OGSA-DAI Overview

Neil P Chue Hong

Goals for this presentation

- Understand data access scenarios on the Grid
- Describe how the Grid influences data access and integration
- Describe an overview of the OGSA-DAI software

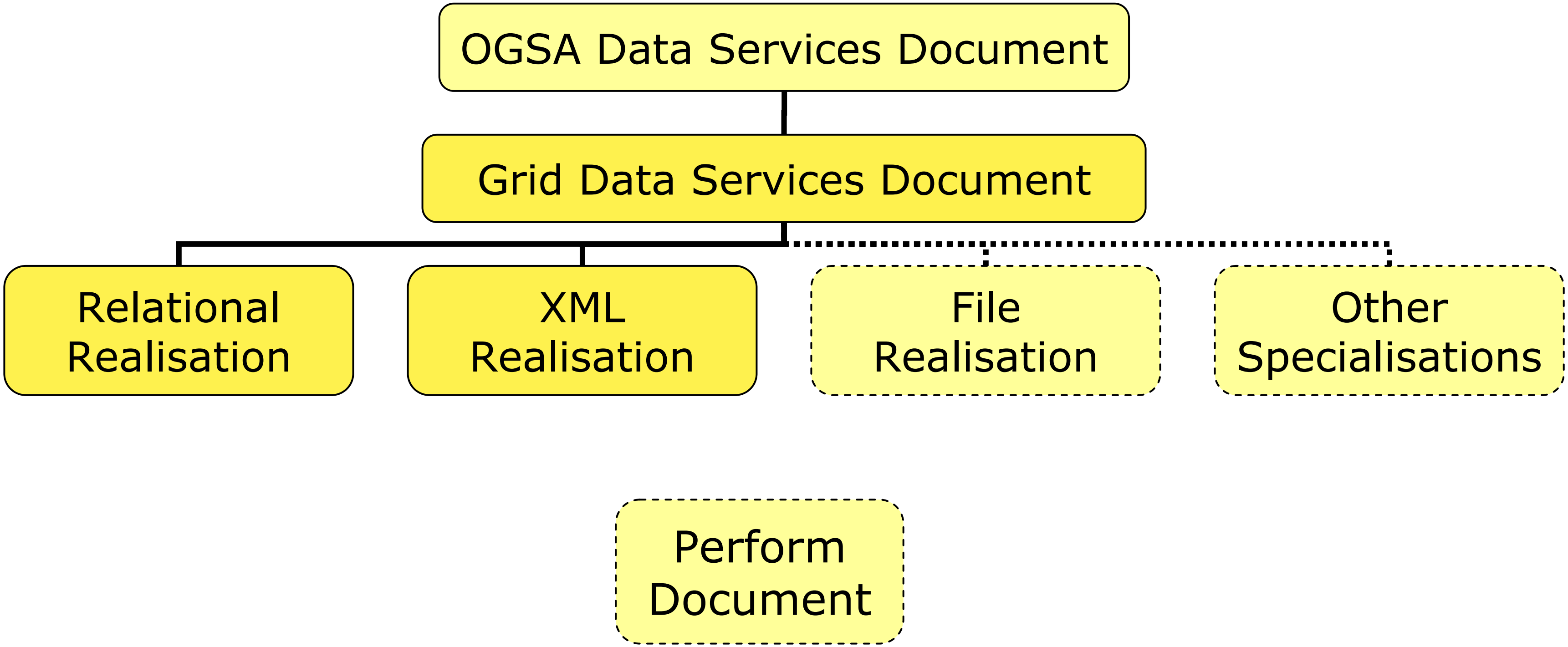
Terminology - Data

- Data Resource
  - Any object that can source/sink data
  - Currently databases in scope
- Data Service
  - Common interface to a data resource
  - Exposes capabilities of data resource
    - SQL Queries, X-Path Queries
  - May provide additional capabilities
    - Data transformations, 3rd party data delivery
- OGSA-DAI
  - Open Grid Services Architecture Data Access and Integration

Motivation

- Entering an age of data
  - Data Explosion
    - CERN: LHC will generate 1GB/s = 10PB/y
    - VLBA (NRAO) generates 1GB/s today
    - Pixar generate 100 TB/Movie
  - Storage getting cheaper
- Data stored in many different ways
  - Data resources
    - Relational databases
    - XML databases
    - Flat files
- Need ways to facilitate
  - Data discovery
  - Data access
  - Data integration
- Empower e-Business and e-Science
  - The Grid is a vehicle for achieving this

DAIS Documents


OGSA Data Services Document

Grid Data Services Document

Relational Realisation

XML Realisation

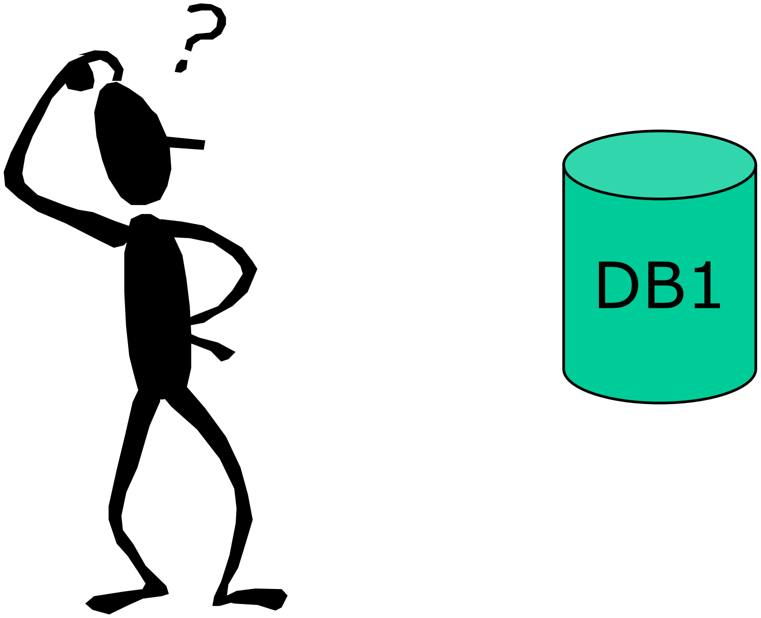
File Realisation

Other Specialisations

Perform Document

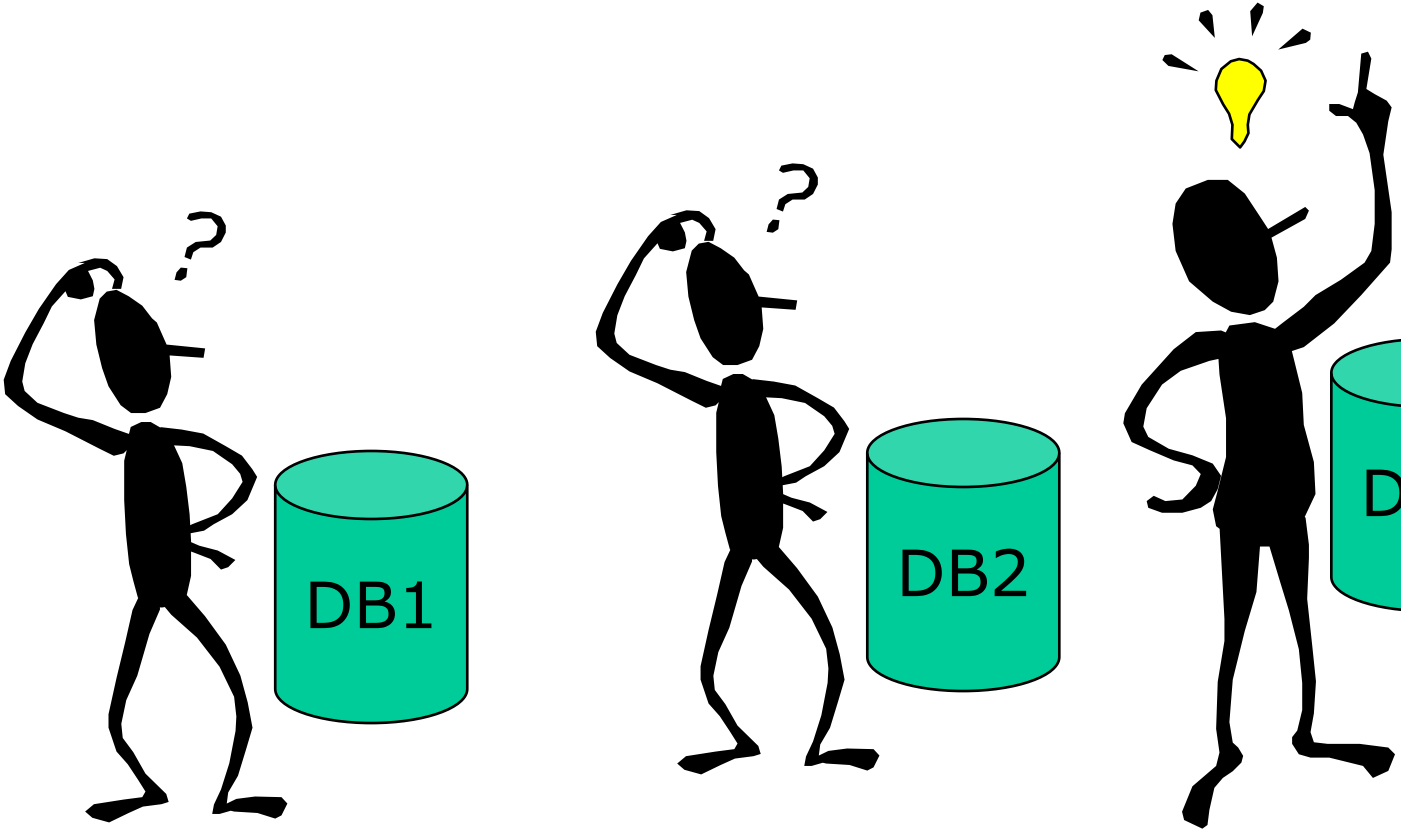
Why is the Grid necessary?

- If I am a researcher with my own database, why do I need the Grid?



It's all about sharing

- You can never have it all...



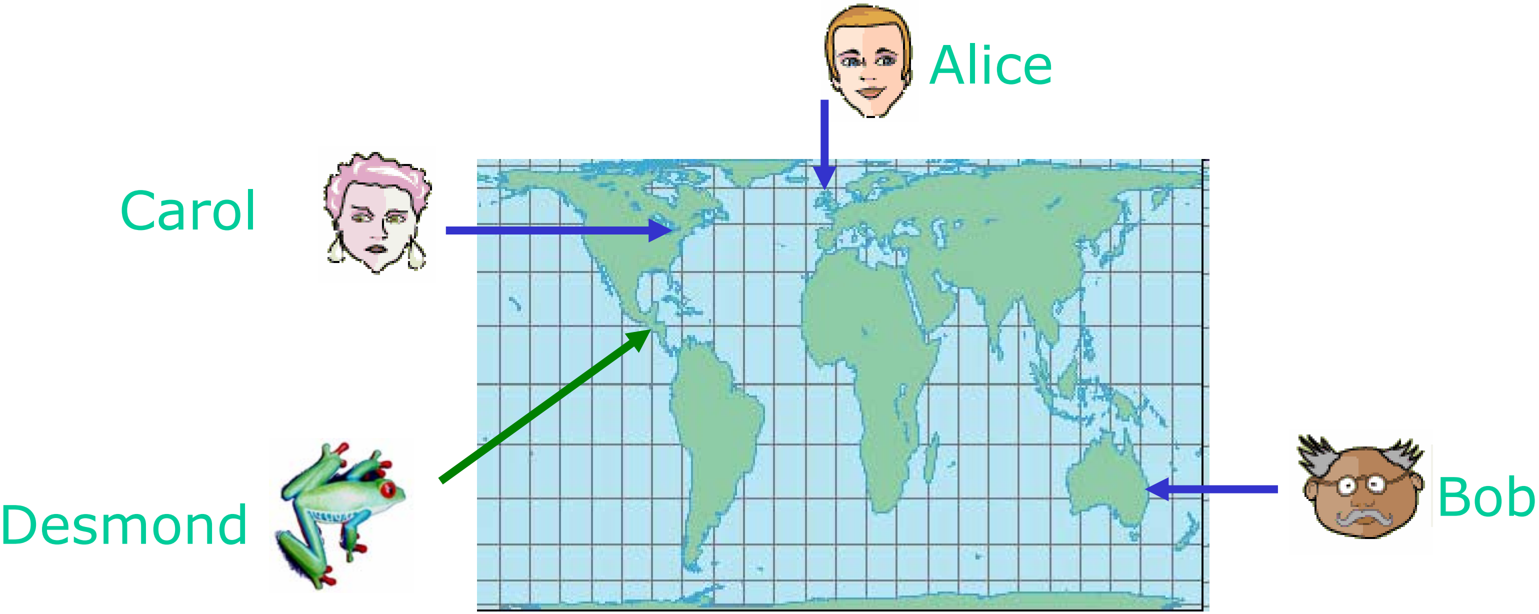
Scenario: Red Eyed Tree Frogs

The story of Alice, Bob, Carol
and a frog called Desmond

Thanks to Tom Sugden and Martin Westhead for the original idea

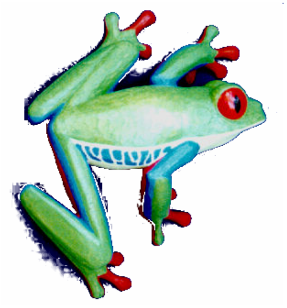
Once upon a time…

- In this story, we will learn how Data Access and Integration Services helped:



Use Case: Publishing

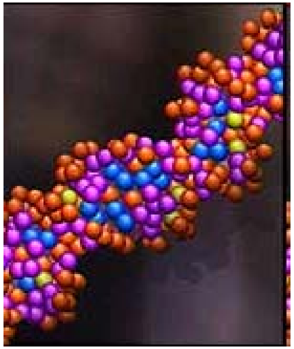
- Alice is a molecular biologist
  - Based at the University of South Edinburgh
  - Mapped the genetic sequence of the Red-Eyed Tree Frog



- Alice wants to make her work available to the scientific community
  - Publish a read-only on-line database
  - Register data resource with a public registry

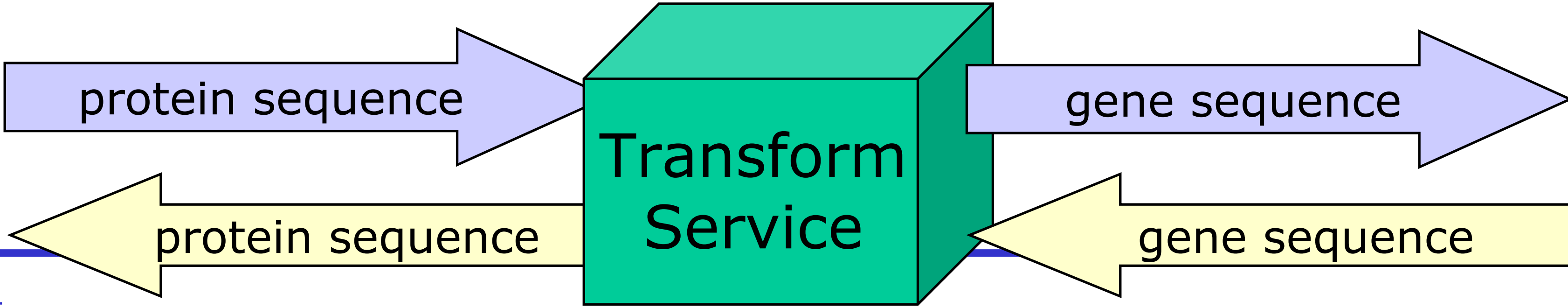
Use-case: Remote update

- Bob is a Professor of Biology
  - Based at the Organisation for Gene Sequencing in Australia
  - Working in collaboration with Alice on the Red-Eyed Tree Frog genome
  - Alice provides a secure private read/write grid data service
- Through Alice's services
  - Bob can contribute new sequences



Use Case: Transformations

- Carroll is a biochemist
  - Works for a small drugs company called DrugsRUs in Aurora, Illinois.
  - Investigating toxin in saliva of Fire Bellied Toad
- Wants to compare proteins with Red Eyed Tree Frog
  - Carroll has a protein sequence
  - Alice's data is encoded as a gene sequence



protein sequence

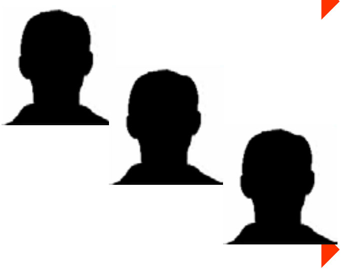
Transform Service

gene sequence

protein sequence

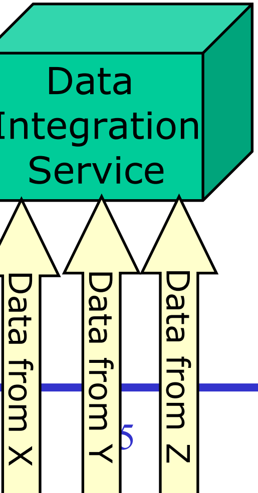
gene sequence

Use Case: Data Integration



- X, Y and Z are other scientists
  - They publish their work as read-only data resources
  - Z only allows specific queries to be run

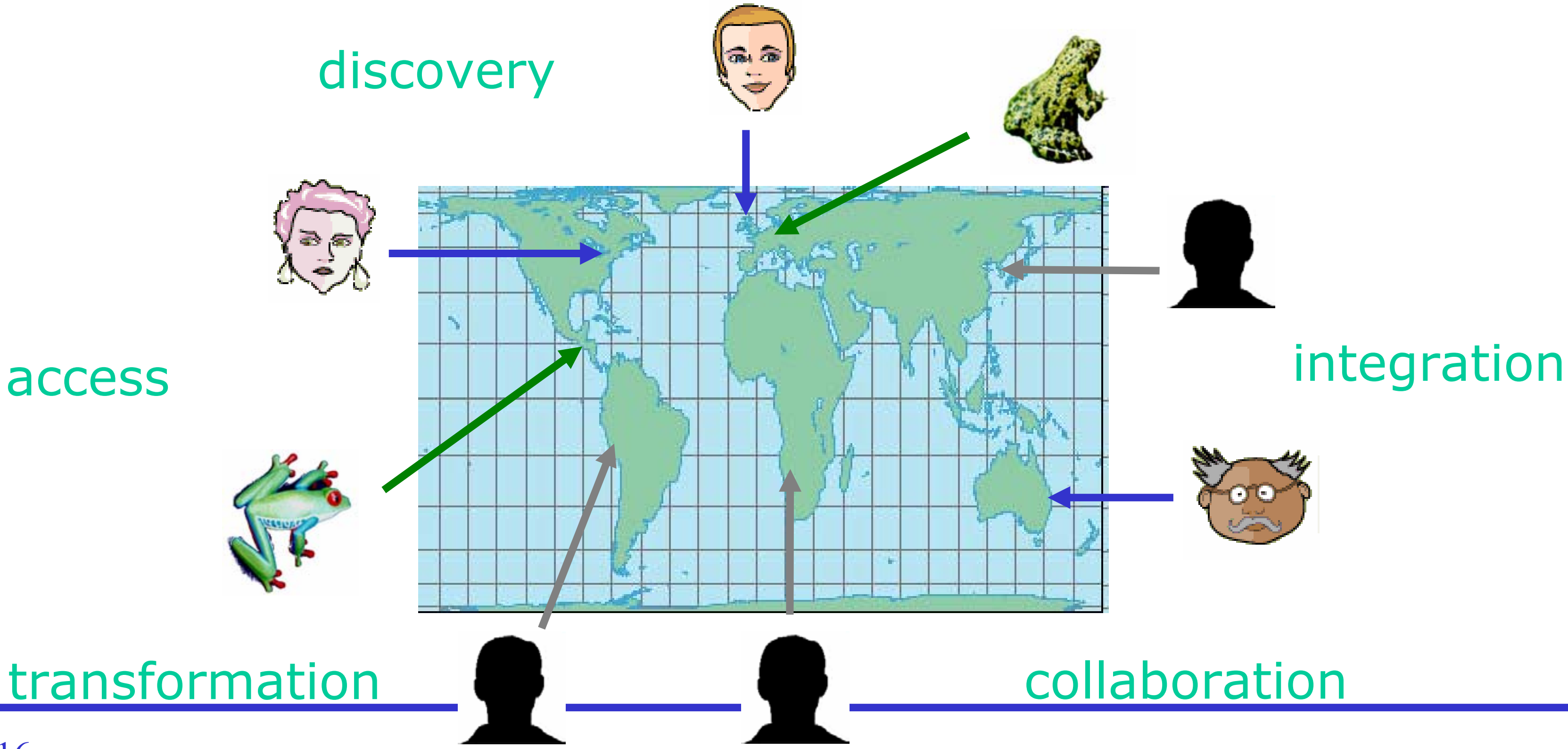
- Alice, Bob and Carol each want to use subsets of data from X, Y, and Z
  - Trying to save the nearly extinct variegated red-eyed tree frog
  - Alice writes a service which exposes a integrated set of data as another virtual data resource
  - Bob and Carol can use this resource as if it were a single data resource



- They find a way to save Desmond!

The End

- Use OGSA-DAI to provide the middleware tools to grid-enable existing databases



discovery

access

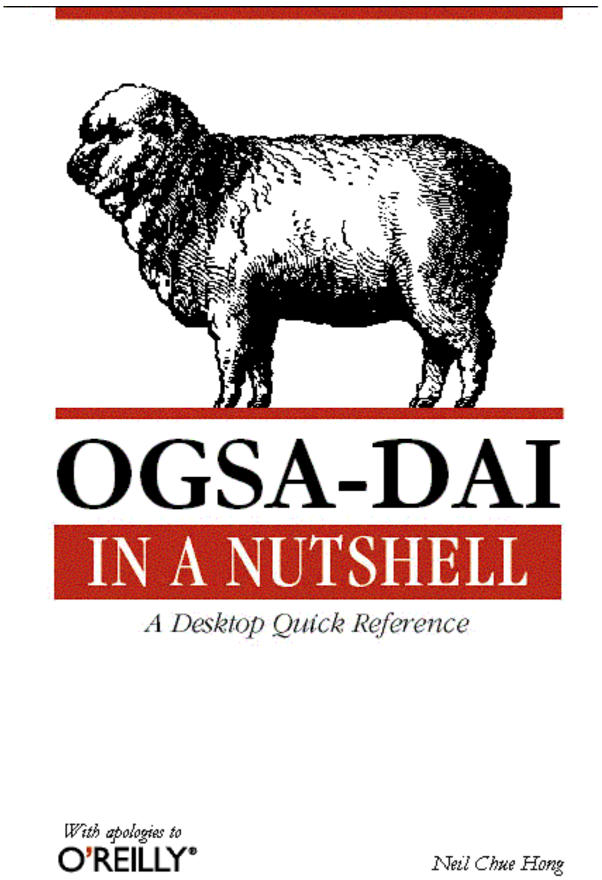
integration

transformation

collaboration

OGSA-DAI in a Nutshell

- All you need to know about OGSA-DAI in a handy pocket sized book!
- Updated for Version 3.1



OGSA-DAI

IN A NUTSHELL

A Desktop Quick Reference

With apologies to
O'REILLY®

Neil Chue Hong

OGSA-DAI Project

- Develop a component library
  - Access and manipulate data in a grid
  - Serve UK and International e-Science communities
- Aims to provide
  - Common interface to data resources
  - Simple integration of distributed queries to multiple data resources
- Contribute to standardisation efforts
  - Input into GGF DAIS WG and other groups
  - Provide a reference implementation of DAIS spec
- Based on Open Grid Services Architecture (OGSA)
  - Globus Toolkit 3 (GT3) “compliant”

Project Partners

Powered by ….

unded by the Grid Core Programme

OGSA-DAI

£3 million, 18 months, from Feb 2002

Three major releases, three interim releases

DAIT (DAI-Two)

Keep the OGSA-DAI brand name

£1.5 million, 24 months, from Oct 2003

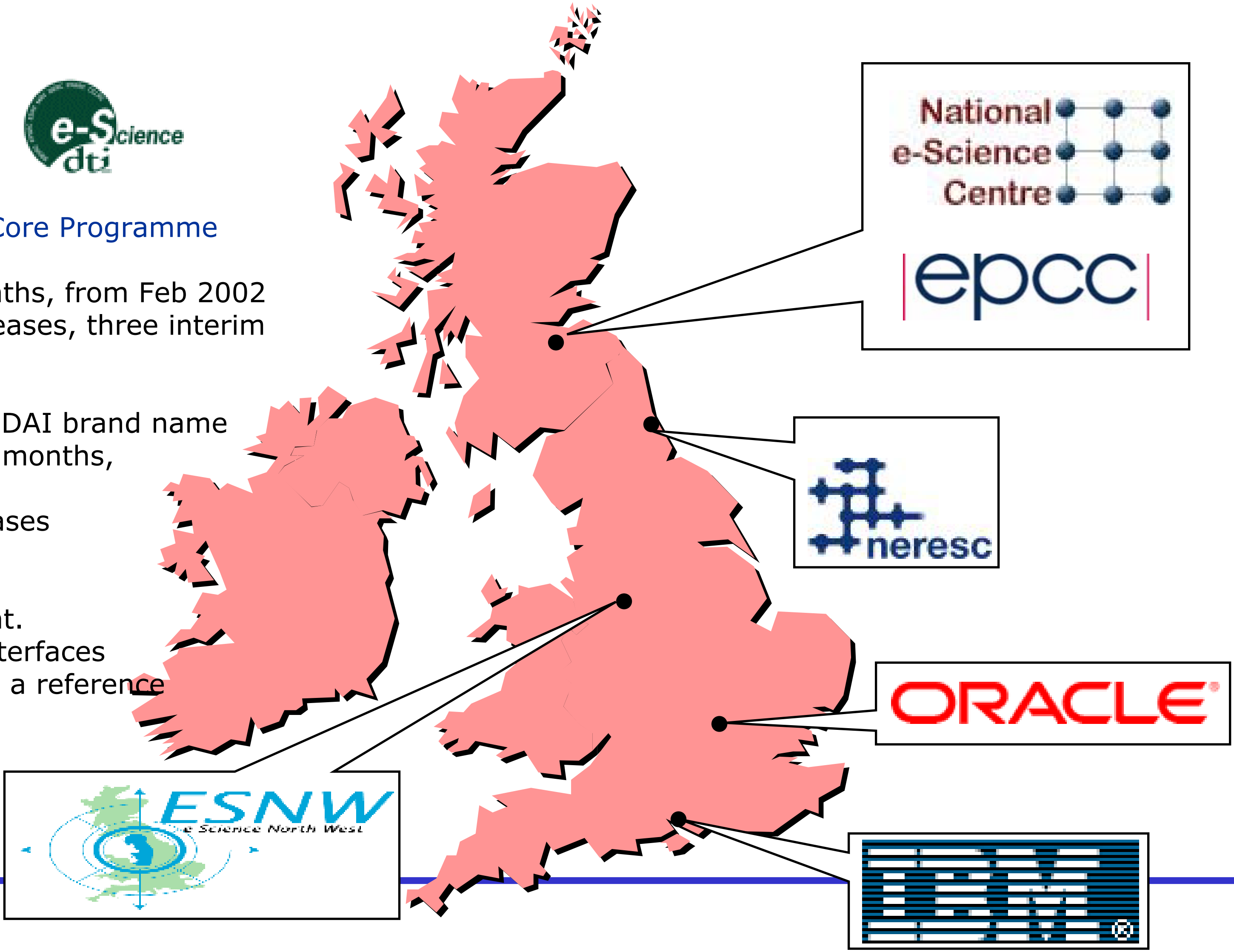
Four major releases

GGF DAIS WG

Strong involvement.

Standardise the interfaces

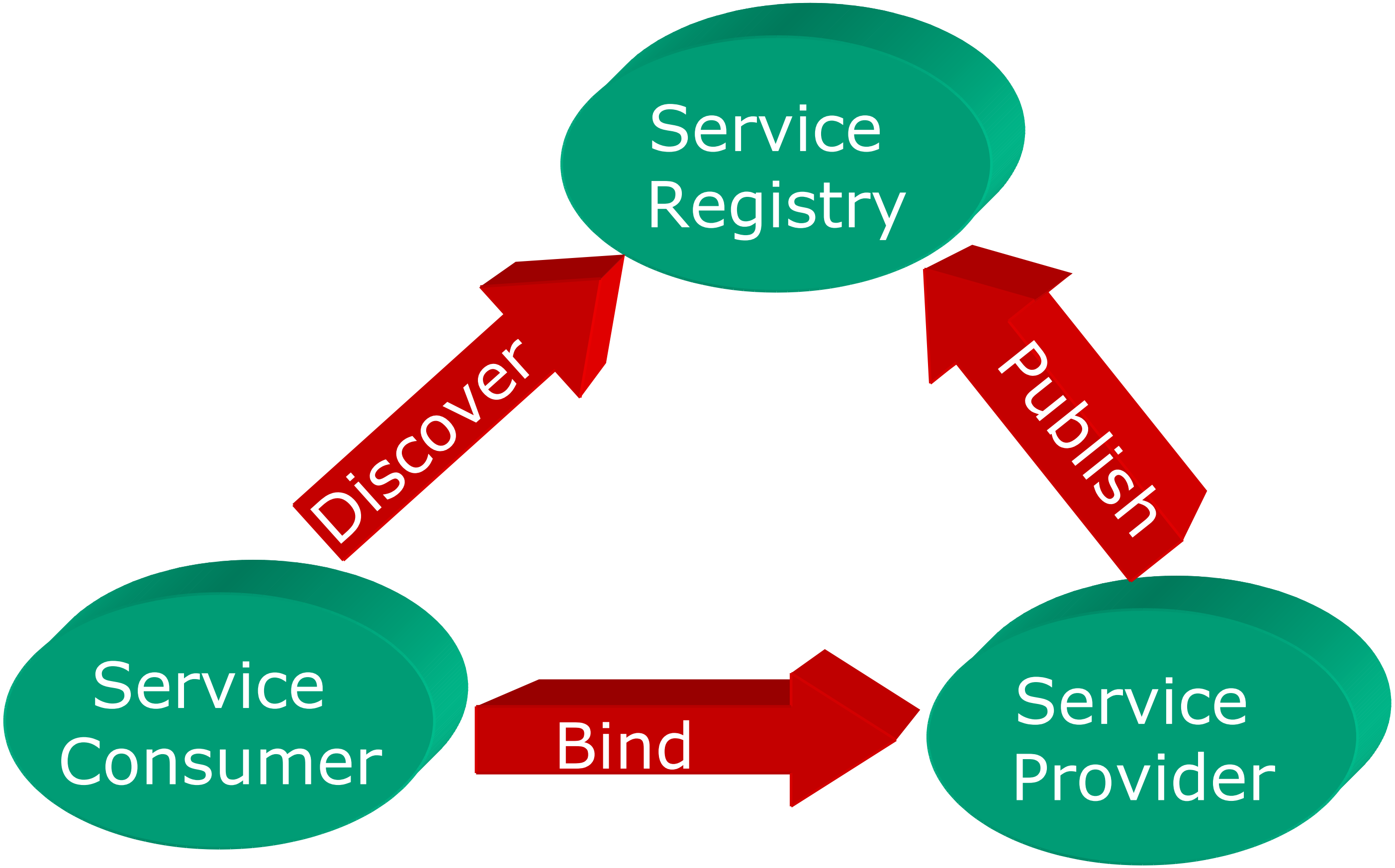
OGSA-DAI to be a reference implementation



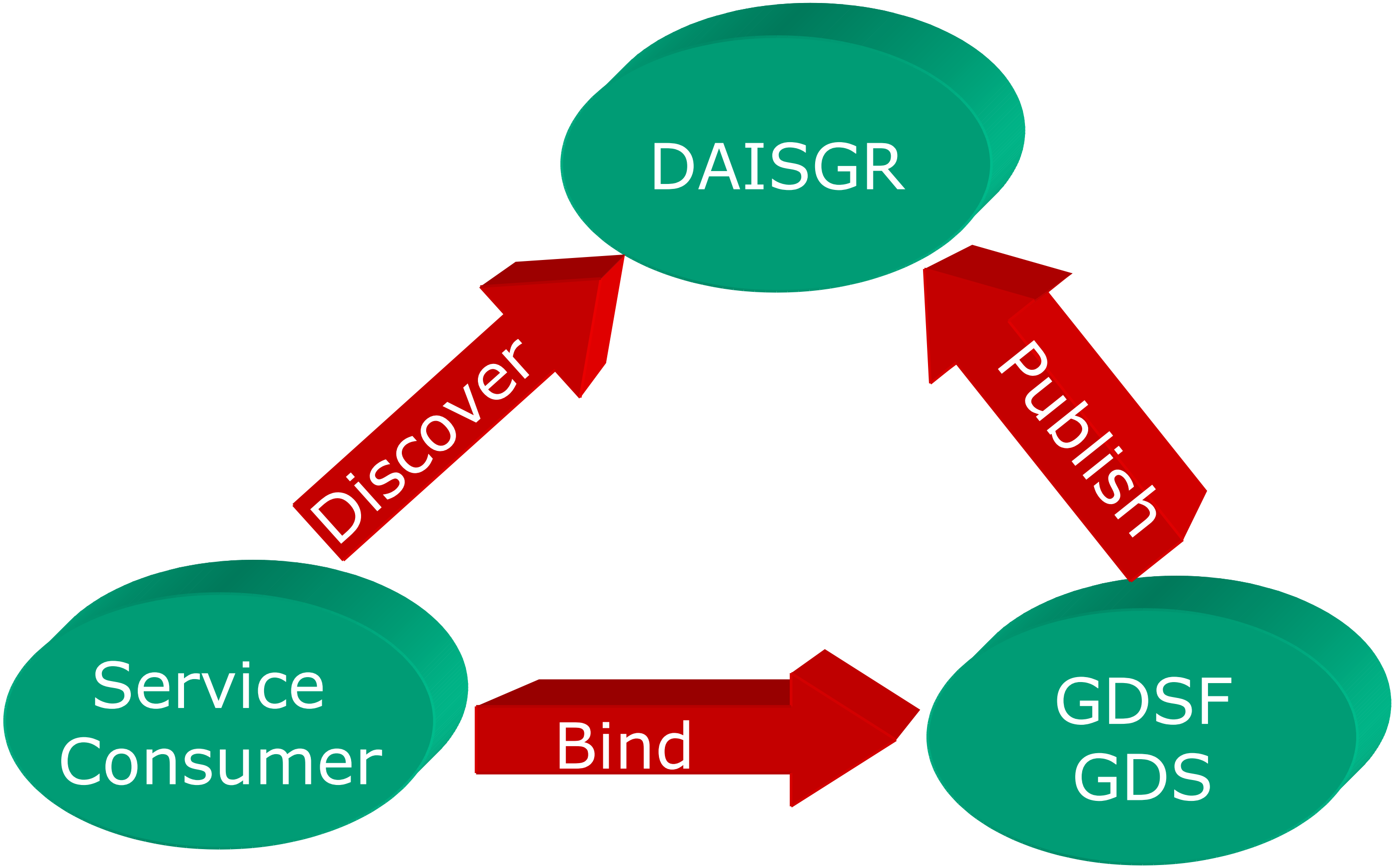
Project Status

- Current release 3.1
  - Globus Toolkit 3.0.2 or 3.2 compliant
  - Platform and language independent
    - Java 1.4
    - Document model
- Work concentrated on data access
  - Wraps data resources without hiding underlying data model
  - Provide base for higher-level services
    - Distributed Query Processing (DQP)
    - Data federation services

Web Service Architecture

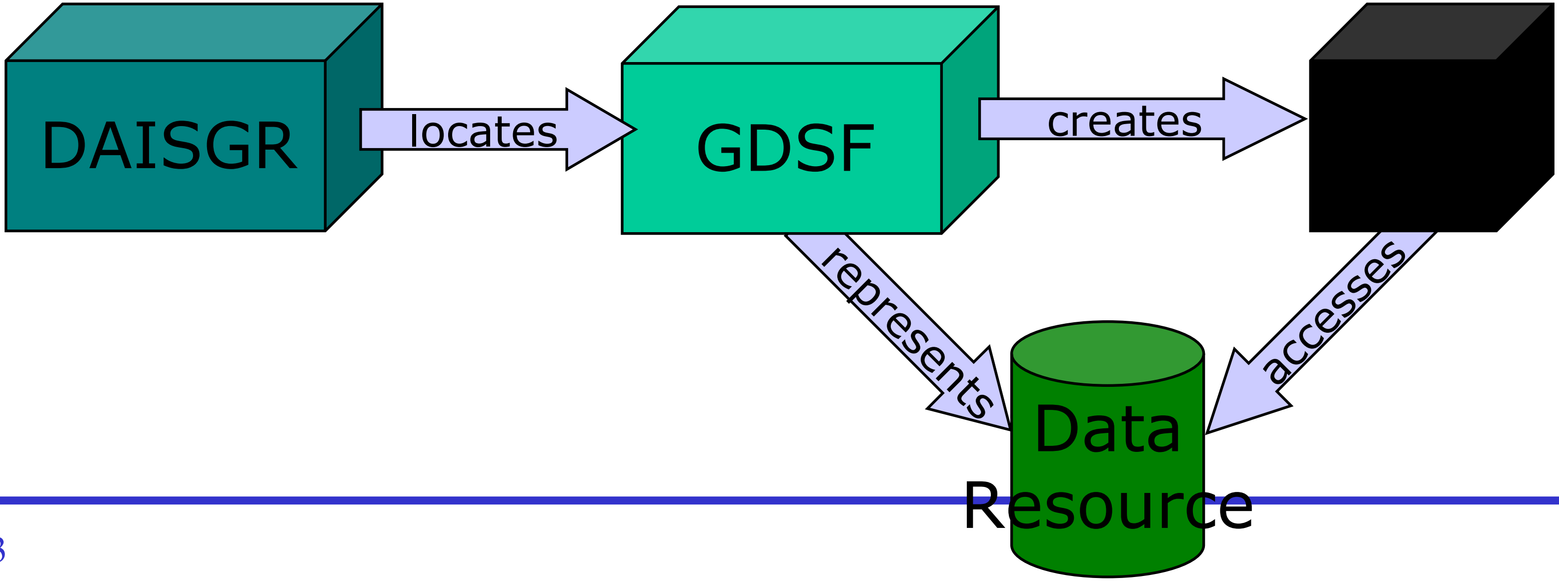


OGSA-DAI Service Architecture



OGSA-DAI Services

- OGSA-DAI uses three main service types
  - DAISGR (registry) for discovery
  - GDSF (factory) to represent a data resource
  - GDS (data service) to access a data resource



GDSF and GDS

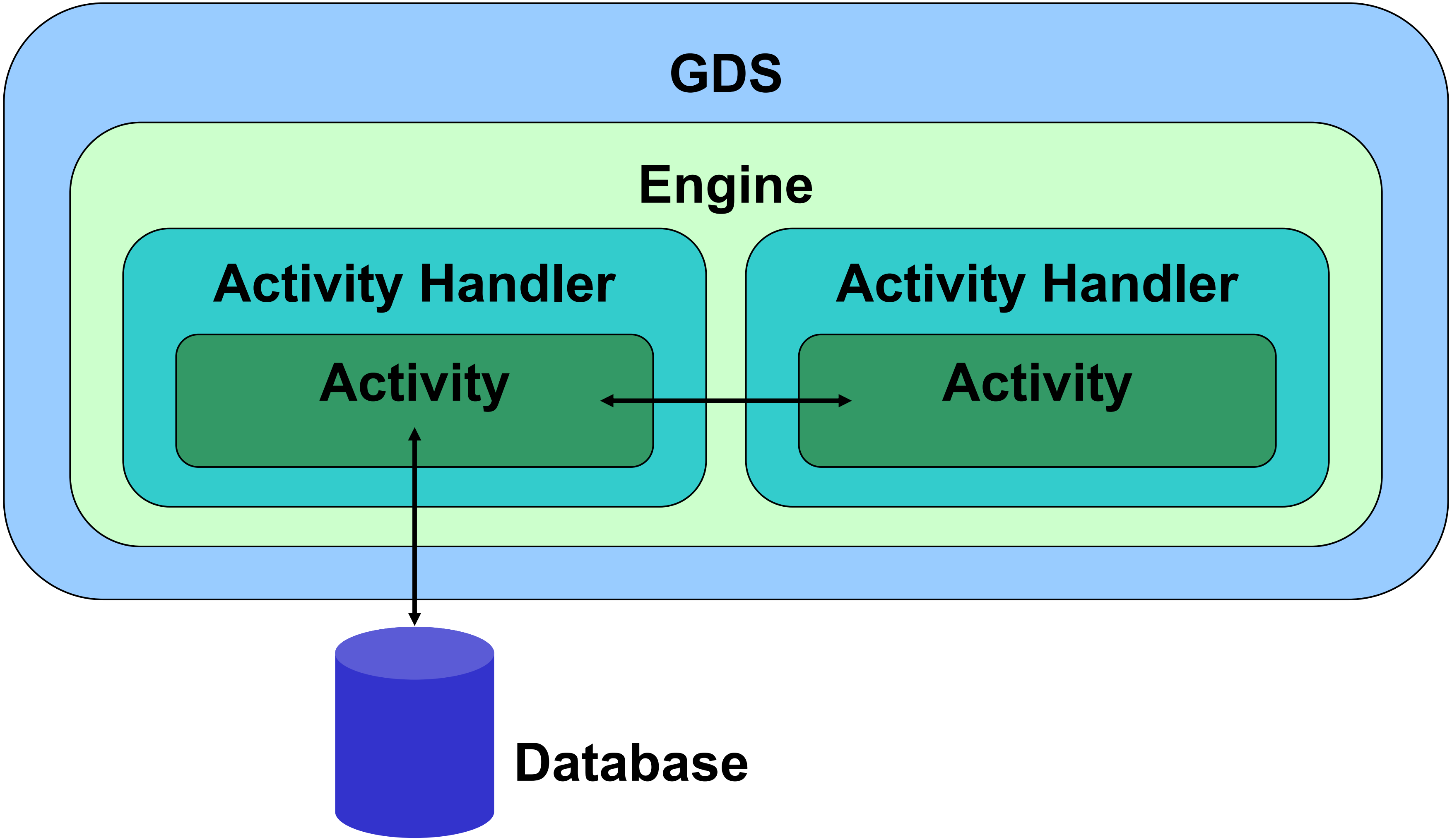
- Grid Data Service Factory (GDSF)
  - Represents a data resource
  - Persistent service
    - Currently static (no dynamic GDSFs)
      - Cannot instantiate new services to represent other/new databases
  - Exposes capabilities and metadata
  - May register with a DAISGR
- Grid Data Service (GDS)
  - Created by a GDSF
  - Generally transient service
  - Required to access data resource
  - Holds the client session

- DAI Service Group Registry (DAISGR)
  - Persistent service
  - Based on OGSI ServiceGroups
  - GDSFs may register with DAISGR
  - Clients access DAISGR to discover
    - Resources
    - Services (may need specific capabilities)
      - Support a given portType or activity

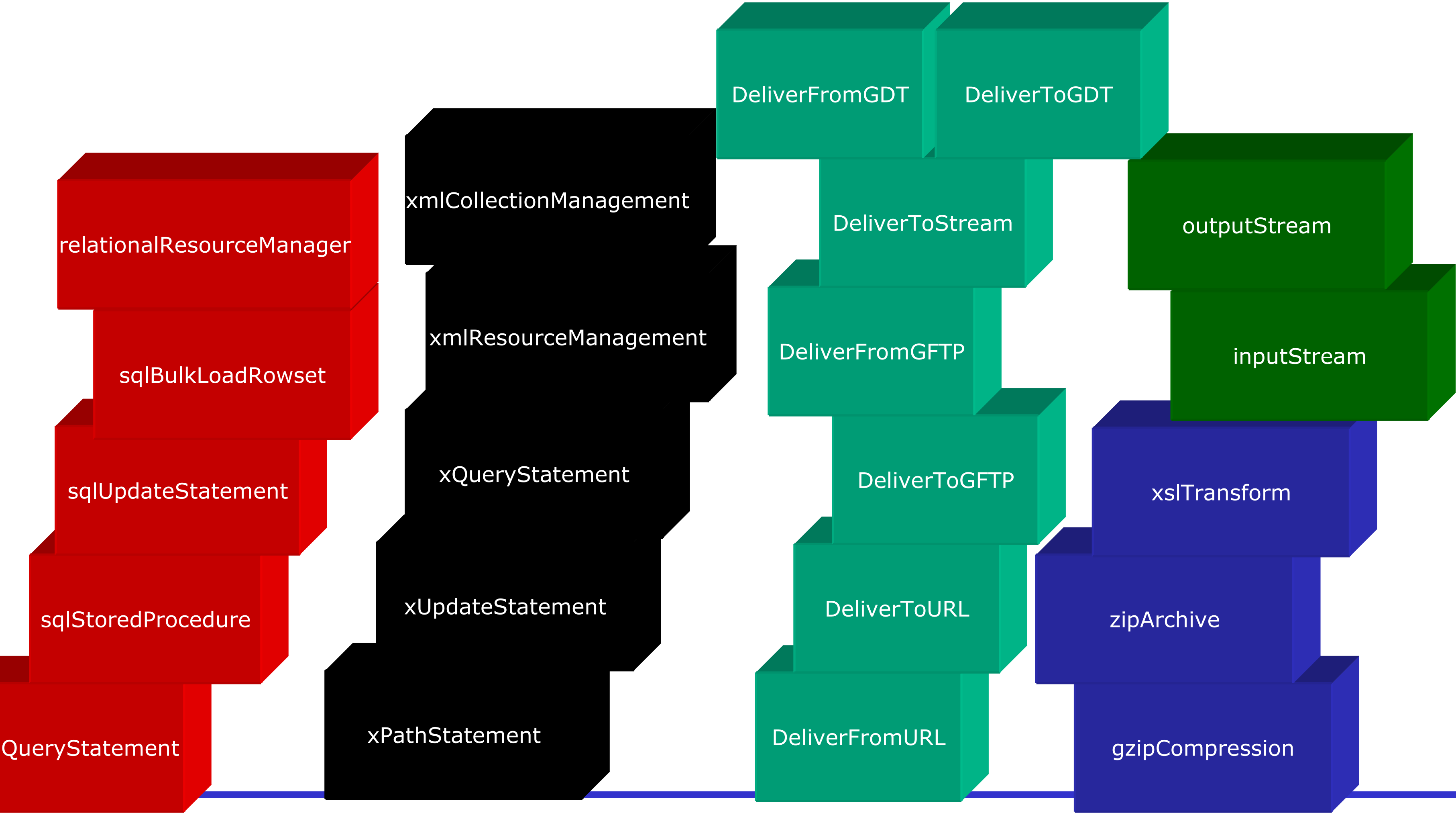
Supported Data Resources

Relational		XML		Other	
MySQL	➡	Xindice	➡	Files	⇨
DB2	➡	eXist	?		
Oracle	➡				
PostgreSQL	⇨				
SQLServer	⇨				

GDS Internals



Predefined Activities



relationalResourceManager

sqlBulkLoadRowset

sqlUpdateStatement

sqlStoredProcedure

QueryStatement

xmlCollectionManagement

xmlResourceManagement

xQueryStatement

xUpdateStatement

xPathStatement

DeliverFromGDT

DeliverToGDT

DeliverToStream

DeliverFromGFTP

DeliverToGFTP

DeliverToURL

DeliverFromURL

outputStream

inputStream

xslTransform

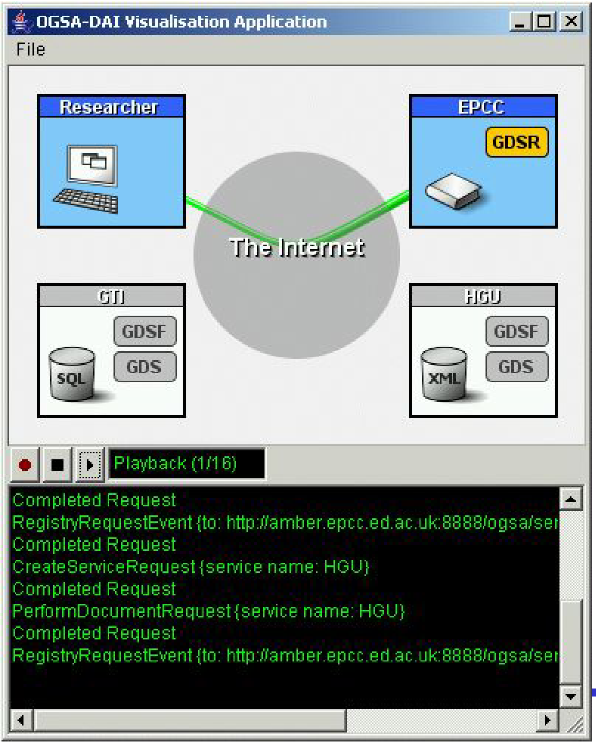
zipArchive

gzipCompression

Client Toolkit

- Why? Nobody wants to write XML!
- A programming API which makes writing applications easier
  - Now: Java
  - Next: Perl, C, C#?

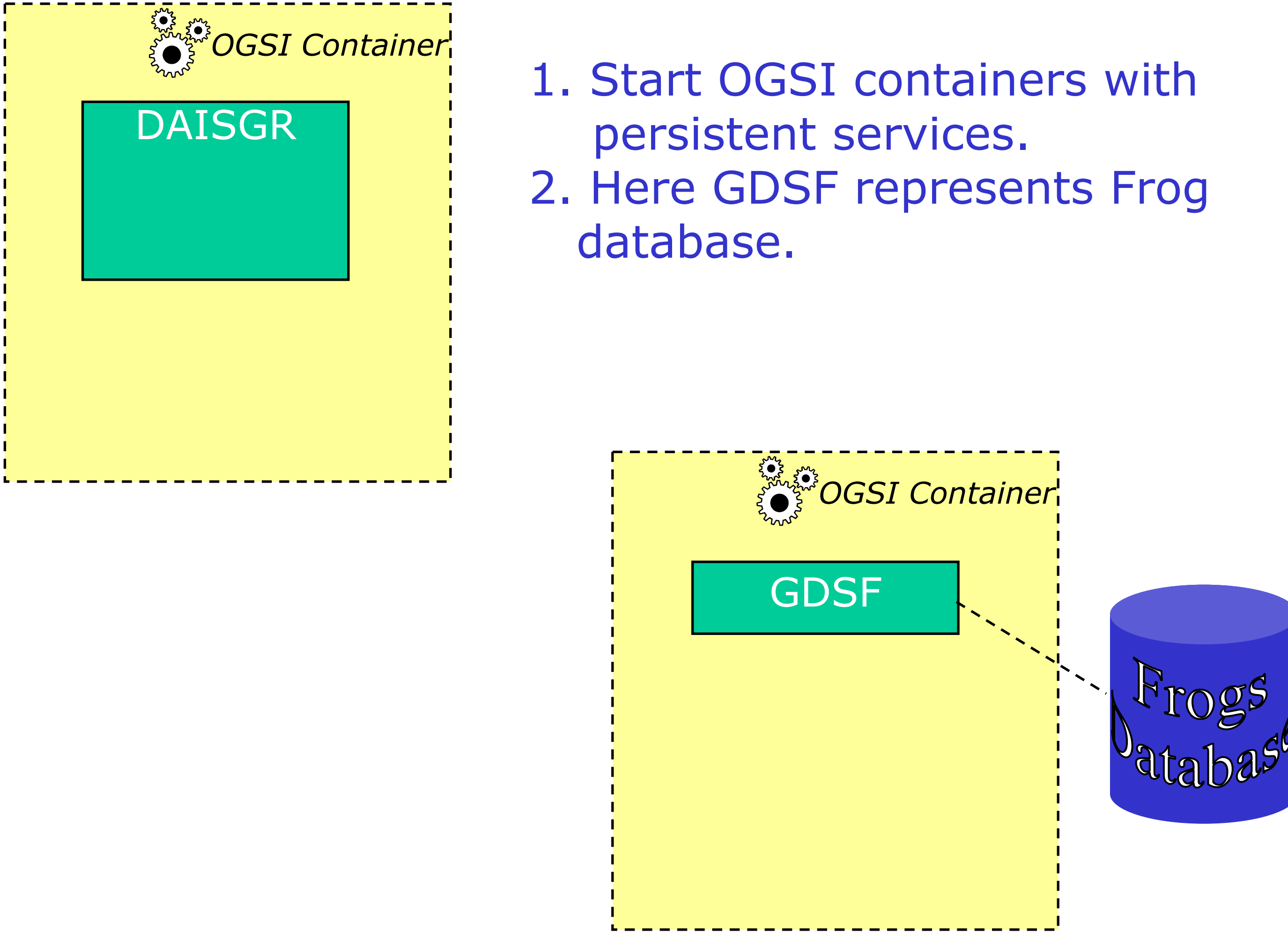
```
// Create a query
SQLQuery query = new SQLQuery(SQLQueryString);
ActivityRequest request = new ActivityRequest();
request.addActivity(query);

// Perform the query
Response response = gds.perform(request);

// Display the result
ResultSet rs = query.getResultSet();
displayResultSet(rs, 1);
```



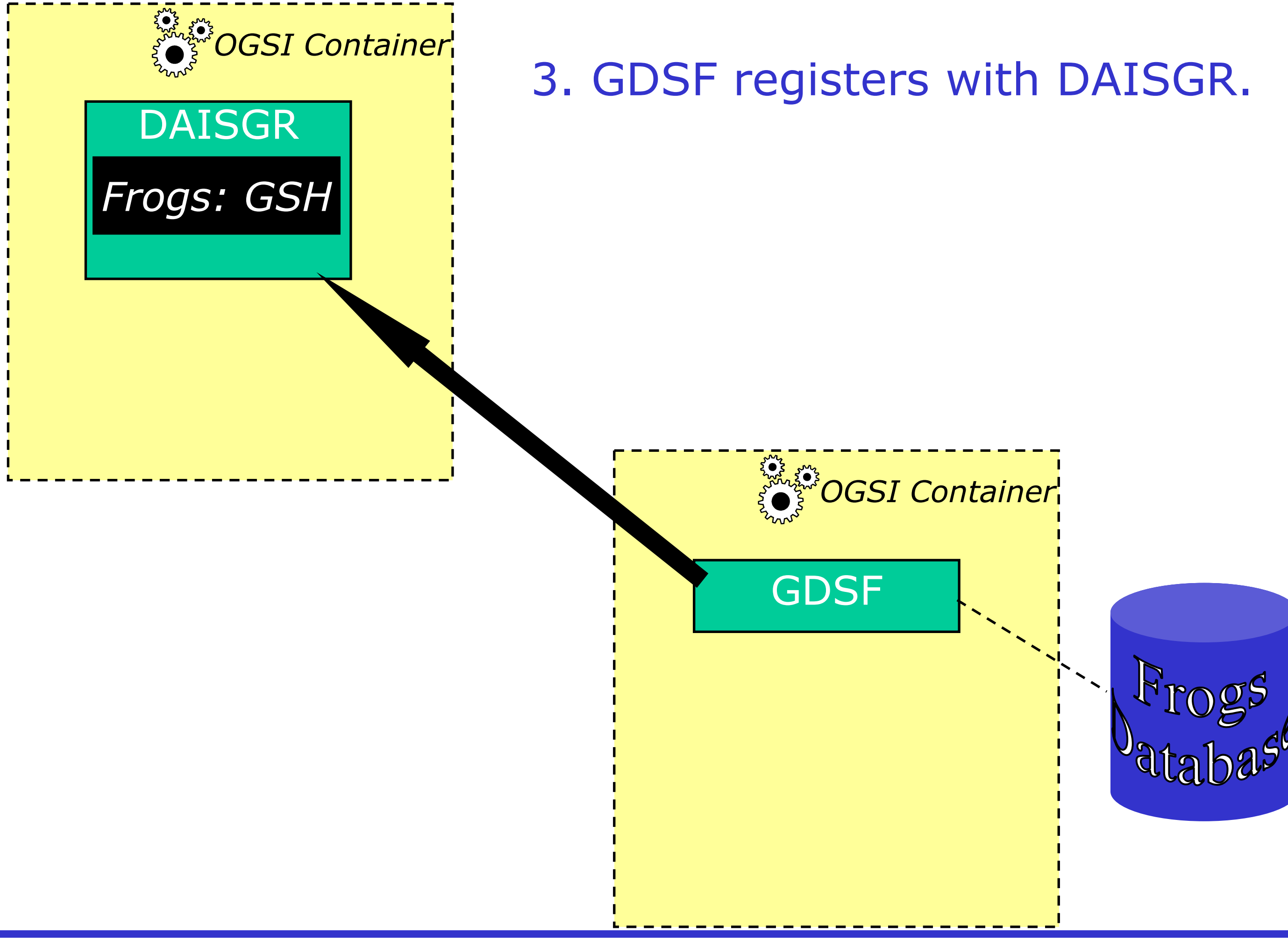
```
Completed Request
RegistryRequestEvent (to: http://amber.epcc.ed.ac.uk:8888/ogsa/ser
Completed Request
CreateServiceRequest (service name: HGU)
Completed Request
PerformDocumentRequest (service name: HGU)
Completed Request
RegistryRequestEvent (to: http://amber.epcc.ed.ac.uk:8888/ogsa/ser
```

Interaction Model: Start up

1. Start OGSI containers with persistent services.
2. Here GDSF represents Frog database.

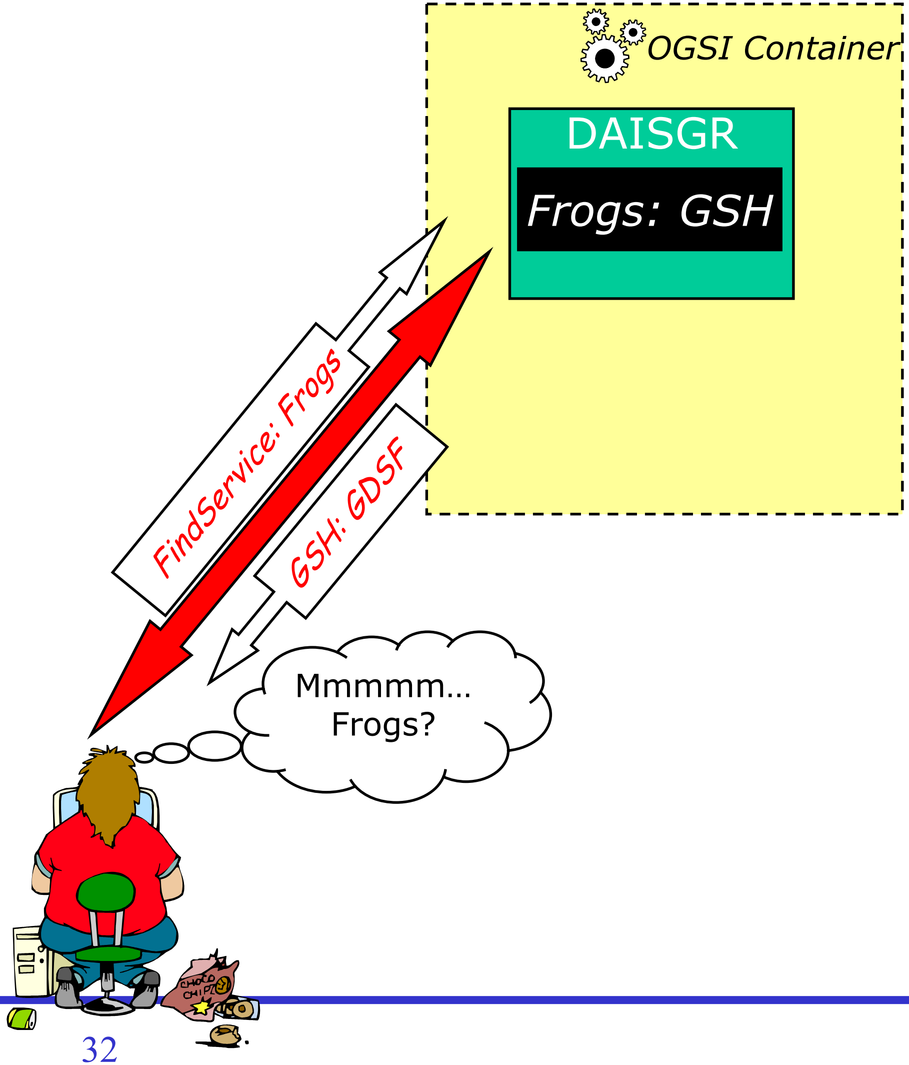


Interaction Model: Registration

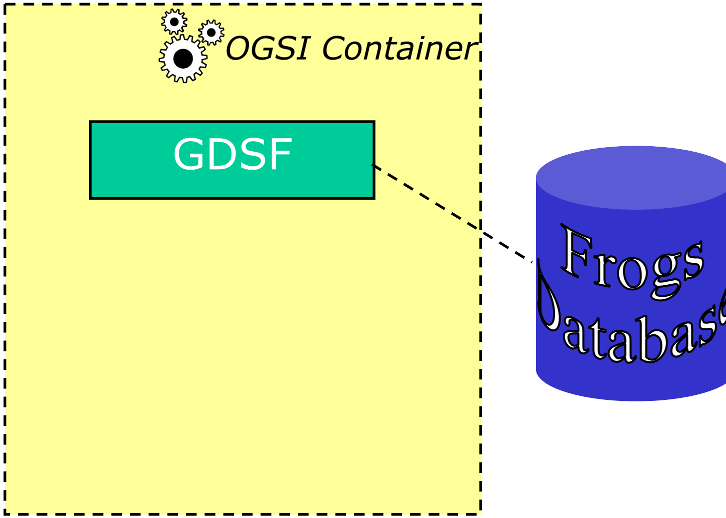


3. GDSF registers with DAISGR.

Interaction Model: Discovery

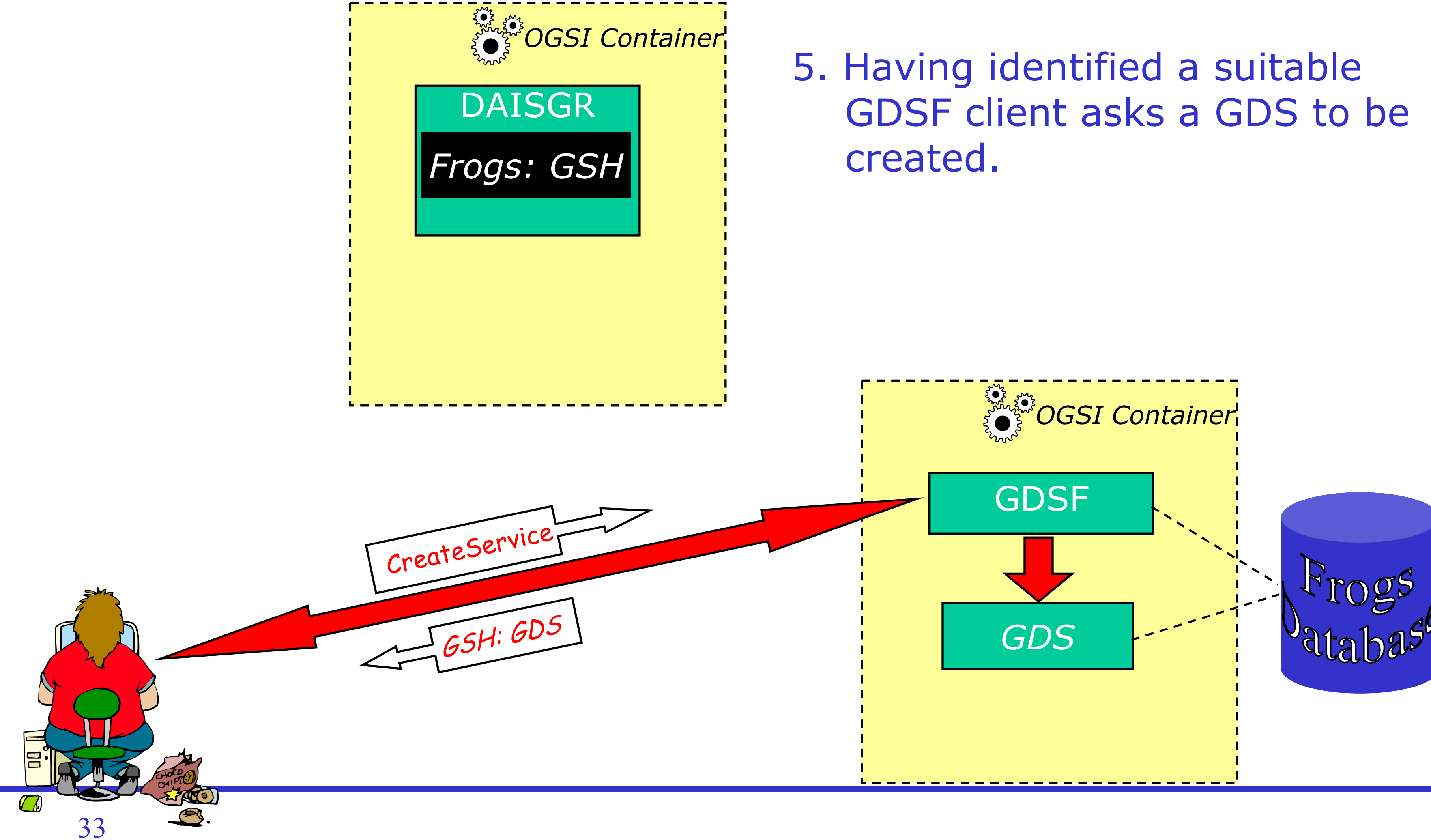


4. Client wants to know about frogs. Can:
   (i) Query the GDSF directly if known or
   (ii) Identify suitable GDSF through DAISGR.

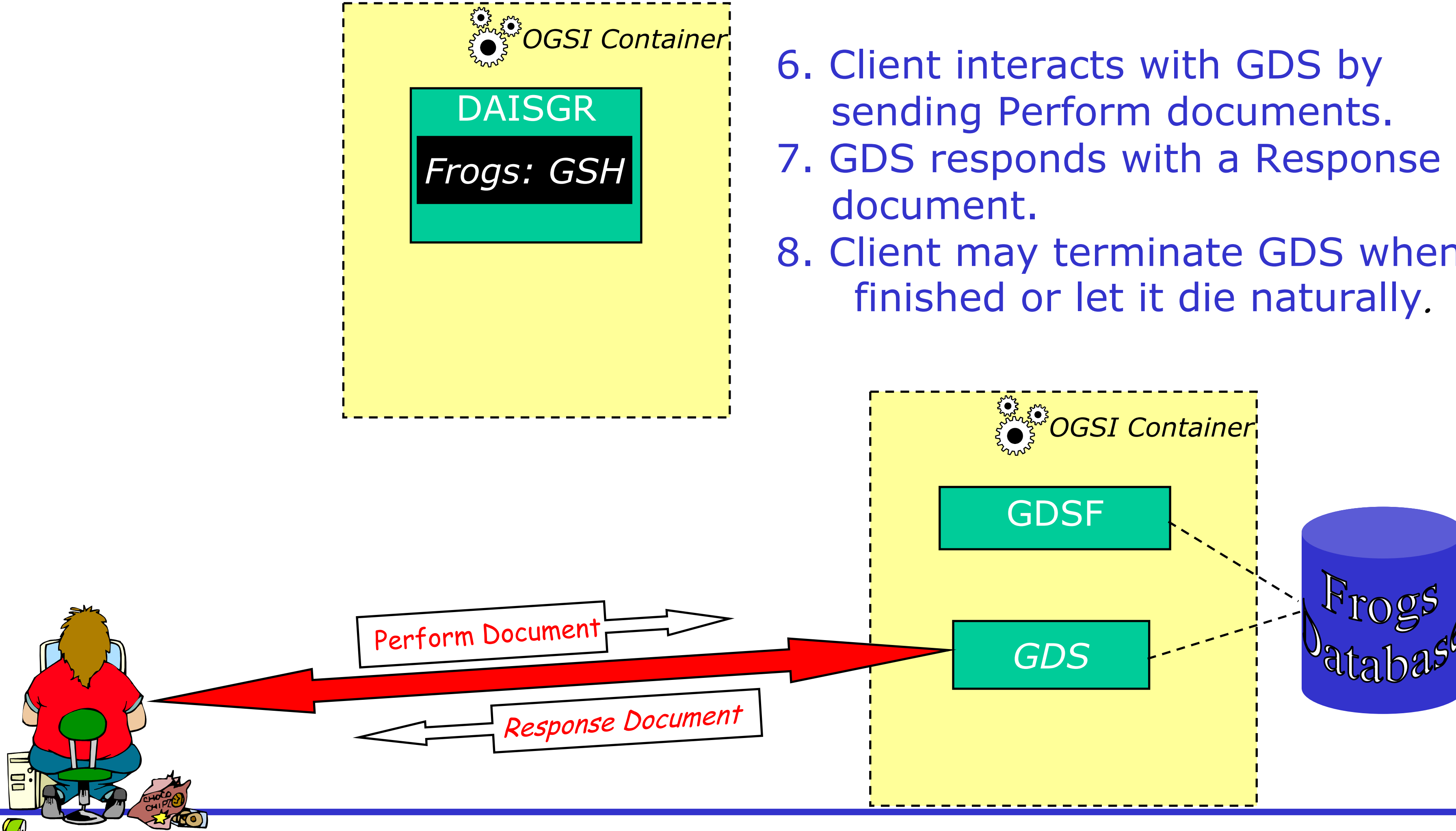


Interaction Model: Service Creation

5. Having identified a suitable GDSF client asks a GDS to be created.



Interaction Model: Perform

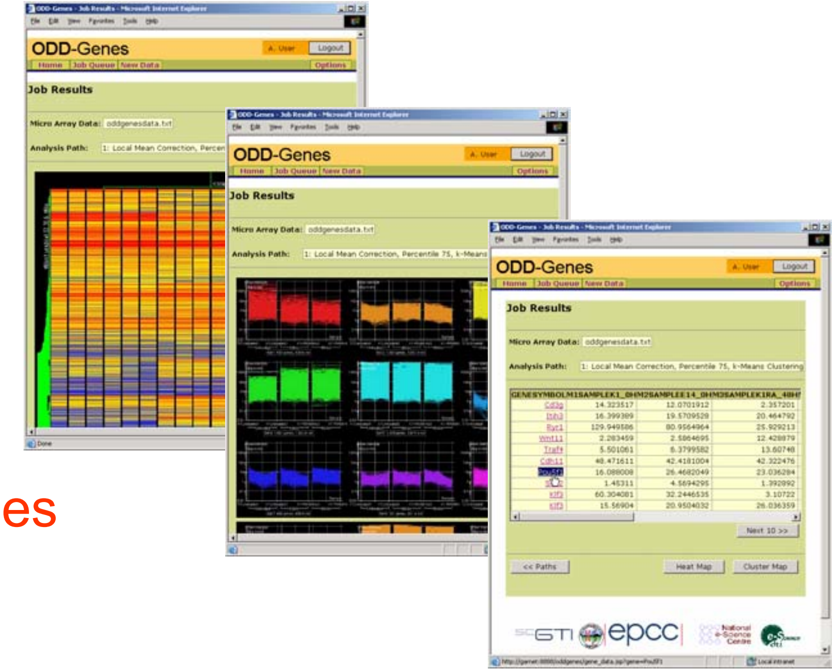


6. Client interacts with GDS by sending Perform documents.
7. GDS responds with a Response document.
8. Client may terminate GDS when finished or let it die naturally.

Interaction Model: Sum up

- Only describe an access use case
  - Client not concerned with connection mechanism
  - Similar framework could accommodate service-service interactions
- Discovery aspect is important
  - Probably requires a human
  - Needs adequate definition of metadata
    - Definitions of ontologies and vocabularies - not something that OGSA-DAI is doing …

- Data Analysis for genetics
  - Sites:
    - GTI (microarray data)
    - HGU (genex data)
    - EPCC (compute server)
  - Software:
    - OGSA-DAI (Data)
    - TOG (Computation)
    - Globus Toolkit 2 and 3
  - http://www.epcc.ed.ac.uk/oddgenes



Job Results

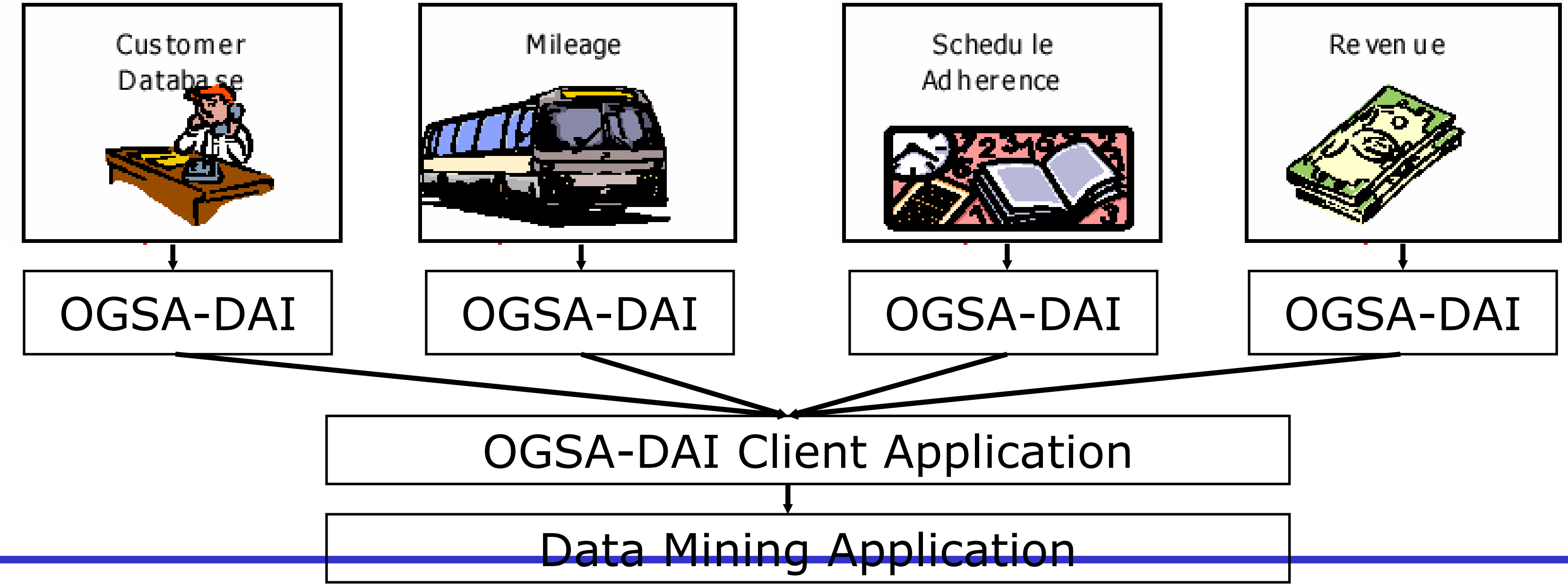
Micro Array Data: oddgenesdata.txt

Analysis Path: 1: Local Mean Correction, Percentile 75, k-Means Clustering

GENESYMBOL	M1SAMPLEK1_0H	M2SAMPLEE14_0H	M3SAMPLEK1RA_48H
Cd3g	14.323517	12.0701912	2.357201
Itih3	16.399389	19.5709529	20.464792
Ryr1	129.949586	80.9564964	25.929213
Wnt11	2.283459	2.5864695	12.428879
Traf4	5.501061	6.3799582	13.60748
Cdh11	48.471611	42.4181004	42.322476
Pou5f1	16.088008	26.4682049	23.036284
[illegible]	1.45311	4.5694295	1.392892
Klf3	60.304081	32.2446535	3.10722
Klf3	15.56904	20.9504032	26.036359

FirstDIG

- Data mining with the First Transport Group, UK
  - Example: “When buses are more than 10 minutes late there is an 82% chance that revenue drops by at least 10%”
  - http://www.epcc.ed.ac.uk/firstdig



Other projects

- MCS on OGSA-DAI
- BioGrid
- OpenSkyQuery

- More projects using OGSA-DAI:
  - http://www.ogsadai.org.uk/projects/