INTEGRATING AND MINING DISTRIBUTED ENVIRONMENTAL ARCHIVES ON GRIDS

Mikhail Zhizhin, Eric Kihn, Rob Redmon, Alexei Poyda, Dmitry Mishin, Dmitry Medvedev and Vassily Lyutsarev
ICSU environmental databases

- International Council for Science (ICSU) World Data Centers system founded in 1948

- Fast growing
  - From increasing number of sensors
  - From computational models
  - Mostly as time series = large arrays of numeric data or images

- Distributed and heterogeneous
  - Text and binary files
  - SQL databases with diverse schemata
Example: Space Physics Interactive Data Resource

- Solar activity, solar wind data, geomagnetic, ionospheric, cosmic ray, radio-telescope ground observations, telemetry
- Time span: 1933 – present
Example: Climate Reanalysis Project

- $2.5^\circ \times 2.5^\circ \times 6\text{hr} \times 100$ weather parameters
- Time span: 1948 – present
Problem statement

- There is lots of potentially interesting information inside. How to make it useful for ecologists, social scientists, general public?
Environmental scenario search engine

- **Common data model**
  - Set of linked multidimensional arrays
  - Read or append
  - Selection = hyperslabing

- **Employ existing standards**
  - WSRF for Grid infrastructures
  - WS-I for more general scenarios

- **User-friendly data mining**
  - “Linguistic” terms instead of numeric criteria: environmental scenarios like
    - Magnetic storm
    - Atmospheric front
**Scenario in fuzzy logic**

\[
\text{atmospheric front} = (\text{high } V\text{-wind speed}) \text{ AND } (\text{low } pressure) \text{ AND } \text{SHIFT } (dt=1\text{day}, (\text{low } U\text{-wind speed}) \text{ AND } (\text{low } V\text{-wind speed})) \text{ AND } (\text{high } pressure)
\]
Fuzzy conditions

- Fuzzy truth: any value between 0 and 1
- Fuzzy logic: operations on fuzzy truth
- Fuzzy conditions: assign membership functions to linguistic terms

![Graph showing membership functions for Very small, Small, Average, Large, and Very Large](image-url)
Implementation

- OGSA-DAI: WSRF/WS-I-compatible web service, extendable, distributed workflow
- EsseDataResource data resource
  - Sampling and averaging/interpolation of time series
- GetMetaData activity
  - Enumerates available parameters with their time intervals
- GetXmlData activity
  - Query data resource and output one time series
- FuzzySearch activity
  - Perform environmental scenario search logic

<table>
<thead>
<tr>
<th>Activity</th>
<th>Representation description</th>
<th>Data size, KB</th>
</tr>
</thead>
<tbody>
<tr>
<td>getNetCdfData</td>
<td>Separate stream with binary NetCDF file</td>
<td>925</td>
</tr>
<tr>
<td>getXmlData</td>
<td>Response document with XML. XSD data type: float list</td>
<td>1771</td>
</tr>
<tr>
<td>getXmlData+gzipCompress</td>
<td>Response document with base64 encoded compressed XML</td>
<td>124</td>
</tr>
</tbody>
</table>
Example workflow

 OGSA-DAI container

DataResource
  getXmlData
  getXmlData
  fuzzySearch
  gzipCompress
  deliverToGDT

 OGSA-DAI container

DataResource
  inputStream
  gzipDecompress
  fuzzySearch
  getXmlData
  getXmlData
  gzipCompress
ESSE architecture (Linux platform)

Open Source

ESSE team

MySQL Database

JVM

Tomcat

OGSA-DAI WS-I

OGSA-DAI Activities

ant, axis, xerces, ...

EsseDataResource

GetXmlData

Fuzzy

GzipCompression

GzipDecompression

MySQL

JDBC

MySql

Fuzzy Engine
ESSE on a Microsoft platform

- Open Source
- ESSE team (common)
- ESSE team (C#)
- Microsoft

.NET Framework 2.0

IIS

OGSA-DAI .NET

OGSA-DAI Activities

Comptibility libraries

EsseDataResource

GetXmlData

Fuzzy

GzipCompression

GzipDecompression

JDBC

ODBC

SQL Server Database

Fuzzy Engine
User interfaces

Web portal

NASA World Wind
Conclusions

- OGSA-DAI framework can be used as an interface to archives of environmental data
  - Accessible as web service and on grids
    - More users
    - Interdisciplinary applications
  - Incorporating non-standard data mining algorithms into data access infrastructure