The Data Portal

Glen Drinkwater
Motivation

- Council for the Central Laboratory of the Research Councils’ has many large scale facilities.
- E.g. ISIS, a pulsed neutron source and Diamond, the UK’s new synchrotron light source.
- Data Portal
  - Making existing scientific data resources accessible through a single interface.
  - Acting as a broker between scientists, facilities and data.
- Benefits
  - Repetition of experiments can be avoided.
  - Collaborations can be built by identifying that someone else is working in a similar area.
  - Data about a related material can be found and used to aid a new analysis.
  - Data can be reanalysed when better analysis tools becomes available.
General Architecture

CCLRC Data Portal

- Local data
- Local metadata

Facility 1

Xml Wrapper

Facility 2

Xml Wrapper

Facility N

Xml Wrapper

Other Data Portal instances

CCLRC Data Portal

- Local data
- Local metadata
Core Modules

- Web Interface, Query and Reply, Lookup and Help.
- Important function grouped into modules, each modules with a web service interface, interface description in WSDL and communicate via SOAP.
Additional Modules

- Access Control, Authentication and Authorisation, Data Transfer, Shopping Cart, User Administration, Facility Administration and Accounting.

- Functions grouped into modules, each with web services interface, interface description in WSDL, communication via soap. These Modules could be shared or exchanged with their own implementation.
External Services

- XML-Wrapper, RasDaMan Insertion and Extraction, Specialised Search, HPCPortal, Visualisation Portal, SRB, other DataPortal Instances.
- Other services that are linked with the DataPortal, but are not integral part of it. Registered with the Portal and accessible via web services interface.
Authentication

1: get certificate(name, passphrase)

2: getUserPrivileges (certificate)

3: startSession(certificate, permissions)

4: return SID

5: getPermissions(SID)

Authentication

User

Web Interface

Session Manager

START: Login(name, passphrase, lifetime)

START: Login(name, passphrase)

MyProxy

Access & Control Facility A

Access & Control Facility B

Session Manager

START: Login(name, passphrase)

SESSION MANAGER DATABASE

FACILITY A ACCESS & CONTROL DATABASE

FACILITY A ACCESS & CONTROL DATABASE

Set permissions in database for all facilities for duration of session

Get permissions from database for facility A

Set permissions in database for all facilities for duration of session

OUTSIDE SERVICE
Basic Search

START: basic query (list of facilities, discipline)

1: getPermissions(SID)

Web Interface

User

FINISH: XML results in CLRC scientific metadata format

FINISH: XML results in CLRC scientific metadata format

Session Manager

Query & Reply

2: getXML (discipline)

XML Wrapper

XML Wrapper

FACILITY A metadata archive

FACILITY B metadata archive

OUTSIDE SERVICE

SQL selects

FACILITY A

FACILITY B
**Shopping Cart**

**START:** Add item to Shopping Cart

1: addToCart(SID, item)

2: getCert(SID)

3: add item to database (DN, item)

**FINISH:** success/fail

4: success/fail
• Another e-Science Centre project to develop a Web portal to search for resources and submit HPC applications to a computational Grid.
• Uses Globus toolkit v2.2
• Functionalities include:
  • Resource Management: GRAM.
  • Information Services: MDS.
  • Data Management: GridFTP and GASS.
  • All use GSI security protocol as the connection layer.
Integrated Portals

All-Hands Meeting
2-4th Sept 2003
e-Science Centre

12

- DataPortal
- Web Services
- HPCPortal
- Web Services
- GridFTP
- Web Services
- Data Systems
- Visualisation
- GSI
- HPC Systems
- Web Services
- Working with GGF Grid Computing Environments
- Research Group
Signal Sign on

- How do you have signal sign on?
- Both HPC and DataPortal have their own Session Managers which rely on Globus Proxy Certificates.
- Integrated session managers communicate over SSL using mutual authentication between the web servers.
- Allows certificates to be transferred between portals allowing single sign on.
- The certificate can then be used for GSI authentication.
Single Sign on

START: Log on to DataPortal then to HPC Portal

1: Login(username, password, lifetime) Via Authentication Module

2: Dataportal SID

3: LoginHPC(SID)

4: isValid(SID)

5: RequestCert(SID)

6: Certificate

7: YES!

8: HPC Session id

FINISH: User is sent to HPC front page to use its services
Scenario

- User logs on to Data Portal and searches for data.
- The data found is added to the persistent shopping cart.
- The user could then transfer the data to another machine using GSI FTP, either using the Data Portal or the HPC Portal.
- Using single sign on, the user could then go directly to the HPC, and the use a remote job run on the data that they have just transferred and using GSIFTP, transfer the results back to their machine for analysis.
• Converting to a new Xml Wrapper architecture
• Problems:
  • DTD limits us to one investigation per study.
  • Any new wrappers should be written using the new standard.
  • Only allows one investigation per study (as using the old format).
  • XSLT time consuming to write and must operate on the whole study.
  • Need to transfer the whole study – this can be large.
  • Data Archive goes down we cannot access study even though it is cached.
Old Xml
Wrapper Architecture

DataPortal

Studies in XML
Csmd v1 format

Old Xml
Wrapper Architecture

SQL ‘like’ where clause

Xmlwrapper

cache

SQL statements and replies

Data Archive
Solution

- DataPortal (via Q&R)
  - XQuery

XML Wrapper Framework

- XMLwrapper: Doc Selector
- Cache

XMLwrapper: Doc Builder

- SQL Selects and Results
- Mapping between Archive schema and CSMD format

- Update Repository with New and changed CSMD XML docs

- Data Archive

XML Document Repository

XQuery & Result Returned
Advantages

• Xml Wrapper still operates if archive is down.
• Concurrency much better supported as selecting from the archive is essentially one thread.
• Only relevant data sent between xml wrapper and Data Portal – also XQuery can include formatting rules which can potentially do away with the need to XSLT.
Questions?