Evaluating Metadata access strategies with the GOME test suite

André Gemünd
Fraunhofer SCAI

www.eu-egee.org

Fraunhofer Institut Algorithmen und Wissenschaftliches Rechnen

Information Society and Media

Motivation

- **Testing the test suite**
  - Sufficiency of specification and utility
- **Investigate AMGA and GRelC as alternatives**
  - Until now we‘ve used OGSA-DAI in NA4
  - Used a java wrapper to access from Python and Perl
  - gLite integration

Fraunhofer SCAI Institut Algorithmen und Wissenschaftliches Rechnen

Introduction

- **DEGREE Project**
  - Dissemination and Exploitation of GRids in Earth sciencE
  - Bridge Earth Science and Grid Community
  - Identify barriers for broader acceptance
  - Identify and assess key requirements
  - Improve communication and collaboration

Fraunhofer SCAI Institut Algorithmen und Wissenschaftliches Rechnen

Introduction

- **Test suites**
  - Specify typical workflows for earth science applications
  - As white papers for testing Grid middleware
  - Organised and grouped into categories (data management, etc.)
  - Consisting of test cases with annotated tested requirements

Fraunhofer SCAI Institut Algorithmen und Wissenschaftliches Rechnen

Introduction

- **GOME-Validation Test Suite**
  - High amount of datasets from two sources
    - GOME satellite measurements
    - LIDAR ground station measurements
  - Correlate by metadata
    - geo-coordinates & date of measurement
  - Target components (as specified):
    - Data management
    - Database access
    - Workflow control

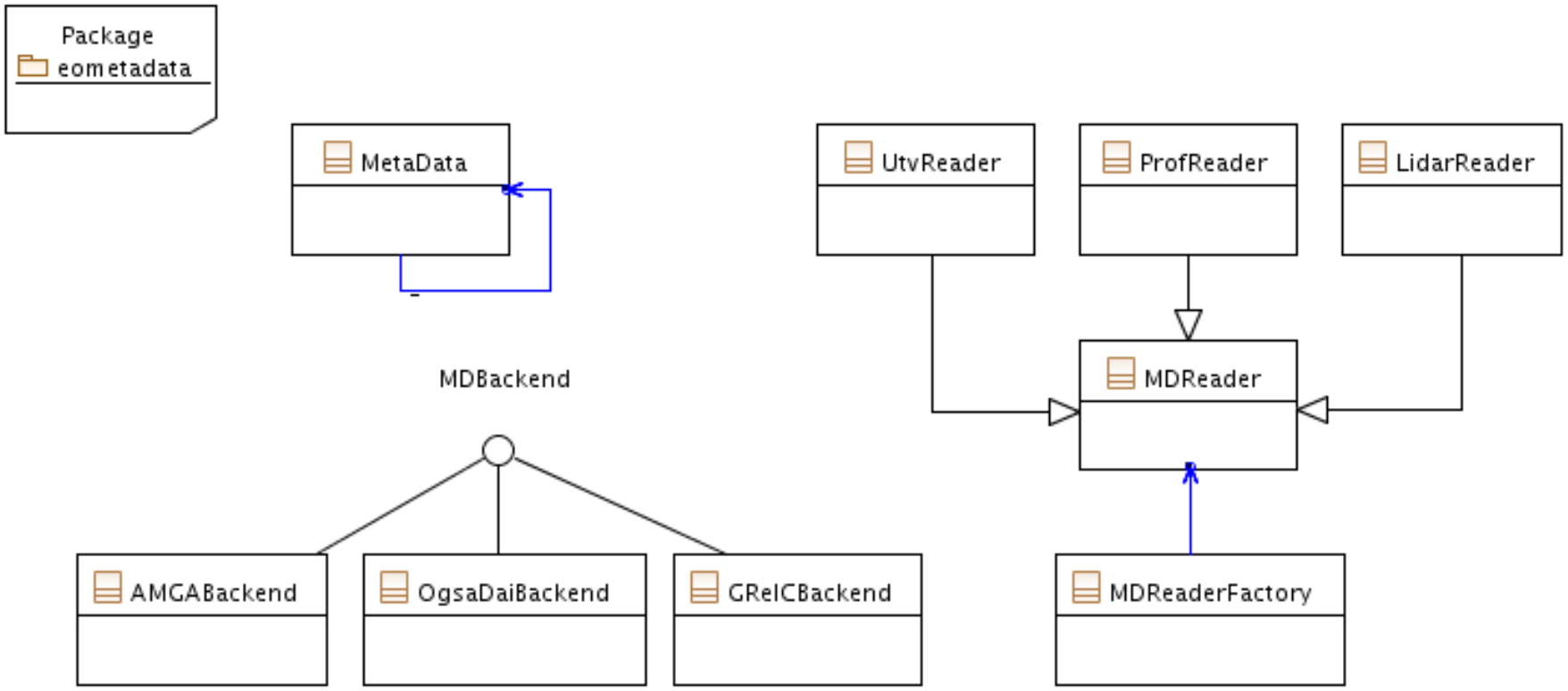
Fraunhofer SCAI Institut Algorithmen und Wissenschaftliches Rechnen

Proceeding

- **What we did**
  - Implement GOME-Validation as a representative workflow
    - Transmission and Grid registration of data files
    - Extraction and archiving of Metadata
    - Bidirectional correlation of files through Metadata
    - Abstraction of Metadata backend

Fraunhofer Institut Algorithmen und Wissenschaftliches Rechnen

- **Software Design**



Fraunhofer SCAI
Institut
Algorithmen und Wissen-
schaftliches Rechnen

Results

- **Problems / Characteristics**
  - Backend Compatibility
  - Data schema and types
  - Query language
  - GIS features
  - Indexing (IDs)
  - Bulk Action support
  - Hierarchical metadata
  - Reuse of Data

Fraunhofer SCAI Institut Algorithmen und Wissenschaftliches Rechnen

Results

- **Database Compatibility**
  - AMGA
    - uses ODBC
      - *MySQL, Oracle, pgSQL, etc.*
    - Extensions and custom Functions need to be added to the Query Parser (Bison Grammar)
  - GRelC
    - C API libraries
      - *Config file states "choose between mysql and pgsql"*
    - Needs pgSQL as configuration backend

Fraunhofer Institut Algorithmen und Wissenschaftliches Rechnen

Results

- **Database Compatibility**
  - OGSA-DAI
    - Unique strength
    - Uses JDBC, eXist and custom drivers
    - Write data providers for arbitrary data sources
      - *Databases and files already included*
    - Combine data from different sources
    - Execute Transformations on data
    - Deliver to Grid-FTP, Gridservice, Client, …

Fraunhofer SCAI Institut Algorithmen und Wissenschaftliches Rechnen

Proceeding

- **Data schema (OGSA-DAI & GRelC)**
  - Raw SQL tables
  - Taken directly from Test suite specification
  - 2 Tables
    - One for LIDAR and one for GOME files
    - Problem: 1 Lidar files hosts n datasets
      - *Different time / coordinates*
      - *Save redundant or introduce relations?*

Fraunhofer SCAI Institut Algorithmen und Wissenschaftliches Rechnen

Proceeding

- **Data schema (AMGA)**
  - We had to devise a modified schema
    - AMGA uses path structures
    - Entity-specific attributes
  - Leverage advantages
    - Dynamic change
    - Inheritance of attributes (hierarchy)

Fraunhofer SCAI Institut Algorithmen und Wissenschaftliches Rechnen

Results

- **Using hierarchies in AMGA example**
  - /gometest/lidar/ano/hgl/30108/
    - /ano/
      - *Identifies station and thus also coordinates*
      - *Here: Andoya, Norway*
    - /hgl/
      - *Author, here: Georg Hansen*
    - /30108/
      - *Identifies file entity*
    - Files in this directory
      - *Real Datasets*

Fraunhofer SCAI Institut Algorithmen und Wissenschaftliches Rechnen

Results

- **Datatypes: Location of measurement**
  - PostGIS Polygon
    - AMGA can use int, float, varchar, timestamp, text, or numeric
      - *But: unknown fieldtypes of database get returned as text*
    - OGSA-DAI & GRelC let you choose
      - *No datatype abstraction*
    - Function to determine containment?
      - *See query language*

Fraunhofer SCAI Institut Algorithmen und Wissenschaftliches Rechnen

Results

- **Datatypes**
  - No additional types offered by the services
  - Desirable
    - Relations
      - *containment, adjacency, …*
      - *Custom relations (ontology-like)*
      - *isResultOf*
      - *isUsedInExperiment*
    - Array types
  - Not only abstraction but extension

Fraunhofer SCAI Institut Algorithmen und Wissenschaftliches Rechnen

Results

- **Query language**
  - OGSA-DAI and GRelC use SQL
    - Highly coupled to table schema
    - Differences in SQL dialect (e.g. pgSQL <-> Oracle)
    - Support for SQL functions, Views, Extensions
      - *Syntax errors if extension is not enabled (e.g. PostGIS)*
  - GRelC add. supports XMLDB query language
    - XPath XQuery
  - AMGA defines own query language
    - Makes for reusable queries / abstractions
    - May possibly limit query power
      - *Add. Functions need source change*

Fraunhofer SCAI Institut Algorithmen und Wissenschaftliches Rechnen

Results

- **Bulk Actions**
  - AMGA additionally supports socket connection instead of document based (SOAP)
    - Low latency
    - Multiple queries without delay
    - High transfer rates possible
  - OGSA-DAI workflows
    - Pipeline, Parallel grouping of activities
    - Powerful but complicated

Fraunhofer SCAI Institut Algorithmen und Wissenschaftliches Rechnen

Results

- **What we would like to have**
  - Integration of external data sources like OGSA-DAI
    - *For custom data sources like swiss-prot etc.*
  - Integration to gLite
    - *Integration with file catalogue*
      - Browsable in both directions
    - *Support for aliases and replicas*
      - Assess best replica for current location
    - *VOMS-based Authorization & Authentication*
  - Extendible for GIS-features and the like
  - APIs for Java, C++, Python & Perl

Fraunhofer SCAI Institut Algorithmen und Wissenschaftliches Rechnen