Uwe Schindler, MARUM – University of Bremen, uschindler@pangaea.de
Benny Bräuer, Alfred Wegener Institute for Polar and Marine Research, benny.braeuer@awi.de
Michael Diepenbroek, MARUM – University of Bremen, mdiepenbroek@pangaea.de

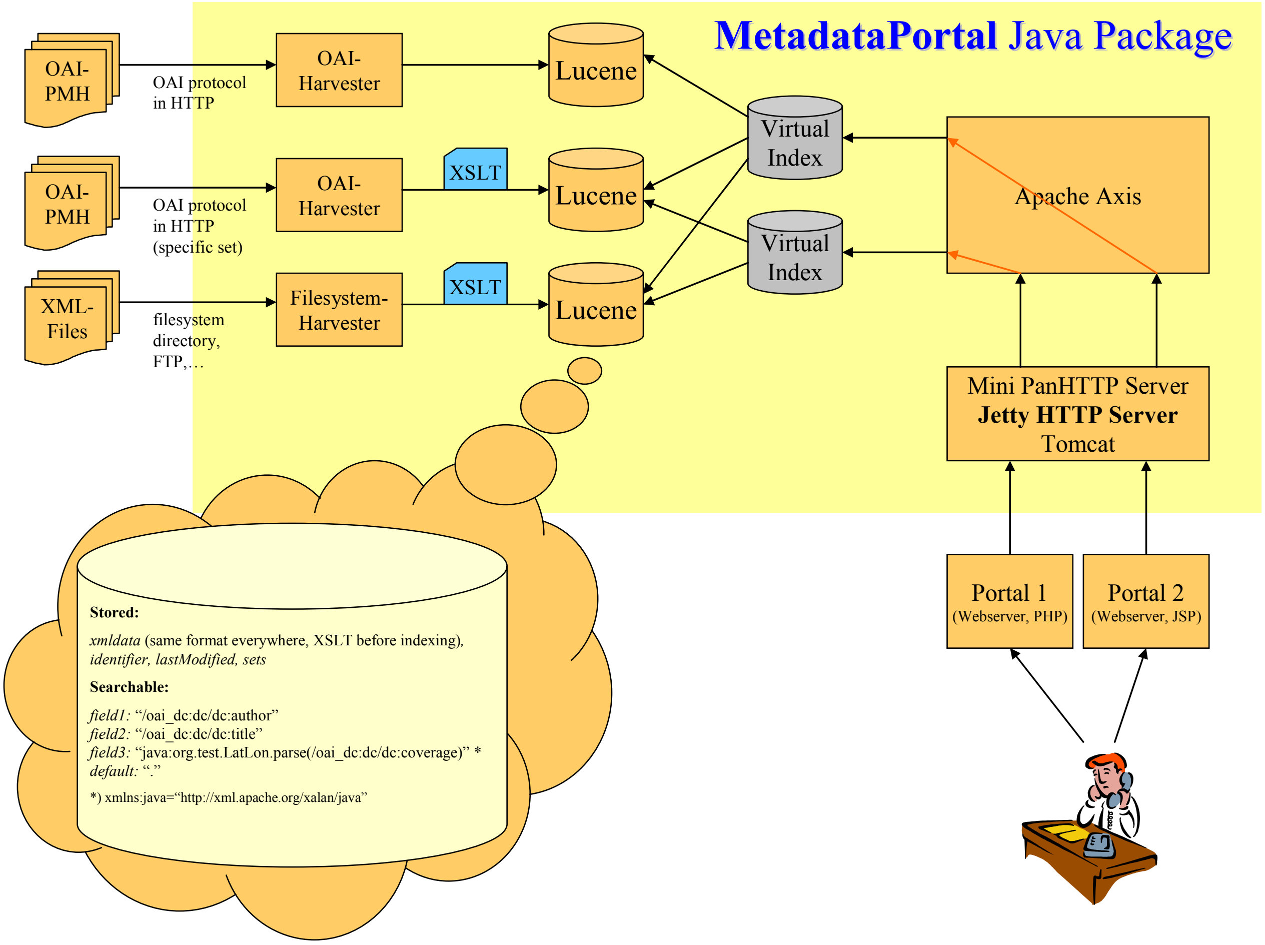
Data Information Service based on Open Archives Initiative Protocols and Apache Lucene

The *World Data Center for Marine and Environmental Sciences (WDC-MARE)* with its data library *PANGAEA* (www.pangaea.de) has experience with data portals for several EU projects (e.g. EUR-OCEANS, CARBOOCEAN,…) to disseminate and publish data and metadata. Based on the needs of these scientific communities we designed a generic portal system architecture suitable for (geo-)scientific data portals.

The software harvests data providers through *Open Archives Initiative (OAI)* protocols using XML metadata in *ISO-19139* or *DIF* format. The current implementations of OAI are limited to Dublin Core metadata only.

The new Java based portal software supports any XML format that can be harvested from OAI-PMH Repositories, file systems or *OGC Catalog Services* (in preparation) and makes them searchable through *Apache Lucene* without any other database software. All datasets are harvested into the index directly without the need to store them separately in the portal.

The open architecture makes it possible to define all searchable fields in several data formats by XPath. This allows not only full text queries, even numerical or date ranges are retrievable. This is achieved by an extension to Lucene that stores the numerical values (even dates are numerical values) in a special format with different precisions in the index. The problem of very slow range queries in the standard Lucene codebase was solved for large indexes with many documents because it is no longer dependent on the index size.

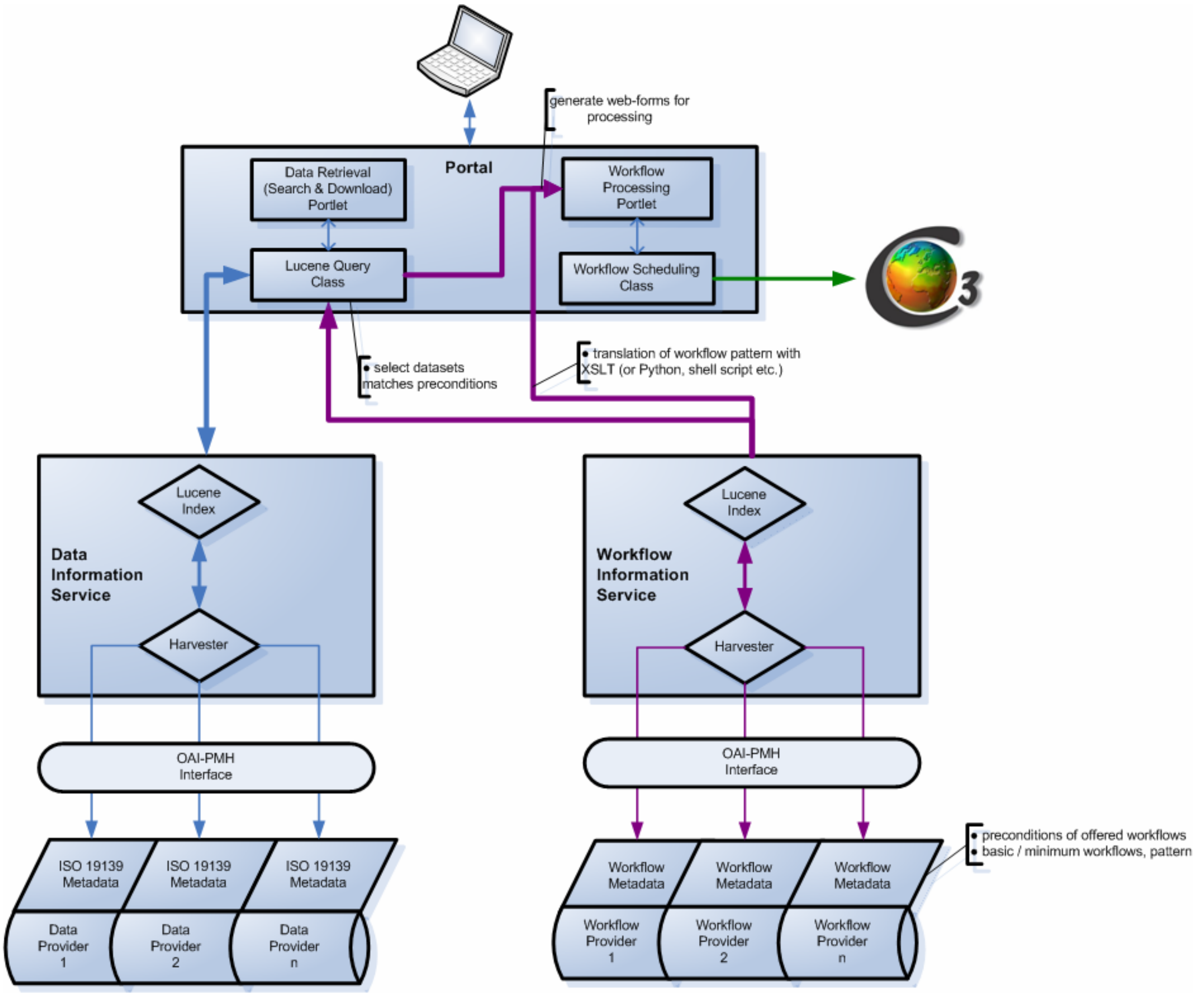

MetadataPortal Java Package
OAI-PMH
OAI protocol in HTTP
OAI-Harvester
Lucene
OAI-PMH
OAI protocol in HTTP (specific set)
OAI-Harvester
XSLT
Lucene
XML-Files
filesystem directory, FTP,…
Filesystem-Harvester
XSLT
Lucene
Virtual Index
Virtual Index
Apache Axis
Mini PanHTTP Server
Jetty HTTP Server
Tomcat
Portal 1 (Webserver, PHP)
Portal 2 (Webserver, JSP)
Stored:
xmldata (same format everywhere, XSLT before indexing), identifier, lastModified, sets
Searchable:
field1: "/oai_dc:dc/dc:author"
field2: "/oai_dc:dc/dc:title"
field3: "java:org.test.LatLon.parse(/oai_dc:dc/dc:coverage)" *
default: "."
*) xmlns:java="http://xml.apache.org/xalan/java"

The metadata of all providers are stored in separate indexes giving the administrator the possibility to manage them separately, but they can be combined to a big "virtual" index for searching. The generic Java-based interface and web service interface allows to support custom front-ends for users and additional visualization in maps.

The portal software will be made freely available through the open source concept when the code base has proven its usability and the design of the programming API for portal implementers is stable.

The *Collaborative Climate Community Data and Processing Grid (C3-Grid)* proposes to link distributed data archives in several German institutions for earth system sciences and to build up an infrastructure for scientists which provides tools for effective data discovery, data transfer and processing. C3-Grid uses the generic portal software for its *Data Information Service (DIS).* The data providers in the grid generate ISO-19139 compliant metadata for the objects in their databases and file systems, make them available by various OAI-PMH repository software (mostly DLESE jOAI software). The grid portal allows the user to search for datasets by full text, variable names, date/time constraints or a bounding box and start jobs on selected datasets.

In a later stage also the possible workflows at various computer centers will be described by metadata and made available in the so called *Workflow Information Service (WFIS).* Likewise for DIS the workflow providers generate a set of workflow information containing a pattern of the offered workflow and the preconditions for the metadata which can be used within the selected workflow(s). This set (XML format) will be also available via OAI-PMH so the generic portal system can retrieve it. The user of the C3-Grid-Portal chooses the workflow he wants to work with, and the WFIS will return all necessary information. The preconditions are based on a DIS-query which returns only those datasets the workflow can handle. The workflow pattern tells the portal how to generate a web-form for the user where he can insert parameters etc. for this workflow. This translation could be made with XSLT, Python or other script languages.


generate web-forms for processing
Portal
Data Retrieval (Search & Download) Portlet
Lucene Query Class
Workflow Processing Portlet
Workflow Scheduling Class
3
select datasets matches preconditions
translation of workflow pattern with XSLT (or Python, shell script etc.)
Lucene Index
Data Information Service
Harvester
OAI-PMH Interface
ISO 19139 Metadata
ISO 19139 Metadata
ISO 19139 Metadata
Data Provider 1
Data Provider 2
Data Provider n
Lucene Index
Workflow Information Service
Harvester
OAI-PMH Interface
Workflow Metadata
Workflow Metadata
Workflow Metadata
Workflow Provider 1
Workflow Provider 2
Workflow Provider n
preconditions of offered workflows
basic / minimum workflows, pattern