an efficient implementation of WCPS we propose the rasdaman-based reference implementation, petascope. NASA considers this interface suitable not only for customers of ground data centers, but also for on-board deployment (see the video available from [10]); OGC standards, then, become the high-level access interfaces while the service stack's maintenance interfaces allow for convenient in-flight administration.

In WCS 2.0, which is currently being proposed for voting and adoption, the core / extension concept of mdularization will be emphasized even more so that focused so-called Application Profiles can be built which resemble groupings of the core with selected extensions to serve particular application domains, such as SCADA (i.e., in-situ sensors), Earth Observation, and meteorology/ocean research.

ACKNOWLEDGEMENT

This work has been funded by ESA under the VAROS contract.

The author gratefully acknowledges the excellent contributions of the many developers engaged in the softdware development: Andrei Aiordachioaie, Constantin Jucovschi, and Sorin Stancu-Mara, and many more contributors who altogether have made petascope possible.

REFERENCES

[1] P. Baumann.: "On the management of multidimensional discrete data". In: VLDB Journal 4(3)1994, Special Issue on Spatial Database Systems, VLDB, pp. 401 – 444.

[2] P. Baumann: "Large-scale raster services: a case for databases" (invited keynote). In Proc. 3rd Intl Workshop on Conceptual Modeling for Geographic Information Systems (CoMoGIS), 2006, LNCS 4231, Springer, pp. 75 – 84.

[3] P. Baumann (ed.): *Web Coverage Processing Service (WCPS) Language Interface Standard*. OGC 08-068r2, 2008.

[4] P. Baumann (ed.): WPS – Web Coverage Processing Service (WCPS) Application Profile. OGC 09-045, 2009.

[5] Baumann, P.: The OGC Web Coverage Processing Service (WCPS) Standard. Geoinformatica, 2009, DOI 10.1007/s10707-009-0087-2.

[6] R. Machlin: "Index-based multidimensional array queries: safety and equivalence". Proc. 27th ACM Symposium on Principles of Database Systems (PoDS), 2007, Beijing, China, pp. 175-184.

[7] Whiteside, A. (ed.): Web Coverage Service (WCS) — Transaction operation extension. OGC 07-068r4, 2009.

[8] N.n.: *Geographic Information - Coverage Geometry and Functions*. ISO 19123:2005, 2005.

[9] N.n.: *Abstract Specification Topic 6: Schema for coverage geometry and functions*. OGC 07-011, 2007.

[10] www.earthlook.org. Last seen on November 15, 2009.

[11] www.opengeospatial.org. Last seen on November 15, 2009.

[12] www.petascope.org. Last seen on November 15, 2009.

[13] www.rasdaman.org. Last seen on November 15, 2009.

AUTHOR BIOGRAPHY

Peter Baumann researches and teaches databases, Web services, and software engineering at Jacobs University Bremen since 2004. He holds a PhD in CS from Technische Universität Darmstadt (1993). His research interest focuses on large-scale multidimensional raster databases; he addresses foundations, architectures, optimization, and applications in earth and life sciences. He has authored and co-authored 70+ book chapters and reviewed publications in this area.

Peter Baumann has patented and architected rasdaman, the first multi-dimensional raster database management system. For its commercialization he has founded and leads rasdaman GmbH. The rasdaman system, which is installed internationally, has received a series of innovation awards, such as the European IT Prize.

In the Open GeoSpatial Consortium (OGC) Peter Baumann co-chairs the raster-relevant working groups. He has authored the WCPS standard for Web-based geo raster processing; at the time of this writing, he is developing the WCS 2.0 specification set. He is principal architect of the WCPS reference implementation, petascope.