GADS*: Using Web Services to access large data sets

Jon Blower, Keith Haines, Chunlei Liu, Andrew Woolf, Kecheng Liu, Adit Santokhee, Ying Zhao

* Grid Access Data Service
Background

- Climate scientists have a need to access large datasets:
  - Model data and satellite observations
  - Data in a variety of formats (netCDF, HDF, GRIB, more), grids, naming conventions
  - Model intercomparisons (MERSEA)
- Existing standards (DODS/OPeNDAP) are limited
Working with large data sets

- Subset / resample
- Transform / regrid / rotate
- Analyse
- Compare
Background (2)

- Woolf et al (2003) presented a prototype version of GADS as an alternative to DODS
- Developed as part of GODIVA (NERC eScience pilot project)
Advantages of GADS

• Data are abstracted from storage
• Data can be exposed with standard variable names, even if data files do not conform to standards
• Data can be delivered in many formats, irrespective of internal storage format
• Behaves as aggregation server
• Deployed as Web Service
  – Platform – independent
  – Compatible with current eScience advances
GODIVA Web Portal

• Allows users to interactively select data for download using a GUI

• Users can create movies on the fly

• cf. Live Access Server
GADS Methods

- `dataQuery` is used for querying the data holdings, e.g. “What datasets are there?”, “What variables are contained in the X dataset?”

- `dataRequest` is used to download data
  - Can easily extract subsets of data
Recent Enhancements

• More portable (now works on Linux)
• More flexible and extensible (object oriented code)
• Improved metadata handling:
  – Metadata manager
  – Metadata loosely coupled from rest of system, so many different representations (e.g. XML, mySQL) are possible.
Architecture

GADS Web Service

Client

dataQuery

dataRequest

DATA FILES

Metadata Web Service

META-DATA

Metadata Manager Utility
Metadata manager
Workflows

- As a Web Service, GADS can be incorporated into workflows (aka orchestration, choreography)
- See presentation by Sastry and Craig of RAL
- A few standards and fewer tools are available (BPEL4WS, WSCI, SCUFL, LabVIEW)
- Question: How pass around large data sets?
  - Don’t pass data, pass *pointers*, e.g. URLs?
Example workflow

Scufl Workbench (myGrid project)
http://taverna.sourceforge.net

• GADS is used as the data extraction step in a simple data processing workflow

• The other web surface is finding the depth of an isotherm in the dataset
... the result

- reference GODIVA, understanding of THC, heat budgets
Future plans

- Potential deployment at SOC, POL, Met Office, ECMWF...
- Use document-style SOAP for richer queries, e.g. “What datasets contain salinity measurements in the South Atlantic?”
- More output formats (GRIB, VTK?)
- Long-term migration path towards Grid Services
  - Asynchronous notification of, e.g. progress