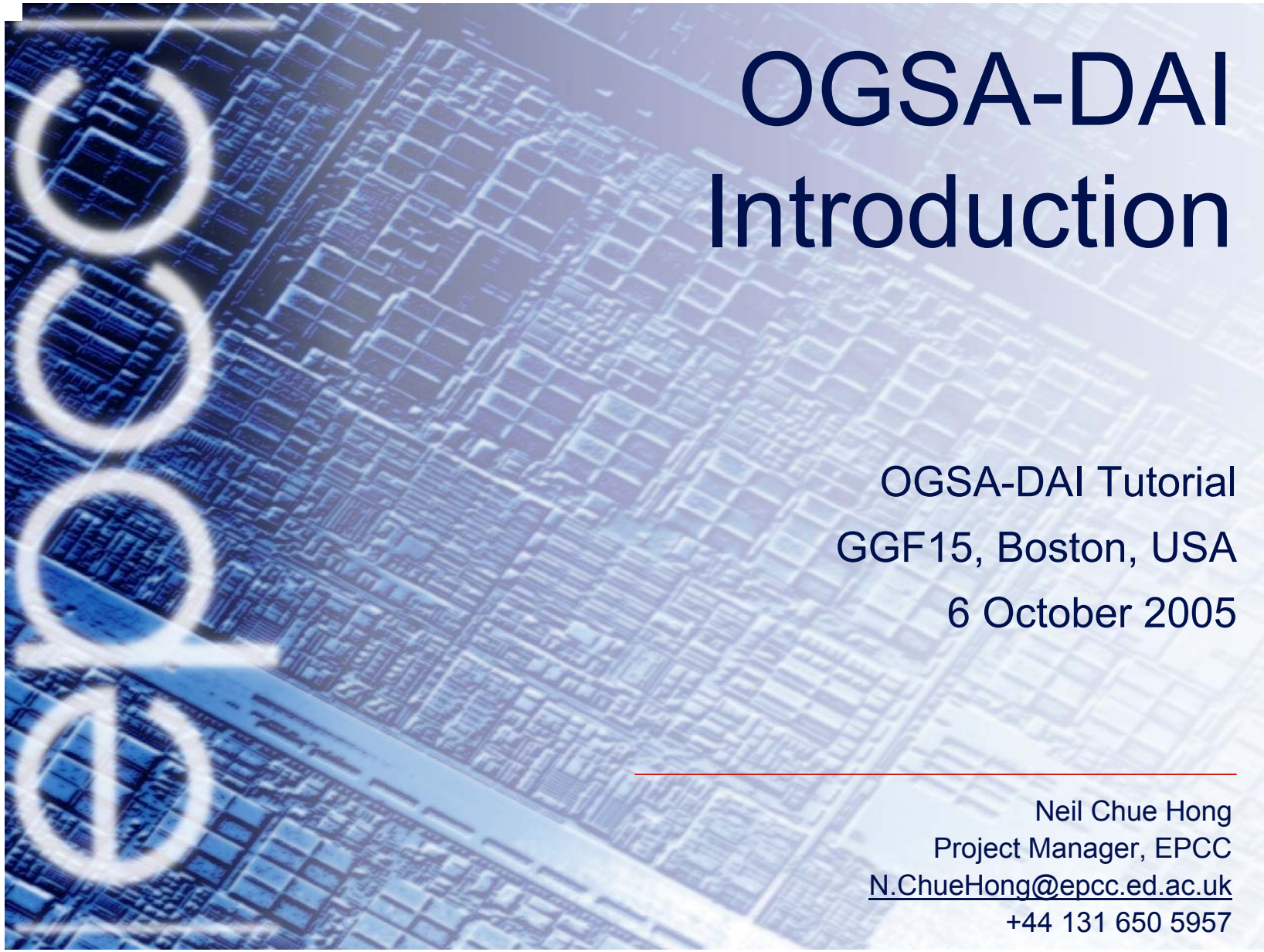


OGSA-DAI Introduction

OGSA-DAI Tutorial

GGF15, Boston, USA

6 October 2005

Neil Chue Hong
Project Manager, EPCC
N.ChueHong@epcc.ed.ac.uk
+44 131 650 5957

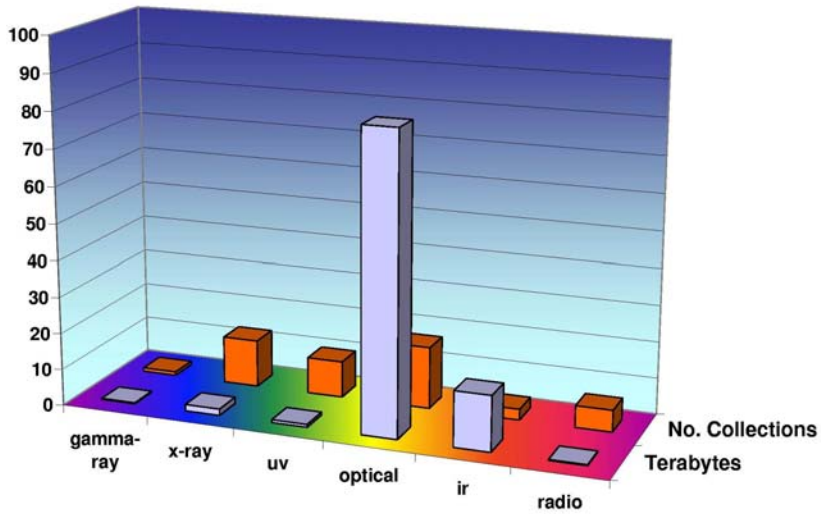
Timetable

- 11:00 Welcome / Introduction
- 11:30 Overview / high level architecture / Roadmap
- 12:30 Lunch
- 14:00 OGSA-DAI low level architecture / Configuring Data Resources
- 15:00 OGSA-DAI Client Toolkit
- 15:15 Wrap-up / feedback / user group & support
- 15:30 Close

Motivation

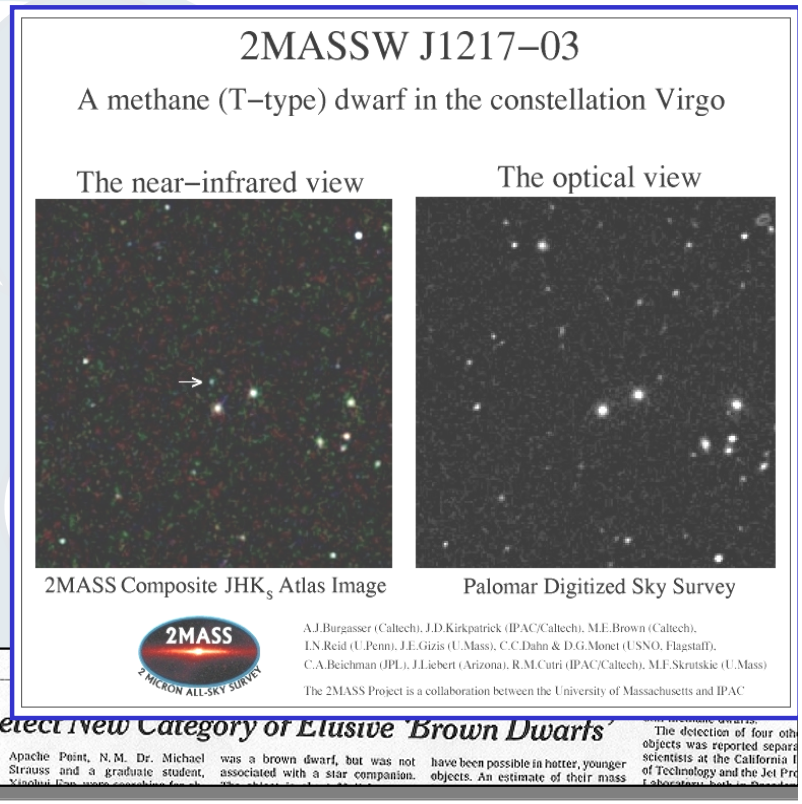
- Entering an age of data
  - Data Explosion
    - CERN: LHC will generate 1GB/s = 10PB/y
    - VLBA (NRAO) generates 1GB/s today
    - Pixar generate 100 TB/Movie
  - Storage getting cheaper
- Data stored in many different ways
  - Data resources
    - Relational databases
    - XML databases
    - 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

Composing Observations in Astronomy



No. & sizes of data sets as of mid-2002, grouped by wavelength

- 12 waveband coverage of large areas of the sky
- Total about 200 TB data
- Doubling every 12 months
- Largest catalogues near 1B objects



Astronomers Detect New Category of Elusive 'Brown Dwarfs'

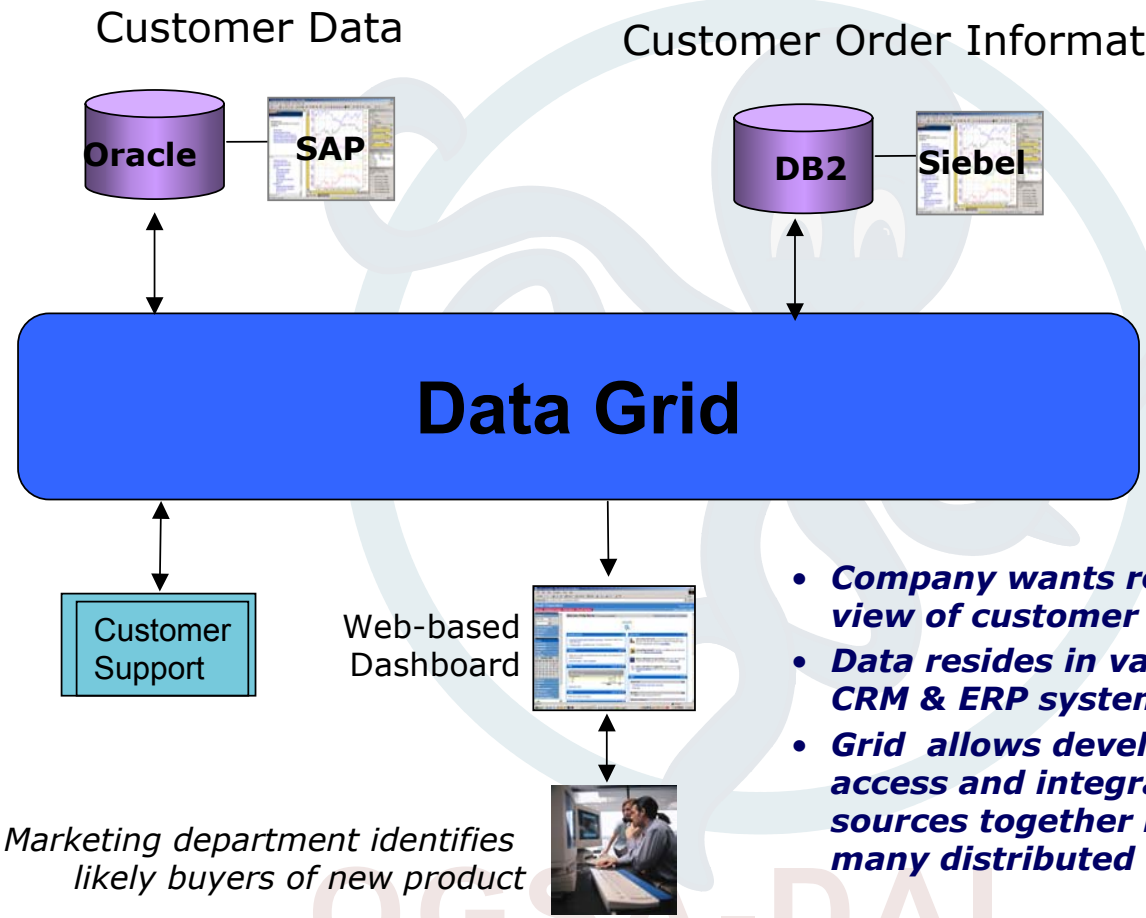
By JOHN NOBLE WILFORD

CHICAGO, May 31 — Ambitious Apache Point, N. M. Dr. Michael Strauss and a graduate student, was a brown dwarf, but was not associated with a star companion. have been possible in hotter, younger objects. An estimate of their mass

The detection of four oth objects was reported separa scientists at the California I of Technology and the Jet Pro

Data and images courtesy Alex Szalay, John Hopkins

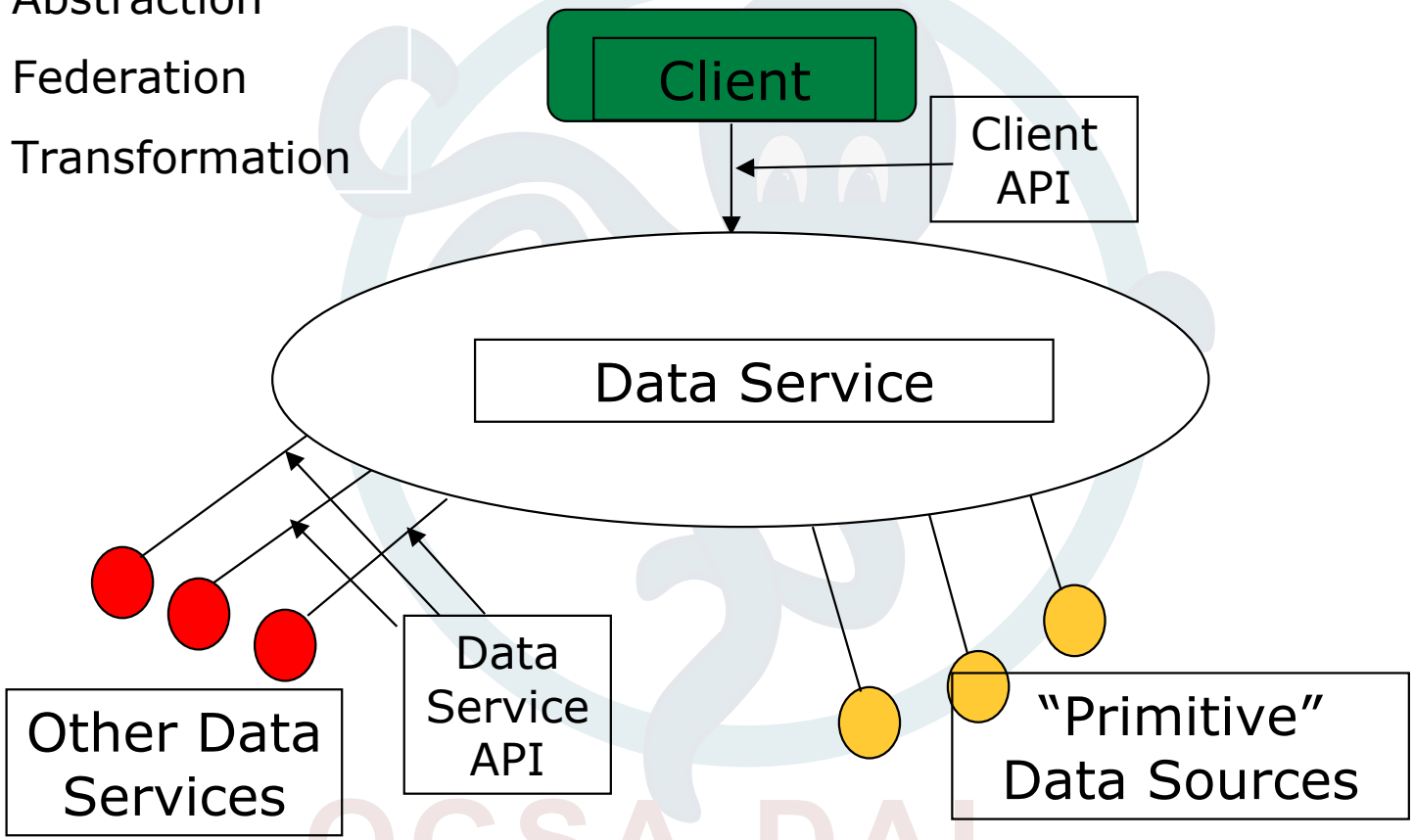
Business Intelligence and Customer Data



- ***Company wants real-time integrated view of customer buying behavior***
- ***Data resides in various distributed CRM & ERP systems***
- ***Grid allows developers and apps to access and integrate customer data sources together in real time--across many distributed databases***

Data Virtualisation

- Abstraction
- Federation
- Transformation



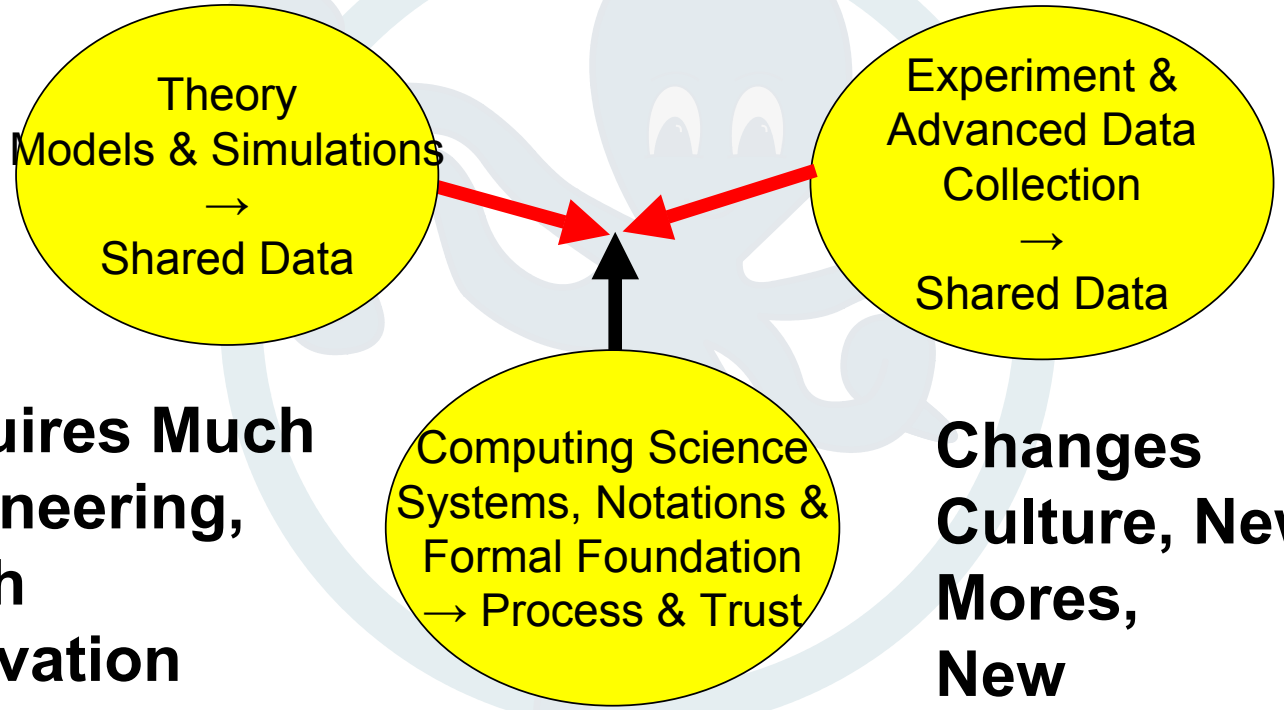
Grand Challenges

Neil Chue Hong
Project Manager, EPCC
N.ChueHong@epcc.ed.ac.uk
+44 131 650 5957

Share and share alike!

- Many challenges:
  - Scalability, performance, heterogeneity, ownership, economics
  - Common schema, data description and semantics, data formats, process and procedure, provenance
- Can be solved only through collaboration and the sharing of:
  - Ideas
  - Efforts
  - Resources
- Perhaps most importantly: sharing of data
  - Beware of data huggers!
- Emerging Open Grid Infrastructures will
  - Allow global collaboration
  - Change the way that we can work

Multi-national, Multi-discipline, Computer-enabled Consortia, Cultures & Societies

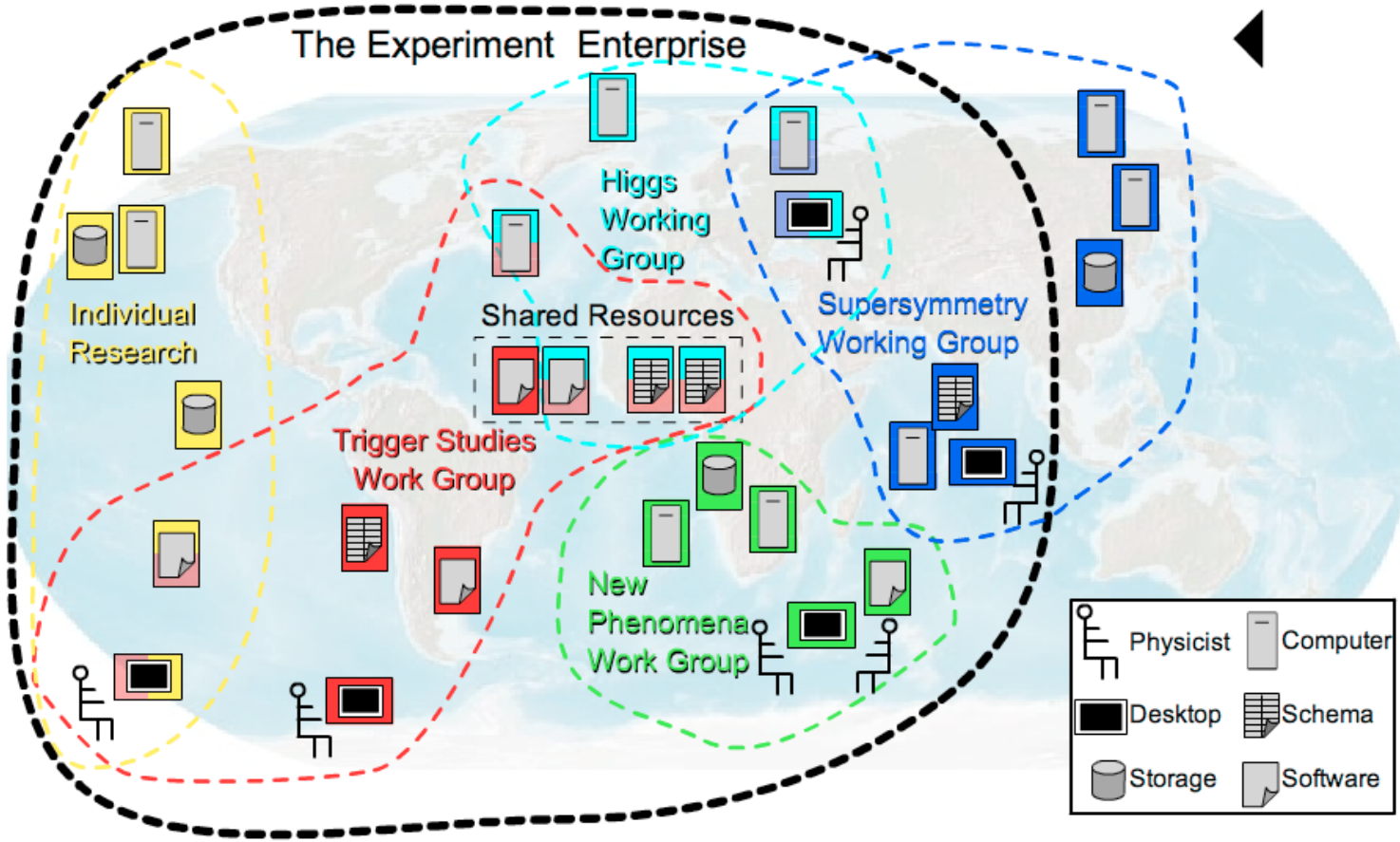


Requires Much Engineering, Much Innovation

Changes Culture, New Mores, New Behaviours

New Opportunities, New Results, New Rewards

Emergence of Virtual Organisations



Data Requirements

- What do we need for effective sharing of data?
  - Structured, organised, annotated & curated data
  - Computable data models
  - Visualisation of data
  - Data provenance
  - Shared distributed systems
  - Networked workplaces, instruments, data sources
  - Metadata, ontologies, standards
  - Authentication, authorisation, accounting, policies

Data Services: motives

- Key to Integration of Scientific Methods
  - Publication and sharing of results
    - Primary data from observation, simulation & experiment
    - Encourages novel uses
    - Allows validation of methods and derivatives
    - Enables discovery by combining data collected independently
- Key to Large-scale Collaboration
  - Economies: data production, publication & management
    - Sharing cost of storage, management and curation
    - Many researchers contributing increments of data
    - Pooling annotation leads to rapid incremental publication
  - Accommodates global distribution
    - Data & code travel faster and more cheaply
  - Accommodates temporal distribution
    - Researchers assemble data
    - Later (other) researchers access data

Data Services: challenges

- Scale
  - Many sites, large collections, many uses
- Longevity
  - Research requirements outlive technical decisions
- Diversity
  - No “one size fits all” solutions will work
    - Primary Data, Data Products, Meta Data, Administrative data, …
- Many Data Resources
  - Independently owned & managed
    - No common goals
    - No common design
    - Work hard for agreements on foundation types and ontologies
    - Autonomous decisions change data, structure, policy, …
  - Geographically distributed
- and I haven’t even mentioned security yet!

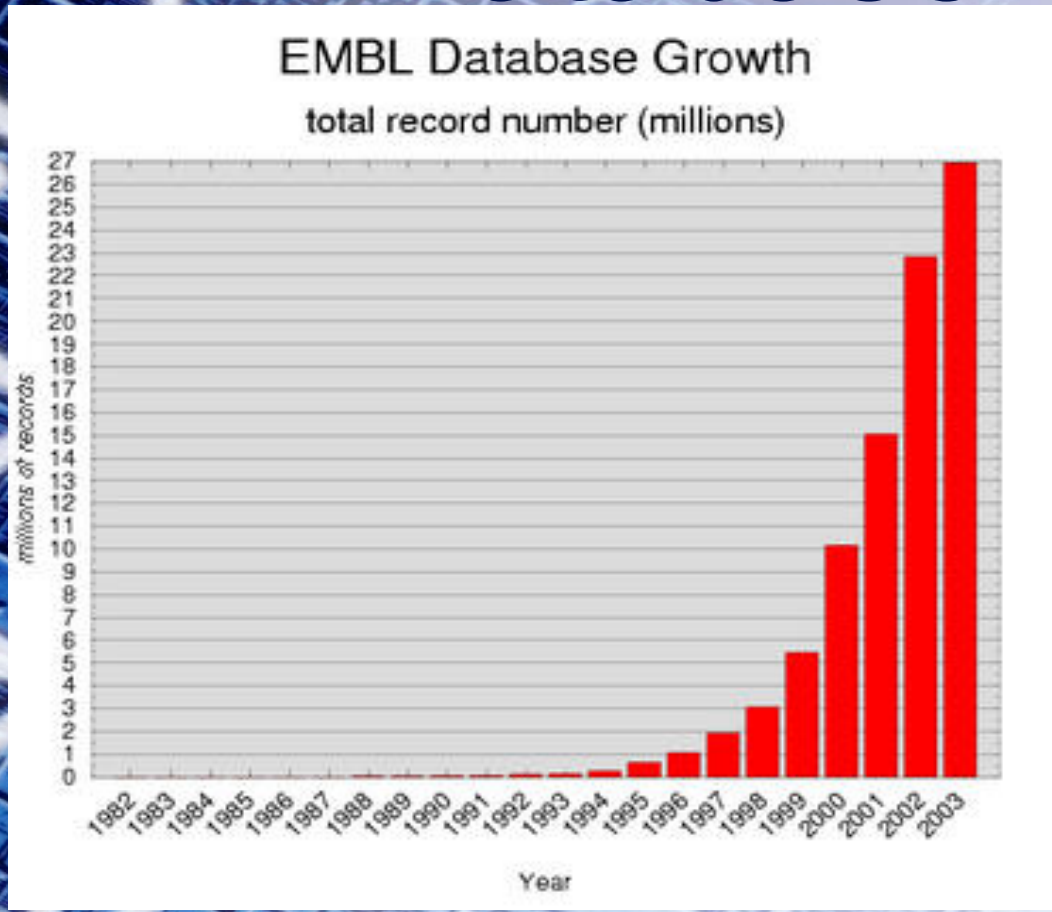
The Discovery Process

- Choosing data sources
  - How do you find them?
  - How do they describe and advertise them?
  - Is the equivalent of Google possible?
- Obtaining access to that data
  - Overcoming administrative barriers
  - Overcoming technical barriers
- Understanding that data and extracting from multiple sources
  - The parts you care about for your research
- Combing them using sophisticated models
  - The picture of reality in your head
- Analysis on scales required by statistics
  - Coupling data access with computation
- Repeated Processes
  - Examining variations, covering a set of candidates
  - Monitoring the emerging details

Small problems

- Not just “Grand Challenges”!
  - Also the small problems
- For instance:
  - What happens to data when a researcher leaves a team?
  - How can a research leader point to “popular” data when a new researcher joins?
  - How can you manage your data when you start to run out of local storage space?
  - How do I get my data from one format/database to another?
  - How do I combine *my* data with *your* data?
- You need to manage your data

Database Growth



Neil Chue Hong
Project Manager, EPCC
N.ChueHong@epcc.ed.ac.uk
+44 131 650 5957

Terabyte → Petabyte

	Terabyte	Petabyte
RAM time to move	15 minutes	2 months
1GB WAN move time	10 hours ($1000)	14 months ($1 million)
Disk cost	7 disks = $5000 (SCSI)	6800 Disks + 490 units + 32 racks = $7 million
Disk power	100 Watts	100 Kilowatts
Disk weight	5.6 Kg	33 Tonnes
Disk footprint	Inside machine	60 m2

Approximately Correct in May 2003 *Distributed Computing Economics*
Jim Gray, Microsoft Research, MSR-TR-2003-24

Mohammed & Mountains

- Petabytes of Data cannot be moved
  - It stays where it is produced or curated
    - Hospitals, observatories, European Bioinformatics Institute
  - A few caches and a small proportion cached
- Distributed collaborating communities
  - Expertise in curation, simulation & analysis
- Diverse data collections
  - Discovery depends on insights
  - Unpredictable or unexpected use of data

Move computation to the data

- Assumption: code size << data size
  - Minimise data transport
- Provision combined storage & compute resources
- Develop the database philosophy for this?
  - Enhanced stored procedures
  - Pre-query analysis for more concise queries
  - Mobile code sandbox
  - Robustness
- Develop the storage architecture for this?
  - Compute closer to disk?
  - System on a Chip using free space in the on-disk controller
- Develop experiment, sensor & simulation architectures
  - That take code to select and digest data as an output control
- Data Cutter a step in this direction
  - Sub-setting and aggregation of datasets using filters executed close to data
  - http://www.cs.umd.edu/projects/hpsl/ResearchAreas/DataCutter.htm

Meta-data: describing data

- Choosing data sources
  - How do you find them?
  - How are they described and advertised?
  - Is the equivalent of Google possible?
- Meta-data is required describing:
  - Structure of data
  - Types of data
  - Operations supported/available
  - Access requirements
  - Quality of service?
- No established standards for heterogeneous data sources

Cultural Challenges

- Changing the way we work?
- Publication and sharing of results
  - Increased volume and diversity = increased opportunity?
  - Allows independent validation of methods and derivatives
  - Responsibility, ownership, credit, citation
- Many distributed data resources
  - Data collected from observation, simulation & experiment
  - Independently owned & managed
    - No common goals or design
    - Work hard for agreements on foundation types and ontologies
    - Autonomous decisions change data, structure, policy, etc
  - Requires negotiations and patience!
- Diversity
  - No “one size fits all” solutions will work

Economic Challenges

- Data production, publication & management
  - Many researchers contributing increments of data
  - Who pays for storage, transport, management and curation?
- Data longevity
  - Research requirements may outlive technical decisions
  - Data does not preserve itself indefinitely!
- Costs must be shared somehow…

Security Challenges: Medical Imaging Data

- Diagnosing based on sensitive patient data
  - Users: a (group of) doctor(s)
  - Retrieve an image, run algorithm, examine result and write diagnosis, maybe re-run another algorithm.
- Secure Data Retrieval
  - Patient data is sensitive, needs to be stored anonymously at all times
  - Site admins are not trustworthy – strip or encrypt patient data from image
  - Replication of data not always allowed
- High security needs
  - Strong authorization
  - Fine-grained access control mechanisms
  - Leaking patient information results in prosecution.

- Obtaining access to that data
  - Overcoming administrative barriers
  - Overcoming technical barriers
- Understanding that data
  - The parts you care about for your research
- Combing them using sophisticated models
  - The picture of reality in your head
- Analysis on scales required by statistics
  - Coupling data access with computation
- Repeated Processes
  - Examining variations, covering a set of candidates
  - Monitoring the emerging details
  - Coupling with scientific workflows

DAIS Working Group

Neil Chue Hong
Project Manager, EPCC
N.ChueHong@epcc.ed.ac.uk
+44 131 650 5957

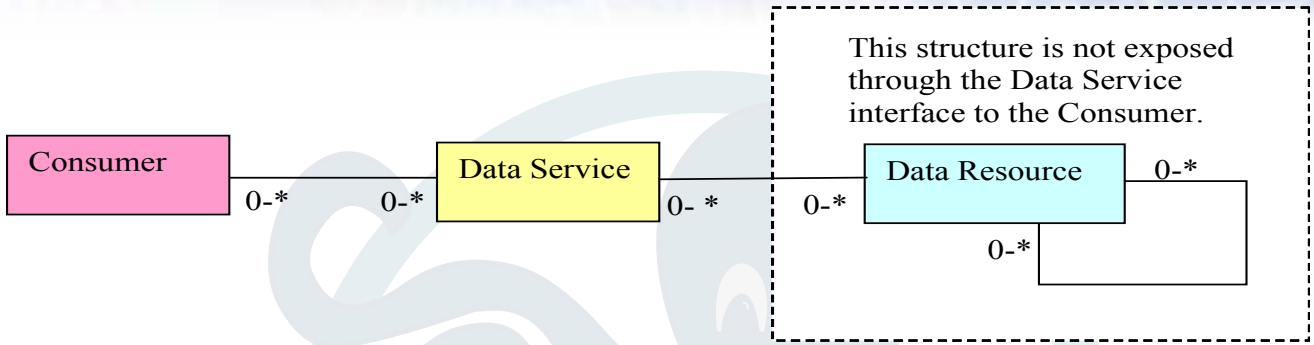
DAIS WG Goals

- Provide service-based access to structured data resources as part of OGSA architecture
- Specify a selection of interfaces tailored to various styles of data access starting with relational and XML
- Interact well with other GGF OGSA specs

DAIS WG Non-Goals

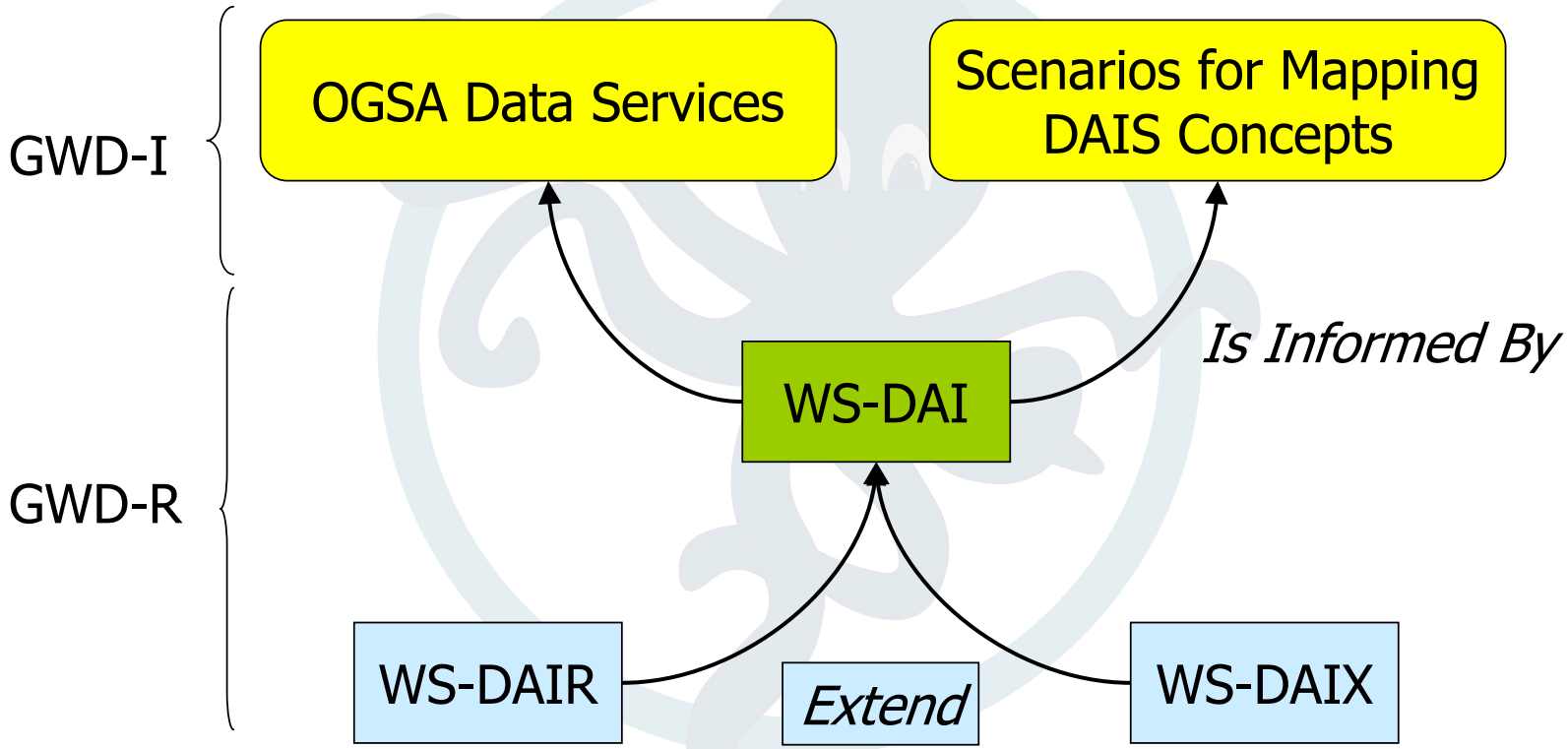
- No new common query language
- No schema integration or common data model
- No common namespace or naming scheme
- No data resource management
  - e.g. starting/stopping database managers
- No push based delivery
  - Information Dissemination WG?

DAIS View Of Data Services Model

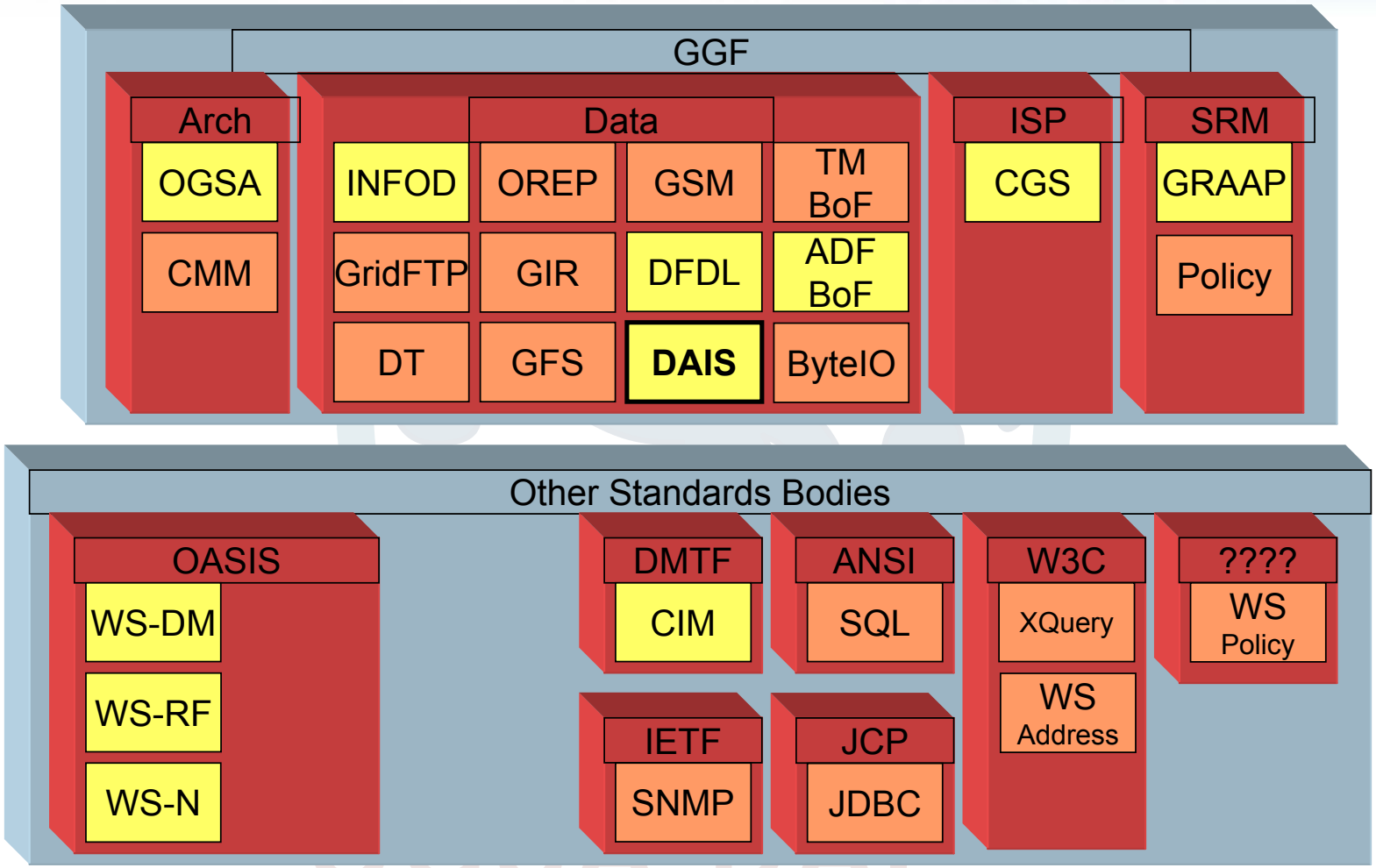


- A Data Service presents a Consumer with an interface to a Data Resource.
- A Data Resource can have arbitrary complexity, for example, a file on an NFS mounted file system or a federation of relational databases.
- A Consumer is not typically exposed to this complexity and operates within the bounds and semantics of the interface provided by the Data Service

DAIS Specification Landscape



DAIS and Other Standards/Specifications



The story so far

- Technology enables Grids and more data
- Distributed systems for sharing information
  - Essential, ubiquitous & challenging
  - Therefore share methods and technology as much as possible
- Collaboration is essential
  - Combining approaches
  - Combining skills
  - Sharing resources
- Structured Data is the language of Collaboration
  - Data Access & Integration a Ubiquitous Requirement
  - Primary data, metadata, administrative & system data
- Many hard technical challenges
  - Scale, heterogeneity, distribution, dynamic variation
  - Intimate combinations of data and computation with autonomous development of both

Overview

Neil Chue Hong
Project Manager, EPCC
N.ChueHong@epcc.ed.ac.uk
+44 131 650 5957

Goals for this presentation

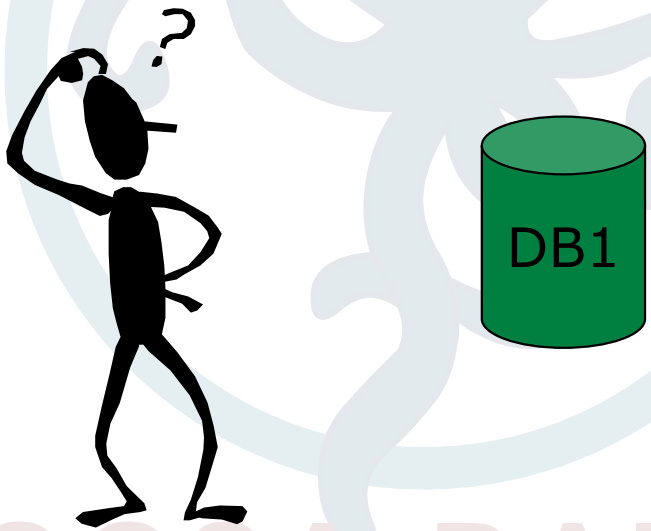
- Understand data access scenarios on the Grid
- Describe how the Grid influences data access and integration
- Describe an overview of the OGSA-DAI software

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

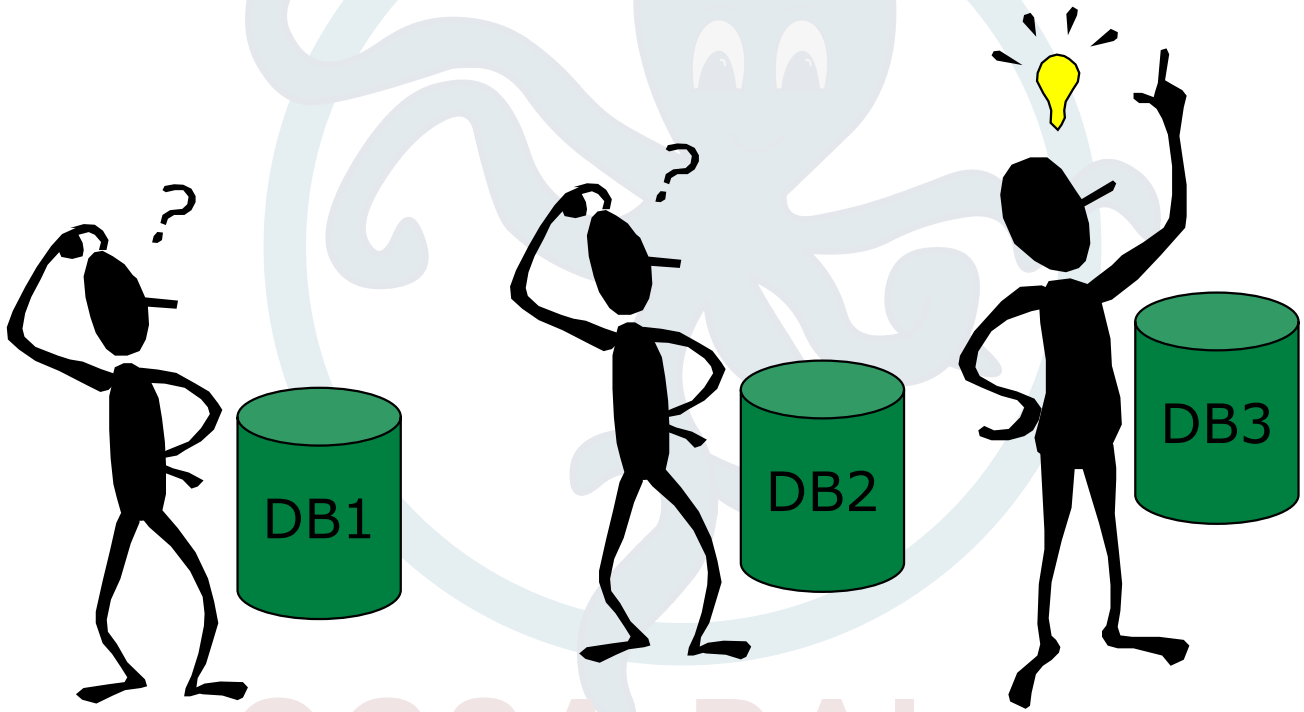
Why is the Grid necessary?

- 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

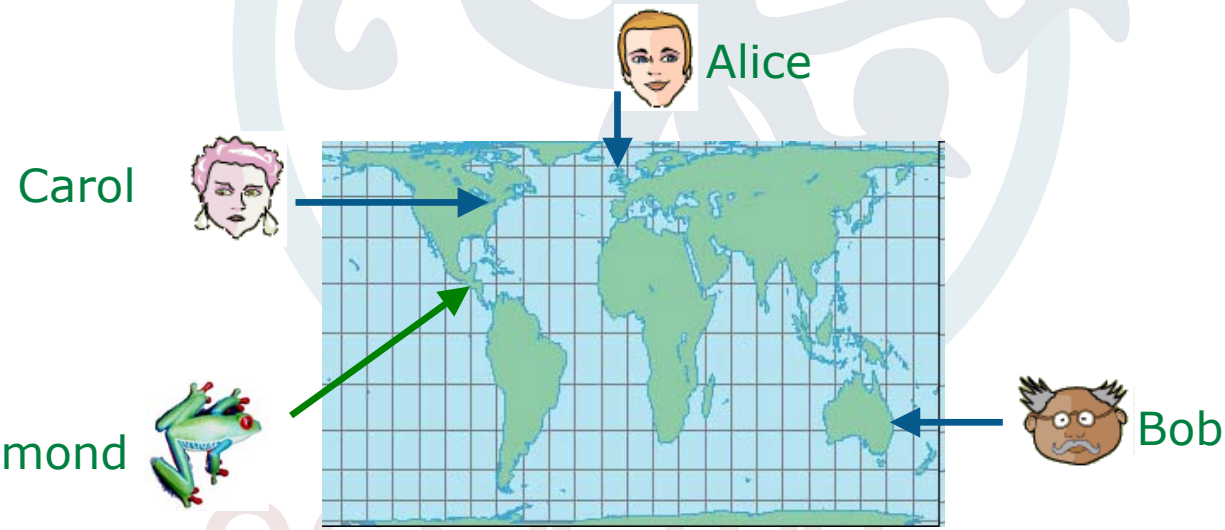
The story of Alice, Bob, Carol

and a frog called Desmond

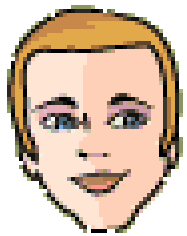
Thanks to Tom Sugden and Martin Westhead for the original idea

Once upon a time…

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

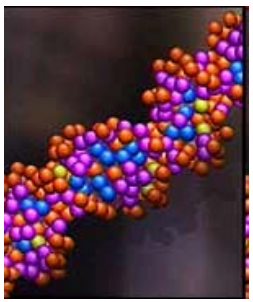


Use Case: Publishing



- 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

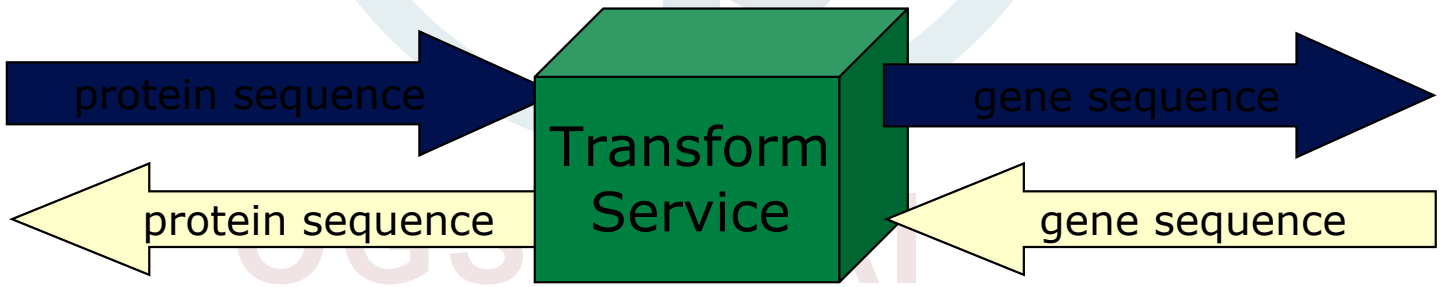
Use-case: Remote update



- 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

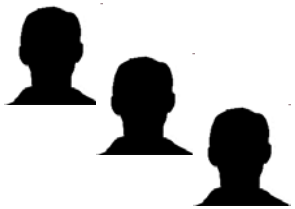
Use Case: Transformations

- 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



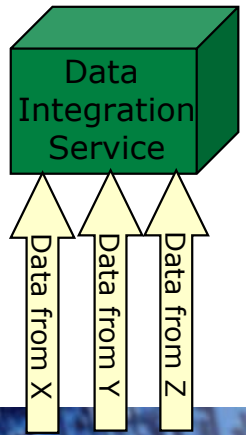
Use Case: Data Integration

- 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

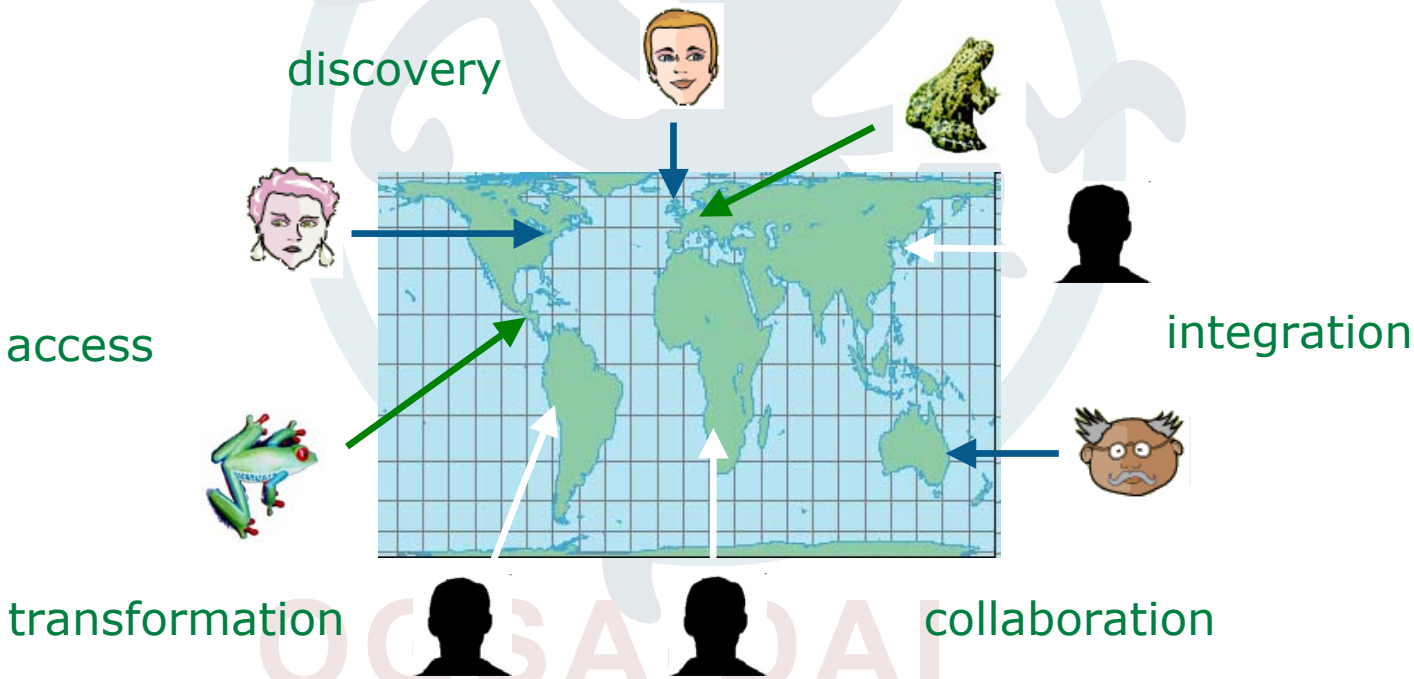
- Trying to save the nearly extinct variegated red-eyed tree frog
- Alice writes a service which exposes a integrated set of data as another virtual data resource
- Bob and Carol can use this resource as if it were a single data resource



- They find a way to save Desmond!

The End

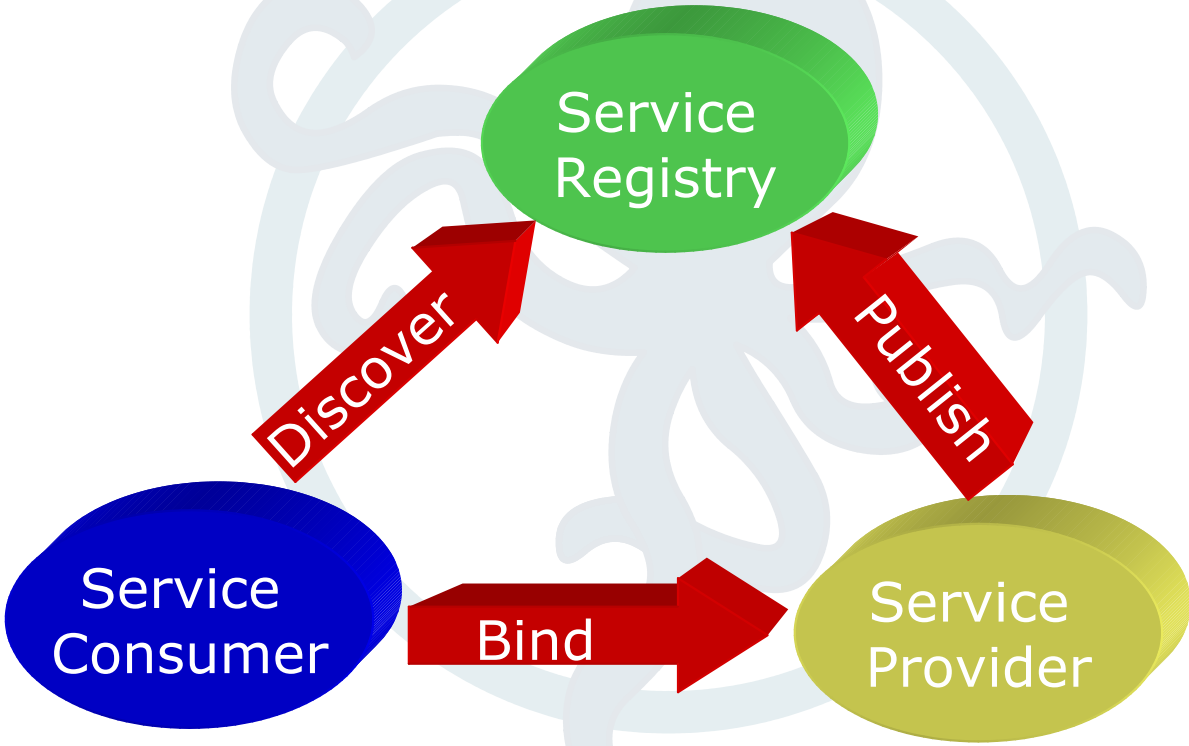
- Use data services to provide the middleware tools to grid-enable existing databases



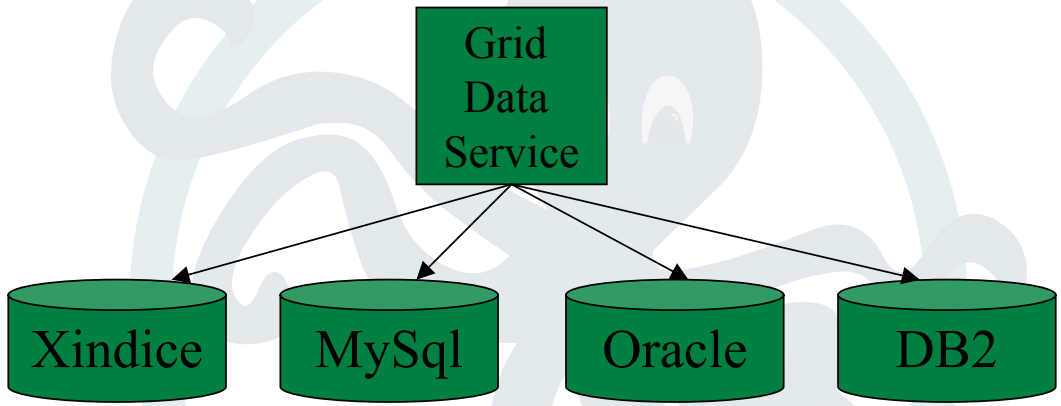
What is a data service?

- An interface to a stored collection of data
  - e.g. Google and Amazon
  - web services
- But the data could be:
  - replicated
  - shared
  - federated
  - virtual
  - incomplete
- Don't care about the underlying representation
  - do care about the information it represents

Web Service Architecture

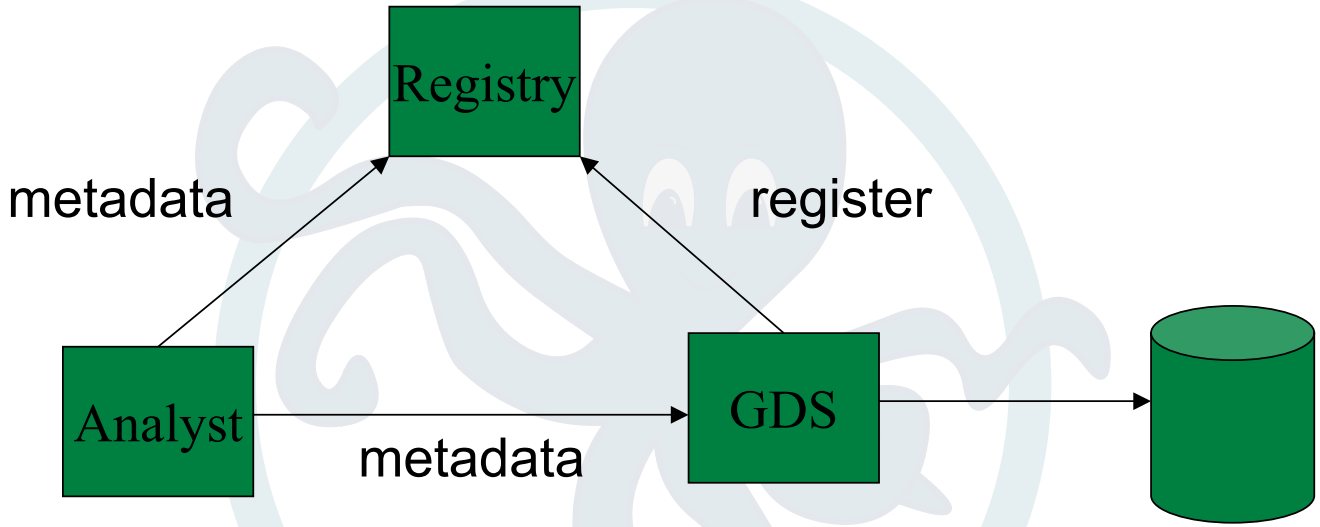


Heterogeneity



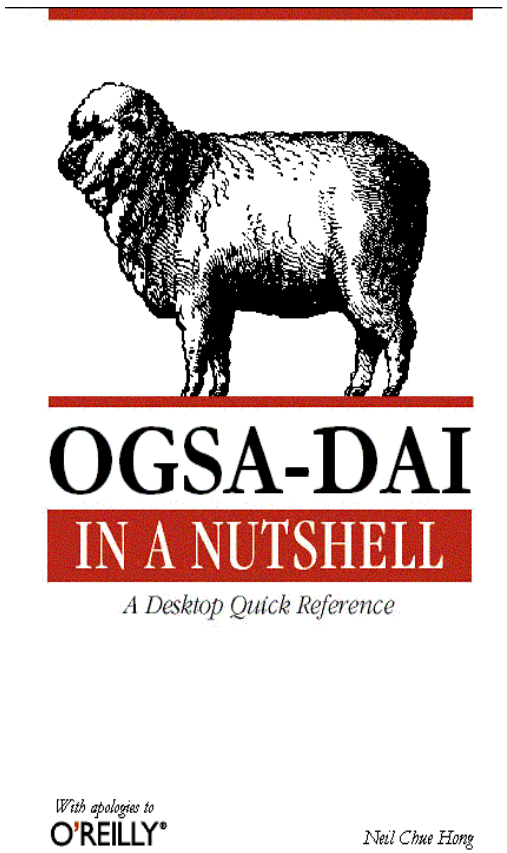
- Data source abstraction behind GDS instance
  - plug in “data resource implementations” for different data source technologies
  - does not mandate any particular query language or data format

Location



- Data resource publication through registry
- Data location hidden by factory
- Data resource meta data available through Resource Properties

OGSA-DAI In One Slide



- An *extensible framework* for data access and integration.
- Expose heterogeneous data resources to a grid through web services.
- Interact with data resources:
  - Queries and updates.
  - Data transformation / compression
  - Data delivery.
- Customise for your project using
  - Additional Activities
  - Client Toolkit APIs
  - Data Resource handlers
- A base for higher-level services
  - federation, mining, visualisation,…

Project Partners

Powered by ….

Funded 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

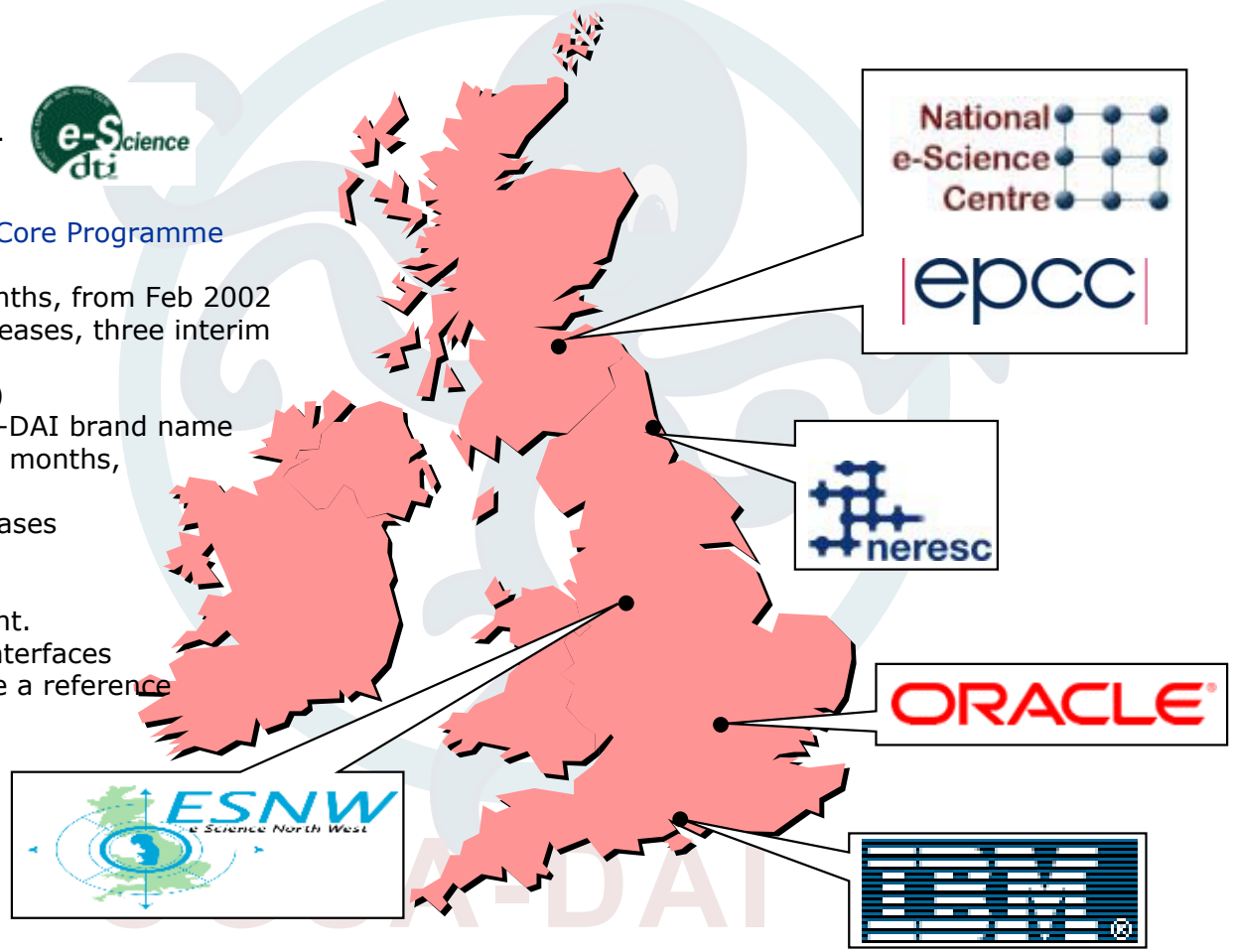
Four major releases

GGF DAIS WG

Strong involvement.

Standardise the interfaces

OGSA-DAI to be a reference implementation



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)
  - Started with Globus Toolkit 3 (GT3)
  - Moved to WS-RF(GT4) and WS-I+(OMII) versions

Project Status

- Current release 6.0
  - GT3.2, GT4.0, OMII_2, Axis 1.2 RC3
  - 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
- Next release 7.0 drops GT3 support

Why OGSA-DAI?

- Why use OGSA-DAI over JDBC?
  - Can embed additional functionality at the service end
    - Transformations, compressions
    - Third party delivery
    - The extensible activity framework
  - Avoiding unnecessary data movement
  - Common interface to heterogeneous data resources
    - Relational, XML databases, and files
  - Usefulness of the Registry for service discovery
    - Dynamic service binding process
    - Provision of good meta-data is necessary
  - Language independence at the client end
    - Do not need to use Java
  - Platform independence
    - Do not have to worry about connection technology, drivers, etc

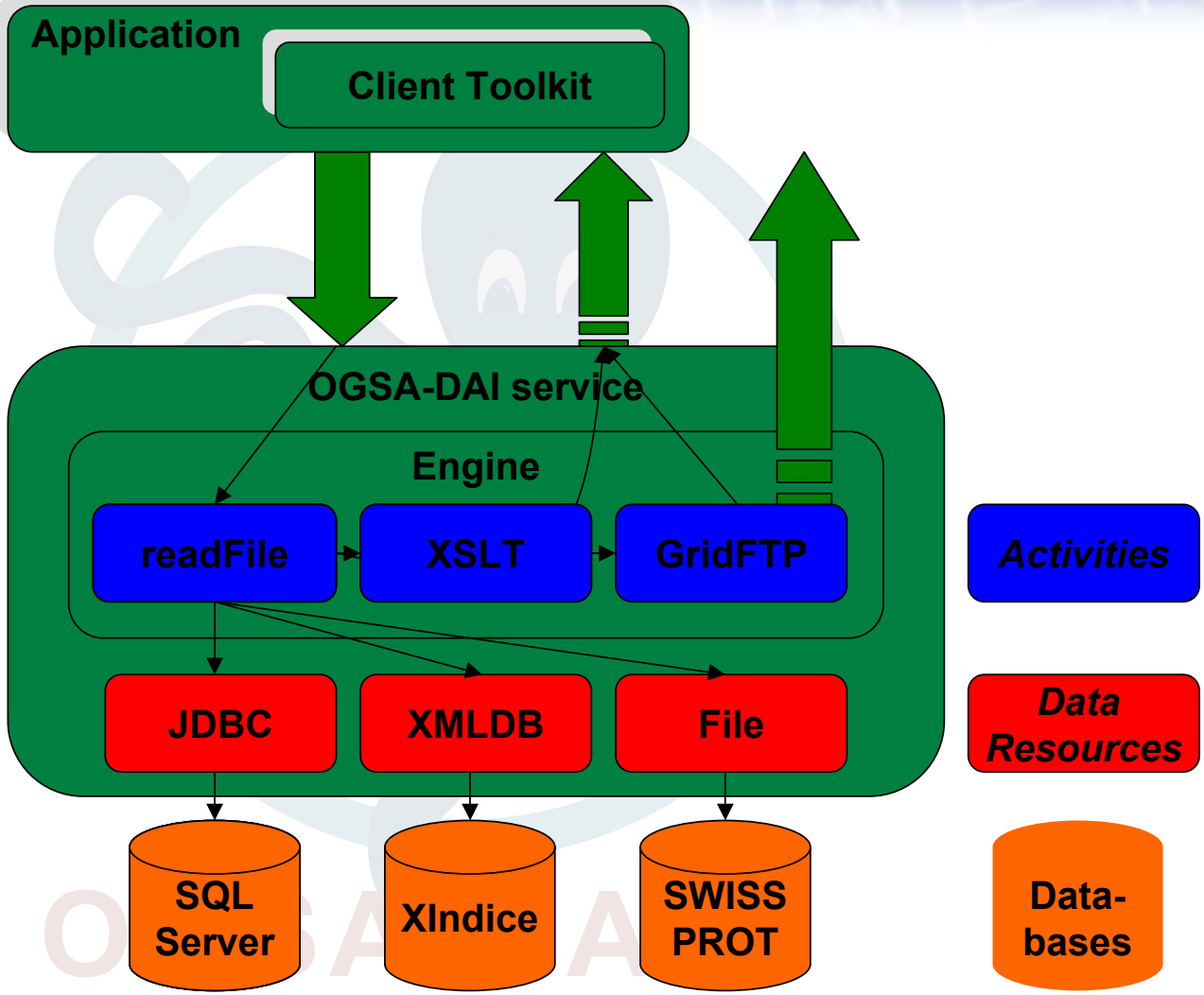
OGSA-DAI Design Principles – I

- Efficient client-server communication
  - Minimise where possible
  - One request specifies multiple operations
- No unnecessary data movement
  - Move computation to the data
  - Utilise third-party delivery
  - Apply transforms (e.g., compression)
- Build on existing standards
  - Fill-in gaps where necessary

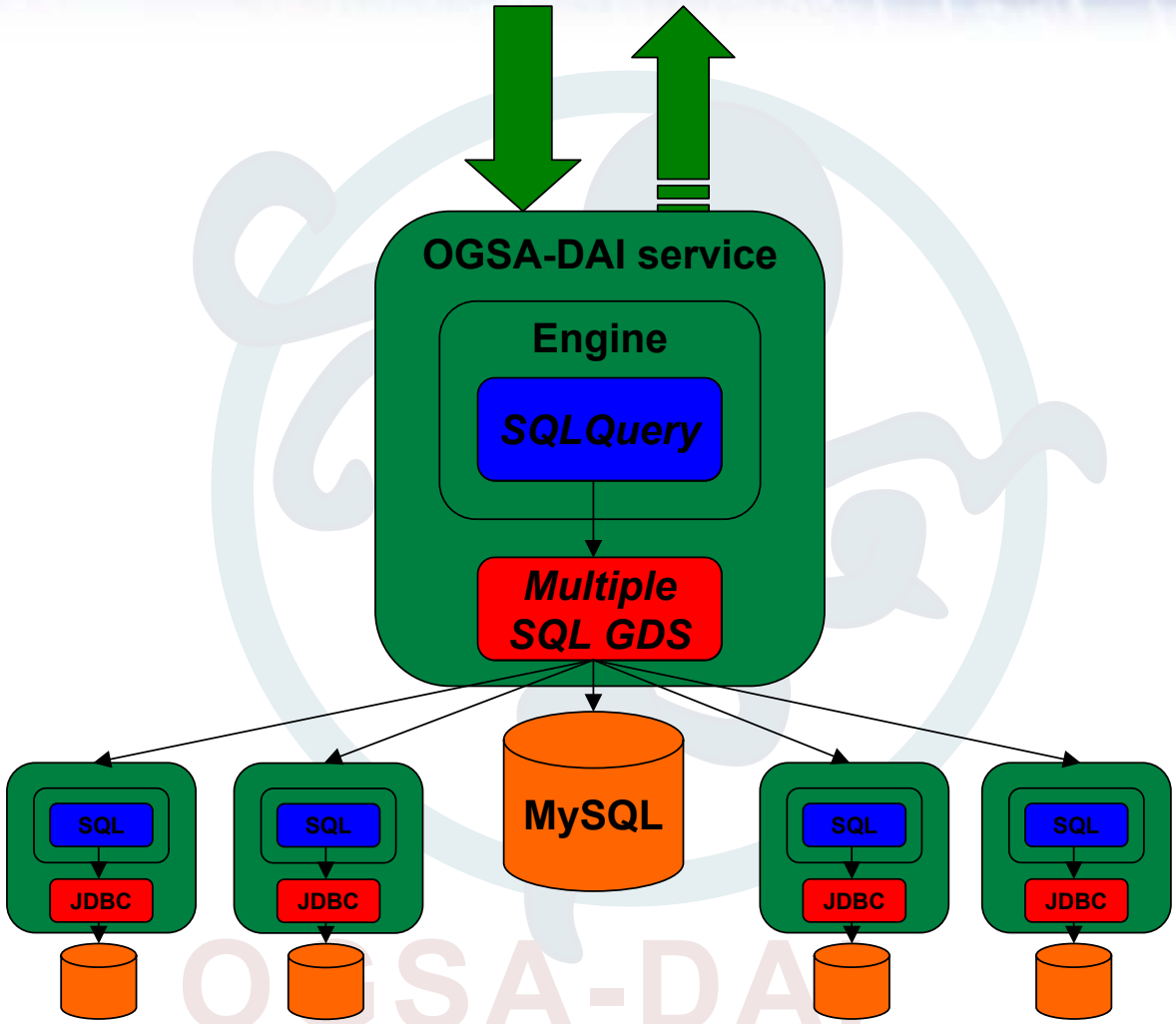
OGSA-DAI Design Principles – II

- Do not hide underlying data model
  - Users must know where to target queries
  - Data virtualisation is hard
- Extensible architecture
  - Modular and customisable
  - e.g., to accommodate stronger security
- Extensible activity framework
  - Cannot anticipate all desired functionality
  - Activity = unit of functionality
  - Allow users to plug-in their own

The OGSA-DAI Framework



Extensibility Example



Core features of OGSA-DAI – I

- A framework for building applications
  - Supports data access, insert and update
    - Relational: MySQL, Oracle, DB2, SQL Server, Postgres
    - XML: Xindice, eXist
    - Files – CSV, BinX, EMBL, OMIM, SWISSPROT,…
  - Supports data delivery
    - SOAP over HTTP
    - FTP; GridFTP
    - E-mail
    - Inter-service
  - Supports data transformation
    - XSLT
    - ZIP; GZIP
  - Supports security
    - X.509 certificate based security

Core features of OGSA-DAI – II

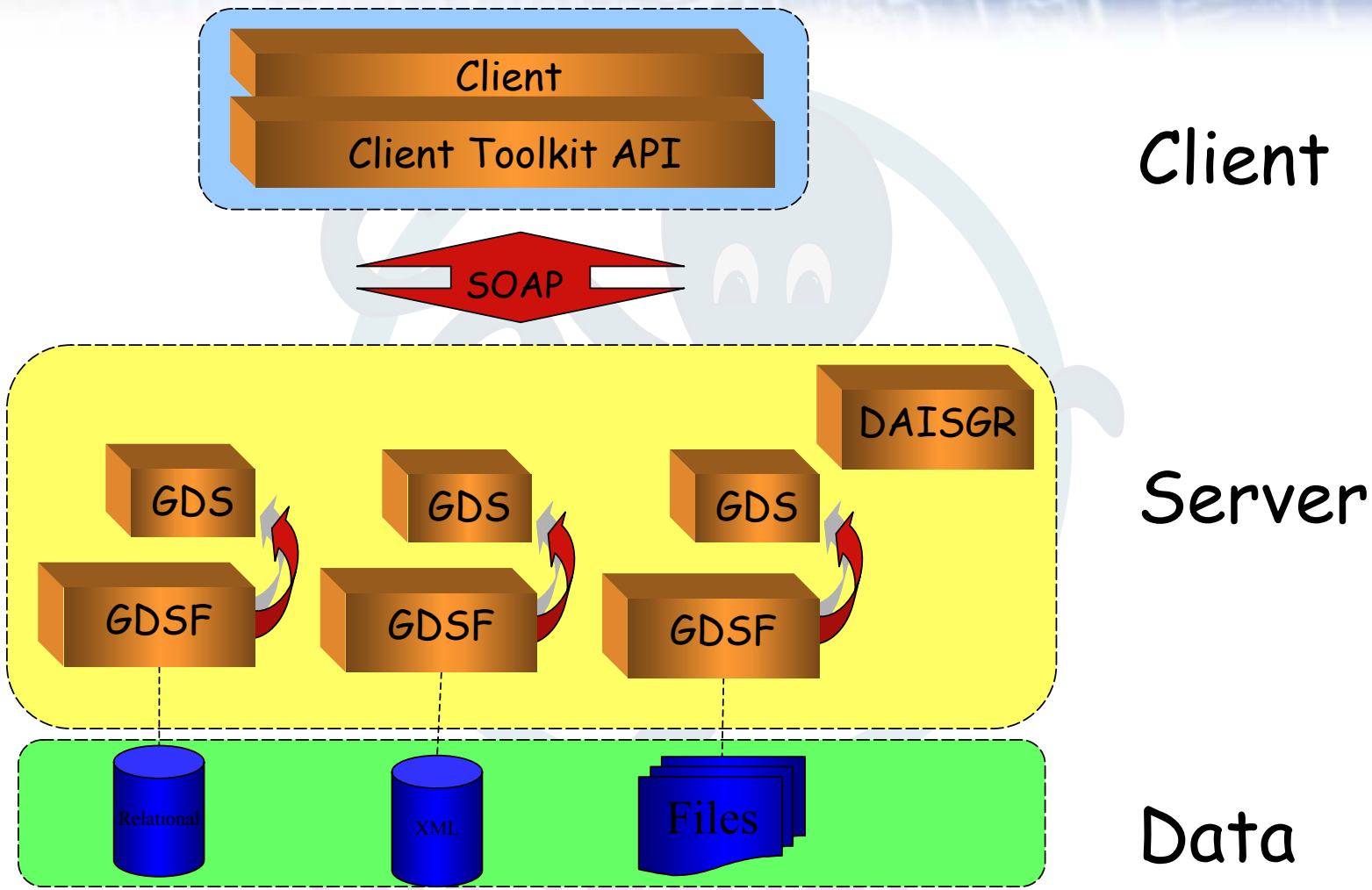
- A framework for building data clients
  - Client toolkit library for application developers
- A framework for developing functionality
  - Extend existing activities, or implement your own
  - Mix and match activities to provide functionality you need
- Highly-extensible
  - Customise our out-of-the-box product
  - Provide your own services, client-side support and data-related functionality
- Comprehensive documentation and tutorials
- Latest release supports GT3.2 (to be deprecated), GT4.0, and Axis 1.2 / OMII_2 using Java 1.4

Our stakeholders

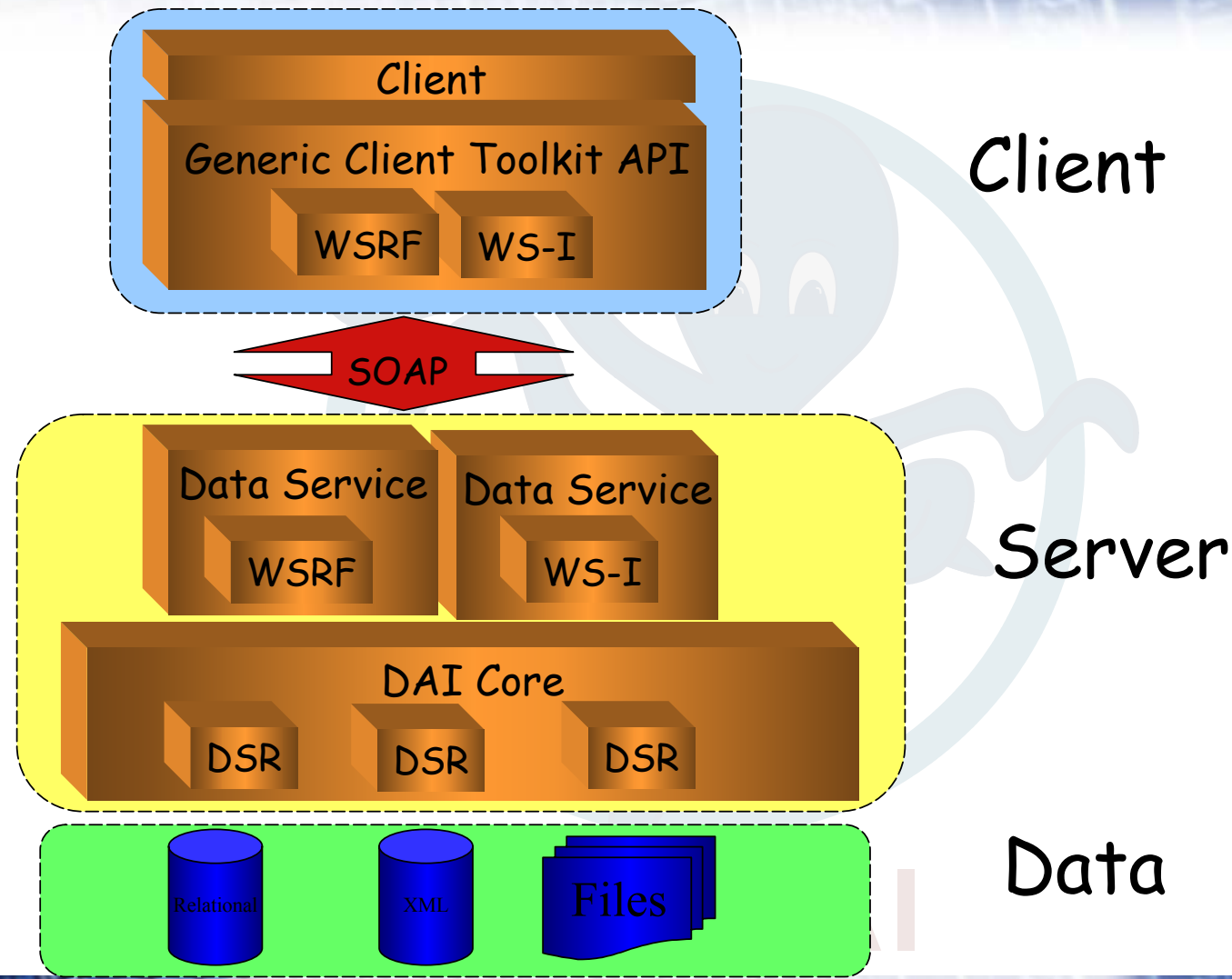
- OMII
  - Current version of OGSA-DAI WS-I distribution runs on OMII_2
    - But tie into OMII security not until OGSA-DAI WSI 2.0
- Globus
  - WSRF 0.9.6 distribution bundled with GT4.0
  - WSRF 1.0 distribution bundled with GT4.0.1
- Projects
  - Number of projects have used/use/will use OGSA-DAI

AstroGrid	Biogrid	BioSimGrid	Bridges	caGrid	DataMiningGrid
eDiamond	FirstDig	GEDDM	GeneGrid	GEON	GridMiner
INWA	IU RGRBench	LEAD	MCS	myGrid	N2Grid
ODD-Genes	OGSA-WebDB	SIMDAT	GOLD		

Out with the old…



… in with the new!



Changes in moving to WSRF/WS-I

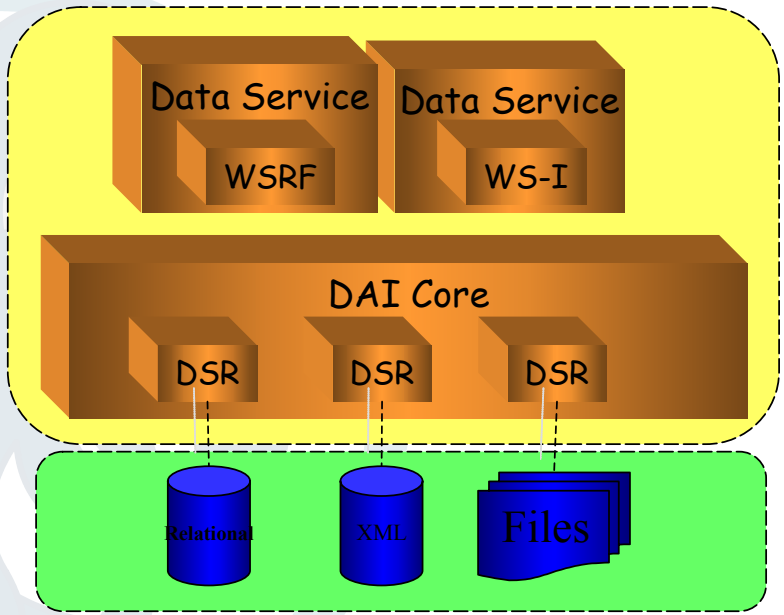
- Registry component (DAISGR) no longer supported
  - Hope to leverage of third party registration services
    - GRIMOIRES (http://www.omii.ac.uk/mp/mp_grimoires.htm)
    - Others …
- GDS/GDSF roles combined
  - Use data services
  - Currently static services but
    - Reconfigurable services
- Improvements to the GDS
  - Data resource abstraction decoupled from the service
  - Renaming (consistent naming across platform versions)
  - Ability to enforce control flow constraints (ordering activities)
  - Refactored exception framework
- Temporary set-backs (we promise we'll fix them)
  - No security model
  - No concurrency
    - Previously used GDSs for concurrency
    - Support now moving to the engine

The Client Toolkit (CTk)

- Provides programmatic abstraction for perform documents
  - Do not have to write XML explicitly
- Abstraction over WSI and WSRF services at client side
  - don't need to know what type of service is at the other end (almost)
    - security model is the remaining issue
- Currently only Java version of CTk
  - Stabilising API
  - Publish an API document
    - Allow 3rd parties to develop CTk for other programming languages

The Server Side

- Server side:
  - Presentation layer:
    - Deal with messaging differences
    - Get one version per distribution
  - Core/Business Logic:
    - Common to all distributions
    - Data Service Resource (DSR)
  - Data Layer:
    - Relational databases
    - XML document repositories
    - File based repositories
- New architecture being rolled out
  - see Malcolm's talk in next session
  - concurrency, sessions and transactions



OGSA-DAI Deck of Activities

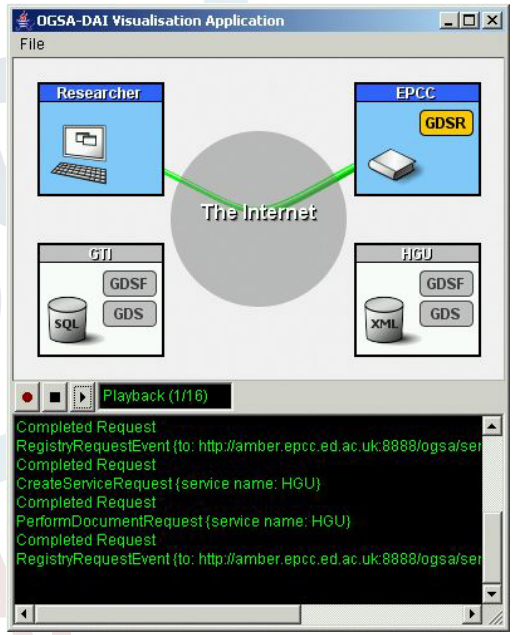
Client Toolkit

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

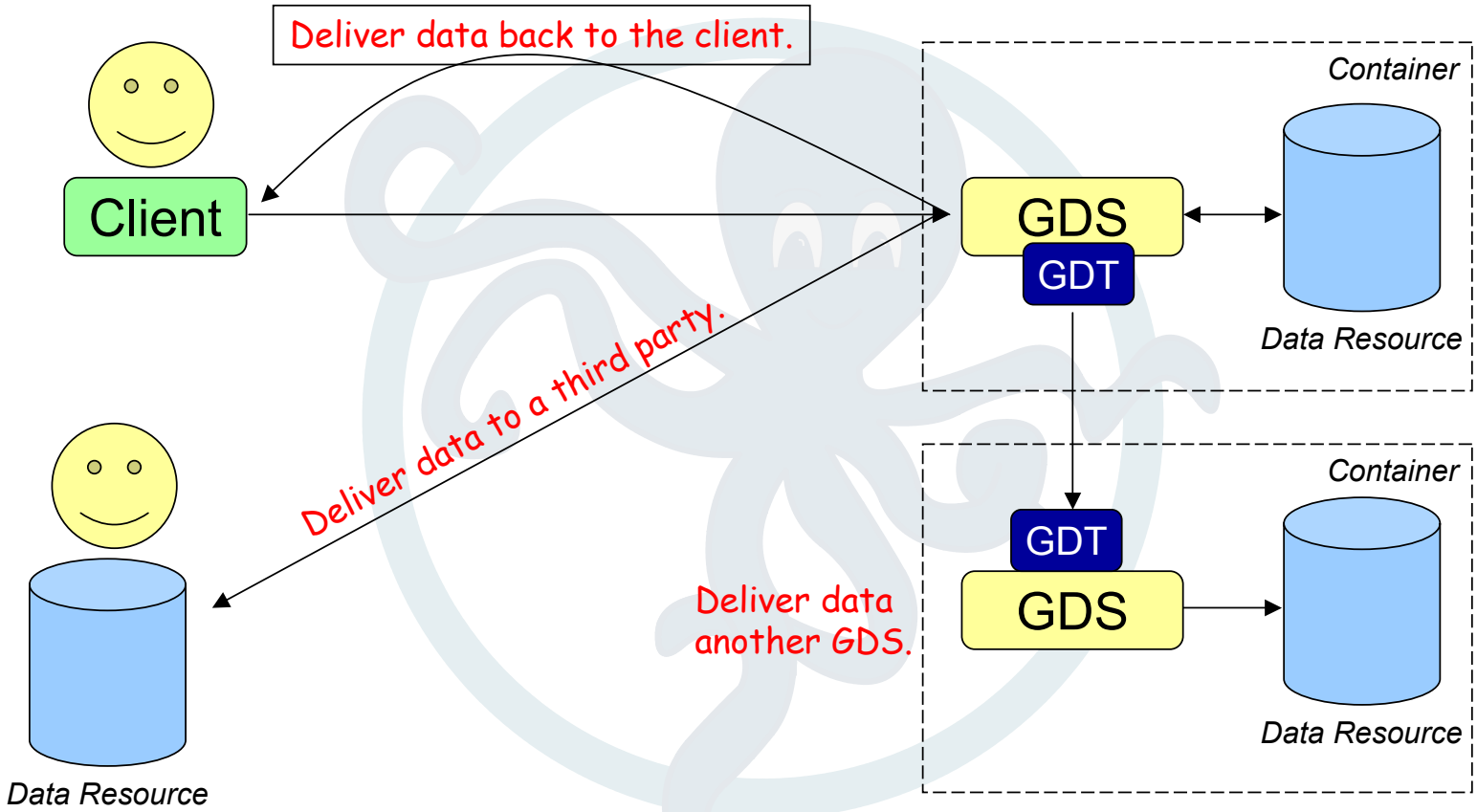
```
// Create a query
SQLQuery query = new SQLQuery(SQLQueryString);
ActivityRequest request = new ActivityRequest();
request.addActivity(query);

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

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

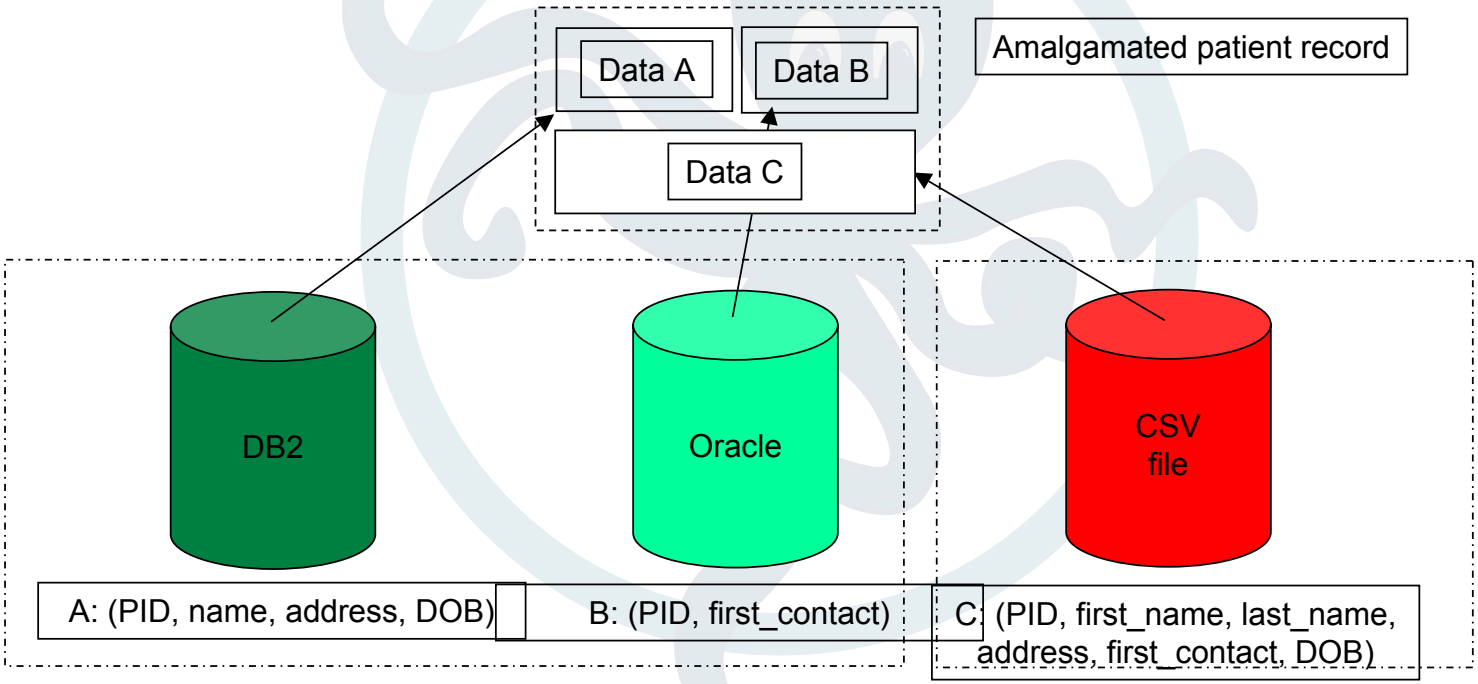


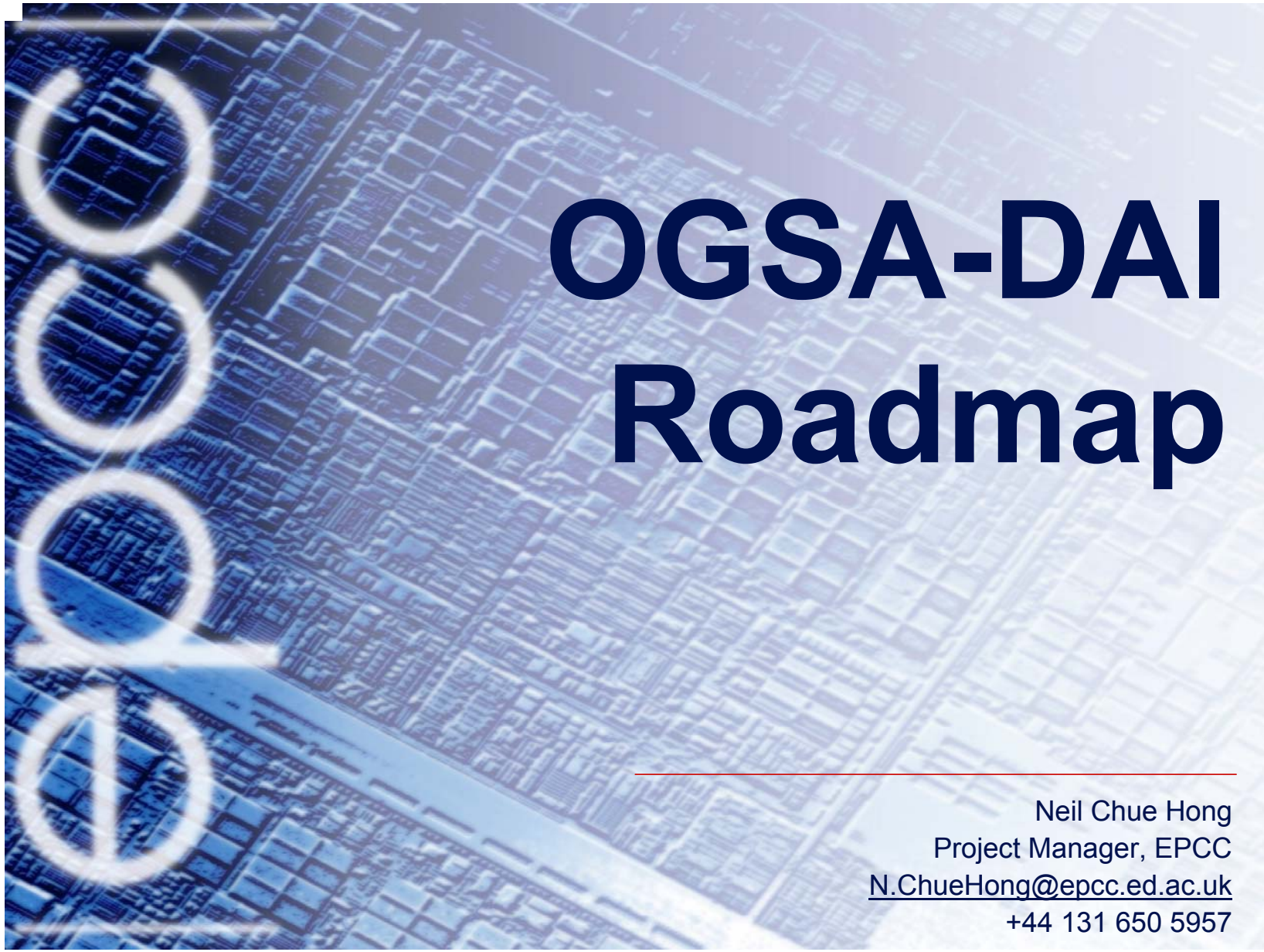
More Complex Behaviour



Integration Scenario

- A patient moves hospital





OGSA-DAI Roadmap

Neil Chue Hong
Project Manager, EPCC
N.ChueHong@epcc.ed.ac.uk
+44 131 650 5957

Roadmap / Workplan

- Roadmap documents available for comment:
  - http://www.ogsadai.org.uk/docs/OtherDocs/OGSA-DAIRoadmapV3.0.pdf
  - User feedback required to drive this document
  - New version of roadmap being produced
- Integrate parts of DQP into OGSA-DAI core
  - Addressing platform dependencies
  - Want to include XML data resources
- Move Computation to Data
  - Java mobile code?
- New version of engine
  - Better support for concurrency and threading

Release 6

- A release to consolidate multiple platforms
  - Release in May 05
- Features:
  - Renaming (consistent naming across platform versions)
  - Exceptions
  - Dynamic Service Configuration
  - Client Toolkit for each platform

Next Release R7.0

- Features:
  - Security (WS-RF, Authorisation interface)
  - Sessions and Transactions support
  - Seamless Multiplatform Client Toolkit
  - Integration of DQP
  - Data Service Concurrency
  - New Engine design
- Expected release in October 2005
- More explanation later…

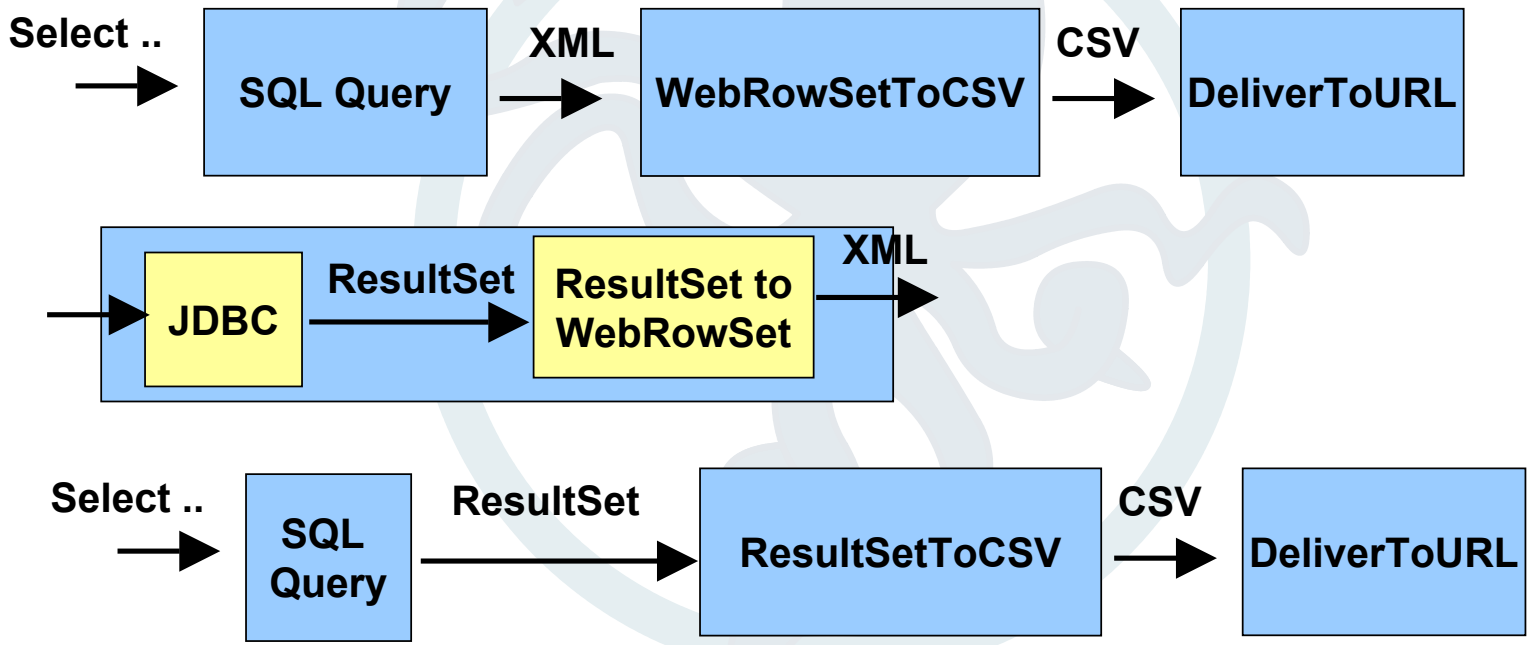
Future versions of OGSA-DAI

- Post September 2005
  - work being done now to prepare for new versions
- New architecture to provide better support for:
  - sessions
  - transactions
  - concurrency
  - security
- Implementing new versions of DAIS specifications
- Key things that we will be addressing after Release 7:
  - Performance
  - A Security Model which can be applied across platforms
  - Full Transactions provision, including implementation of compensatory activities, distributed transactions
  - More data integration facilities
  - Better abstraction over DBMS variation

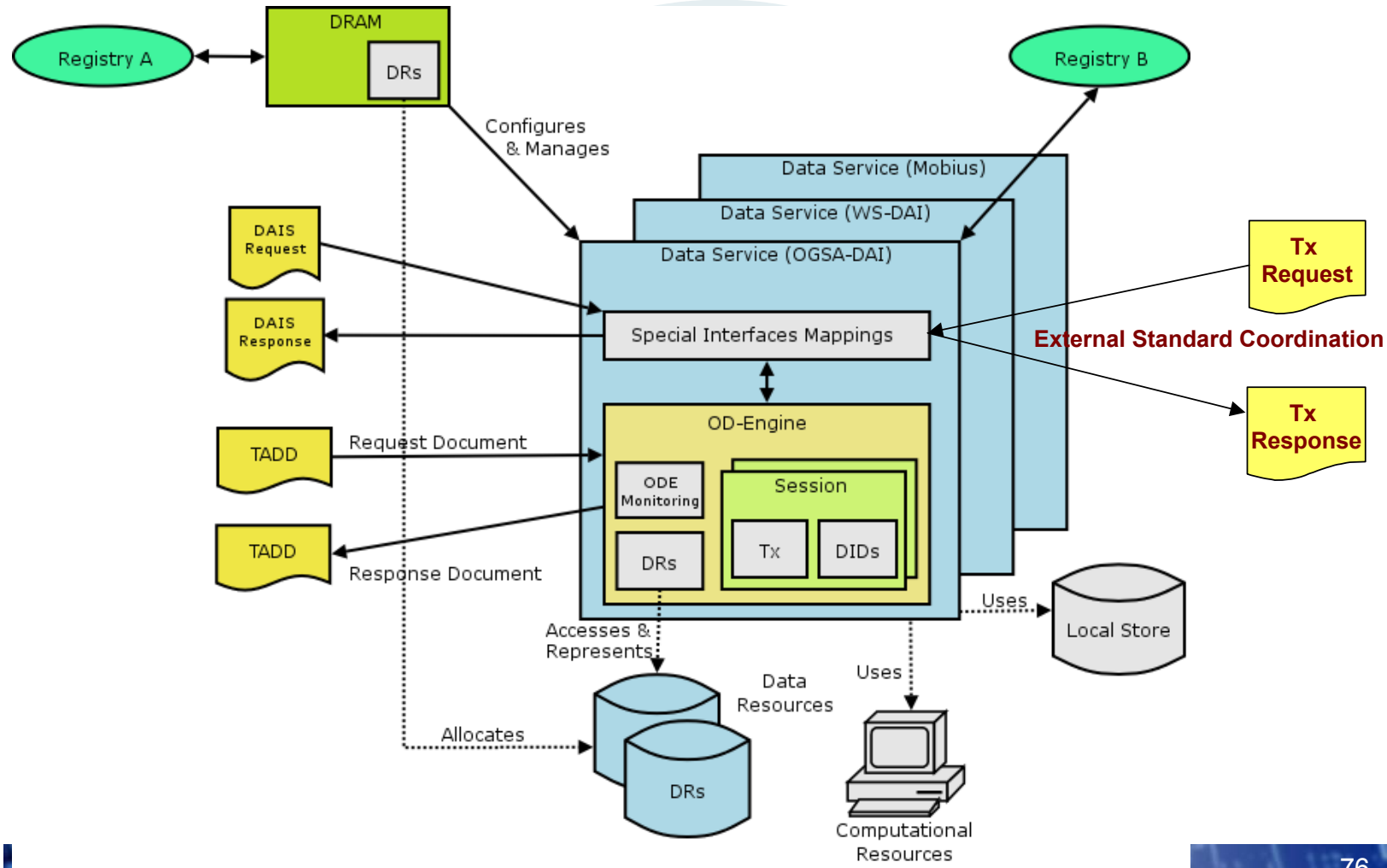
Challenges

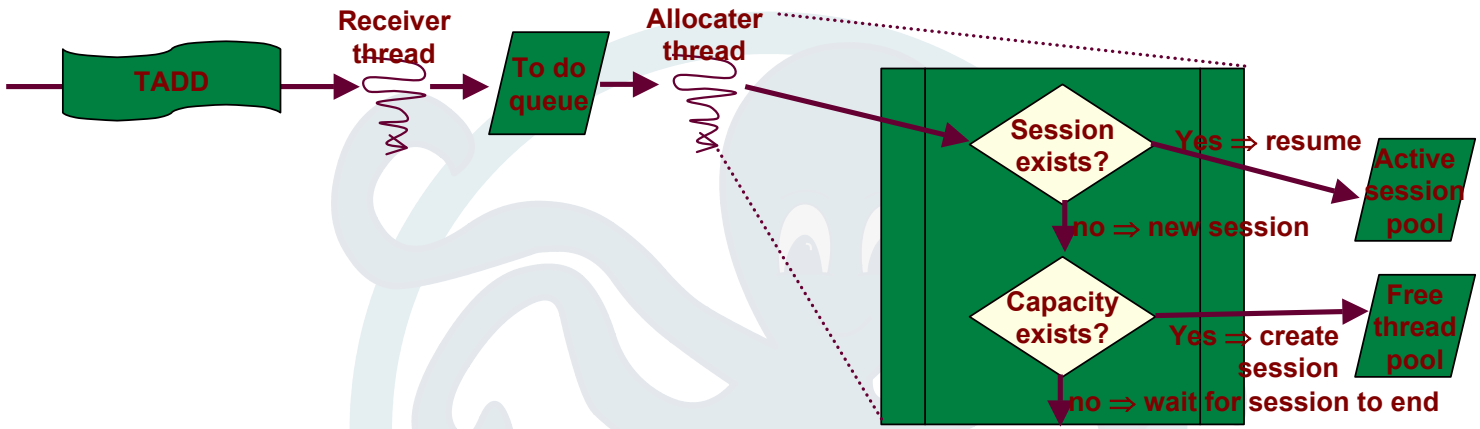
- Metadata extraction
  - define a common model for e.g. database schema?
- Intermediate representation
  - between multiple models (relational, XML,…)
  - XML WebRowSet is flexible (c.f. GridMiner) but expansive
  - DFDL and GridFTP/parallel HTTP?
- Query definition
  - translation of queries
- Data transport and workflow
  - workflow is typically compute driven
- Move computation to data
  - mobile code activities?
  - data services hosted on DBMS?

User controlled format conversions

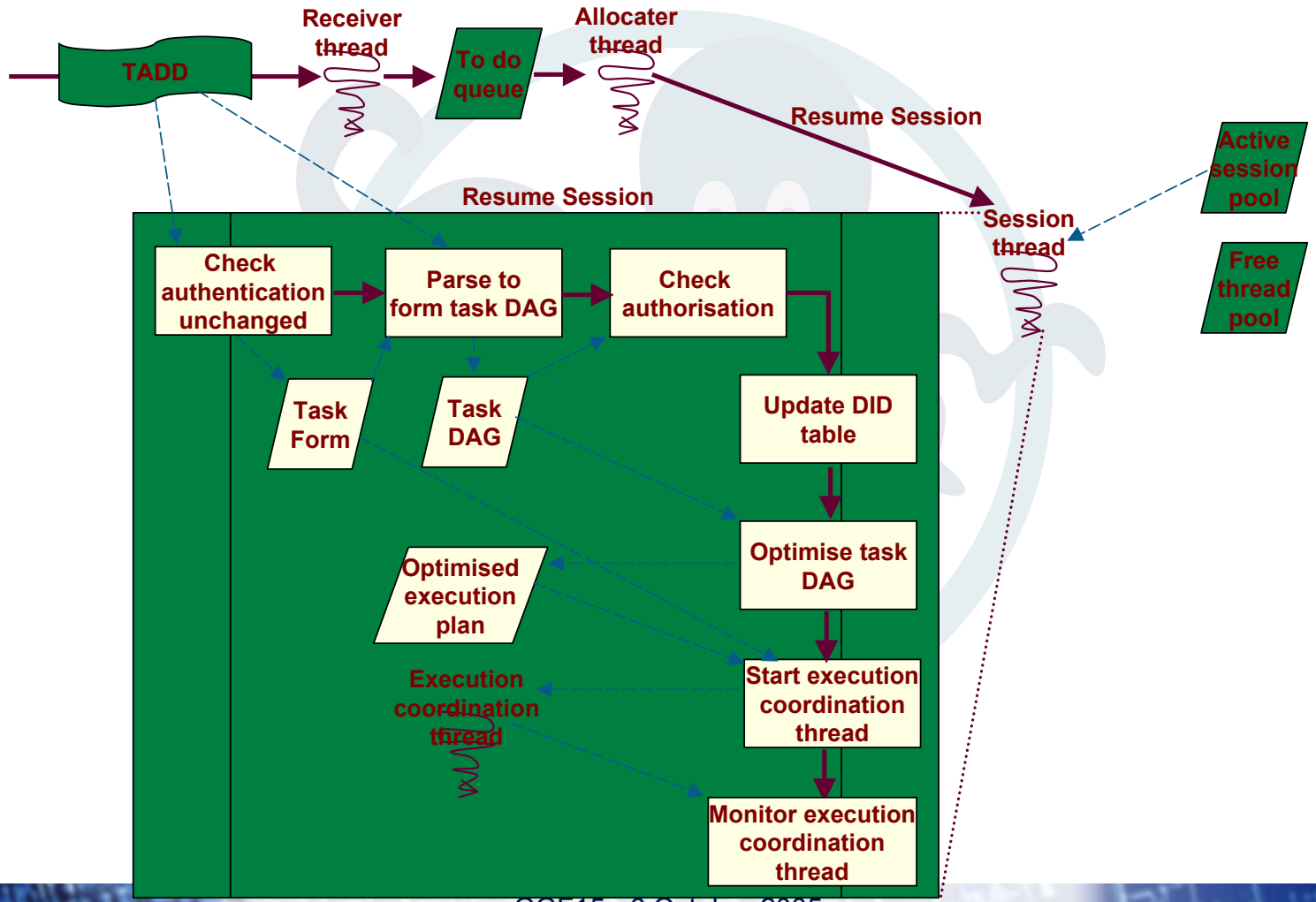


New OGSA-DAI Architecture

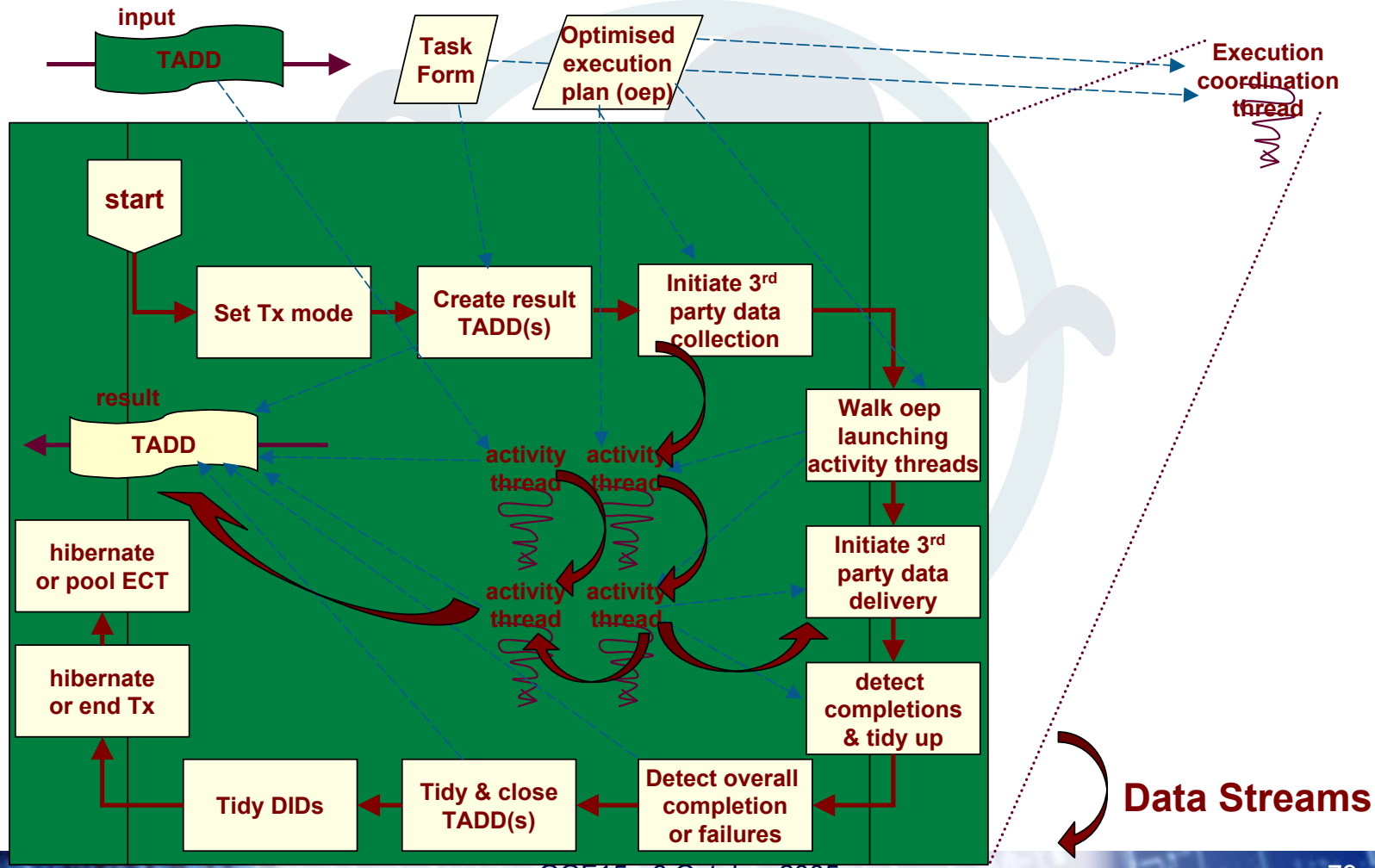



TADD
Receiver thread
To do queue
Allocater thread
Session exists?
Yes ⇒ resume
Active session pool
no ⇒ new session
Capacity exists?
Yes ⇒ create session
Free thread pool
no ⇒ wait for session to end

Execution Model – Execution Planning



Execution Model – Processing one Request



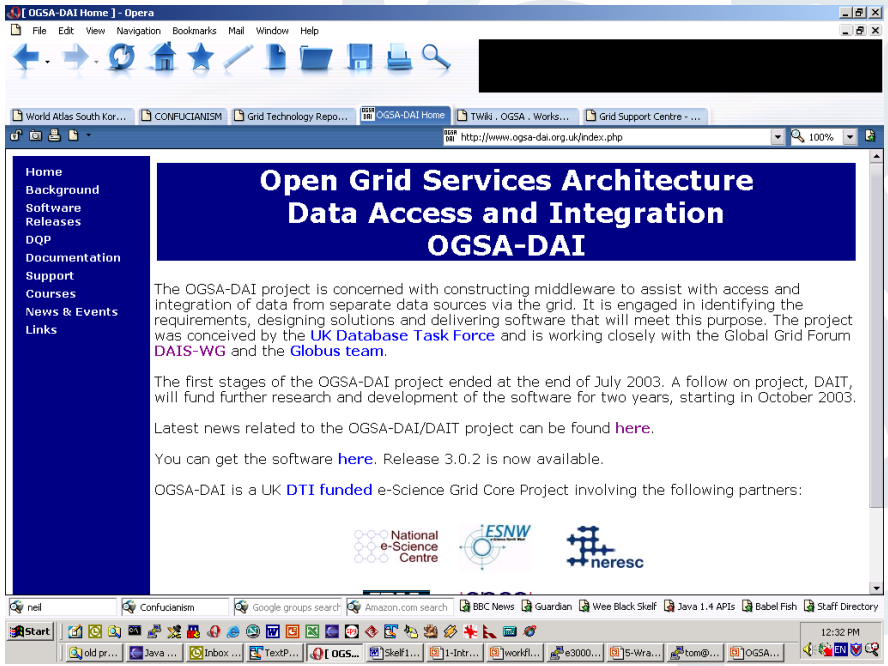
Contributing to OGSA-DAI

- Additional functionality:
  - Provide activities which implement specific functionality
  - Provide extra client functionality
  - Provide different security mechanisms
  - Provide higher level components and applications

- Different levels of contributions
  - Based on OGSA-DAI?
  - Works with OGSA-DAI?
  - Part of OGSA-DAI?

OGSA-DAI Project Webpage

- http://www.ogsadai.org.uk



Background

News & Events

Software Releases

Documentation

On-line Tutorials

Support

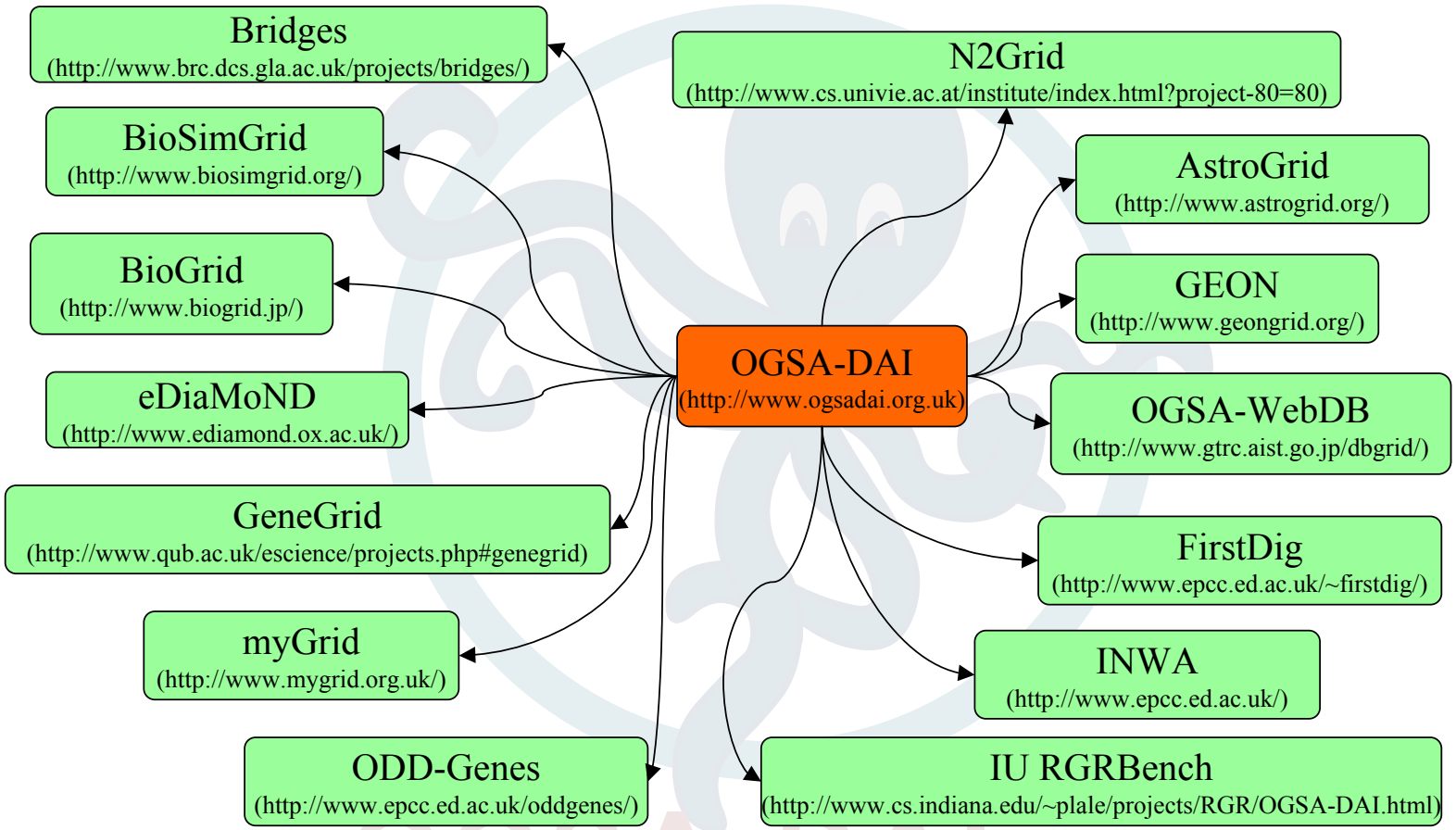
Training Courses

Links

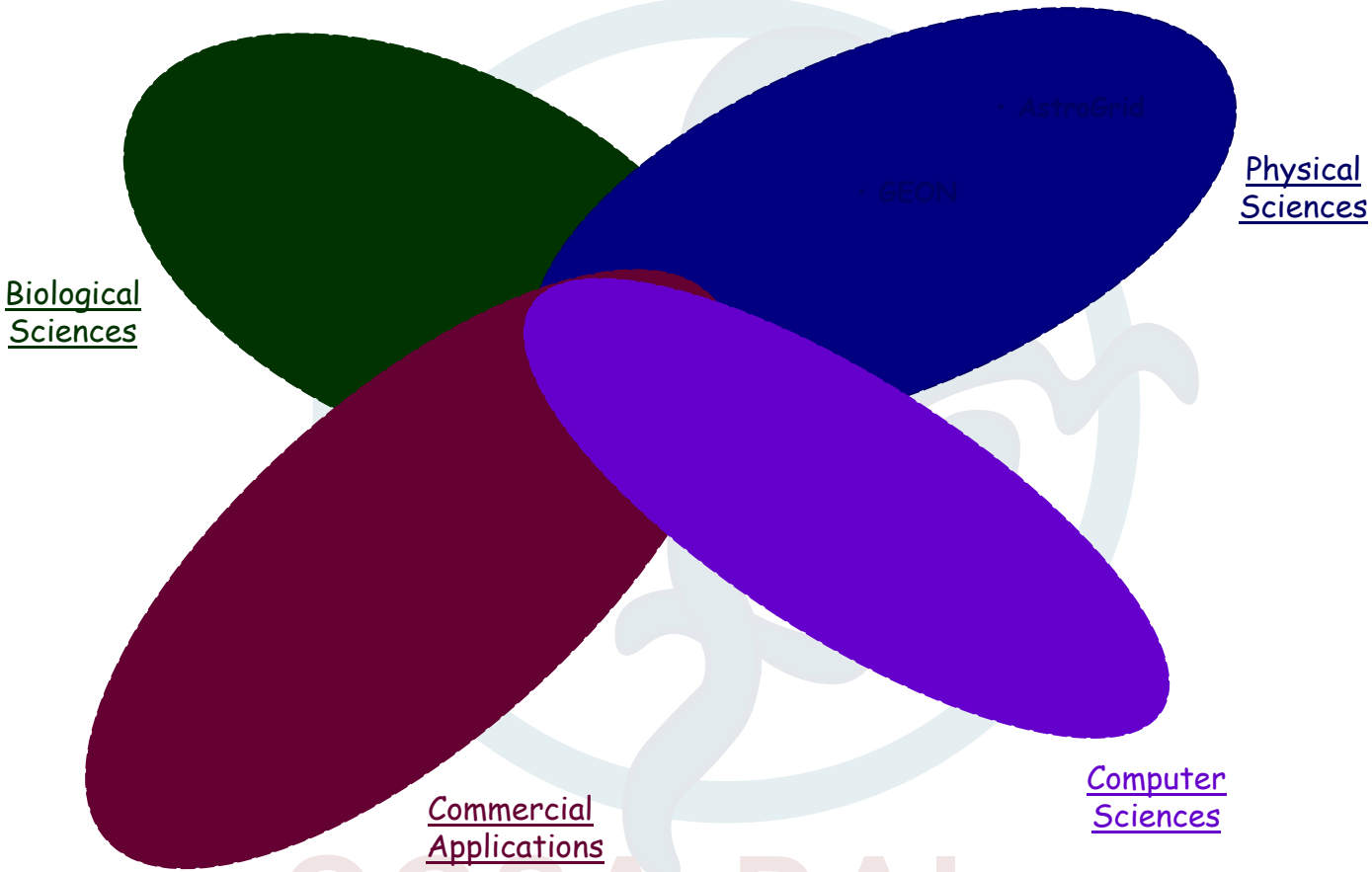
Projects Using OGSA-DAI

Neil Chue Hong
Project Manager, EPCC
N.ChueHong@epcc.ed.ac.uk
+44 131 650 5957

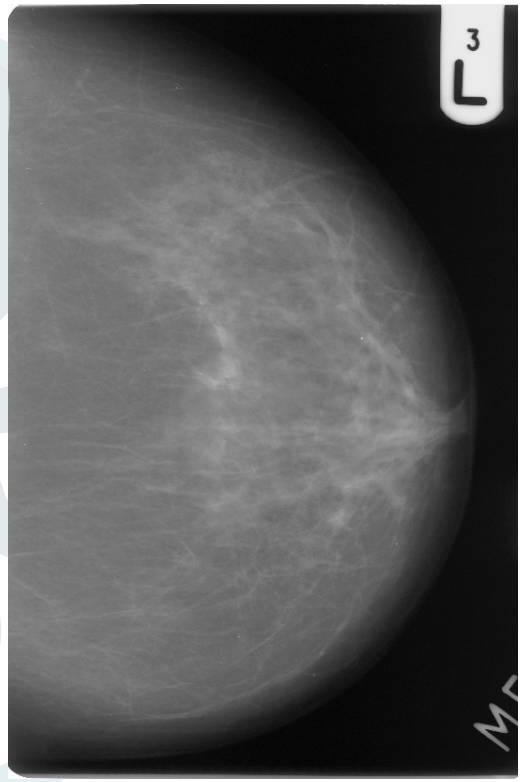
Projects Using OGSA-DAI



Project classification

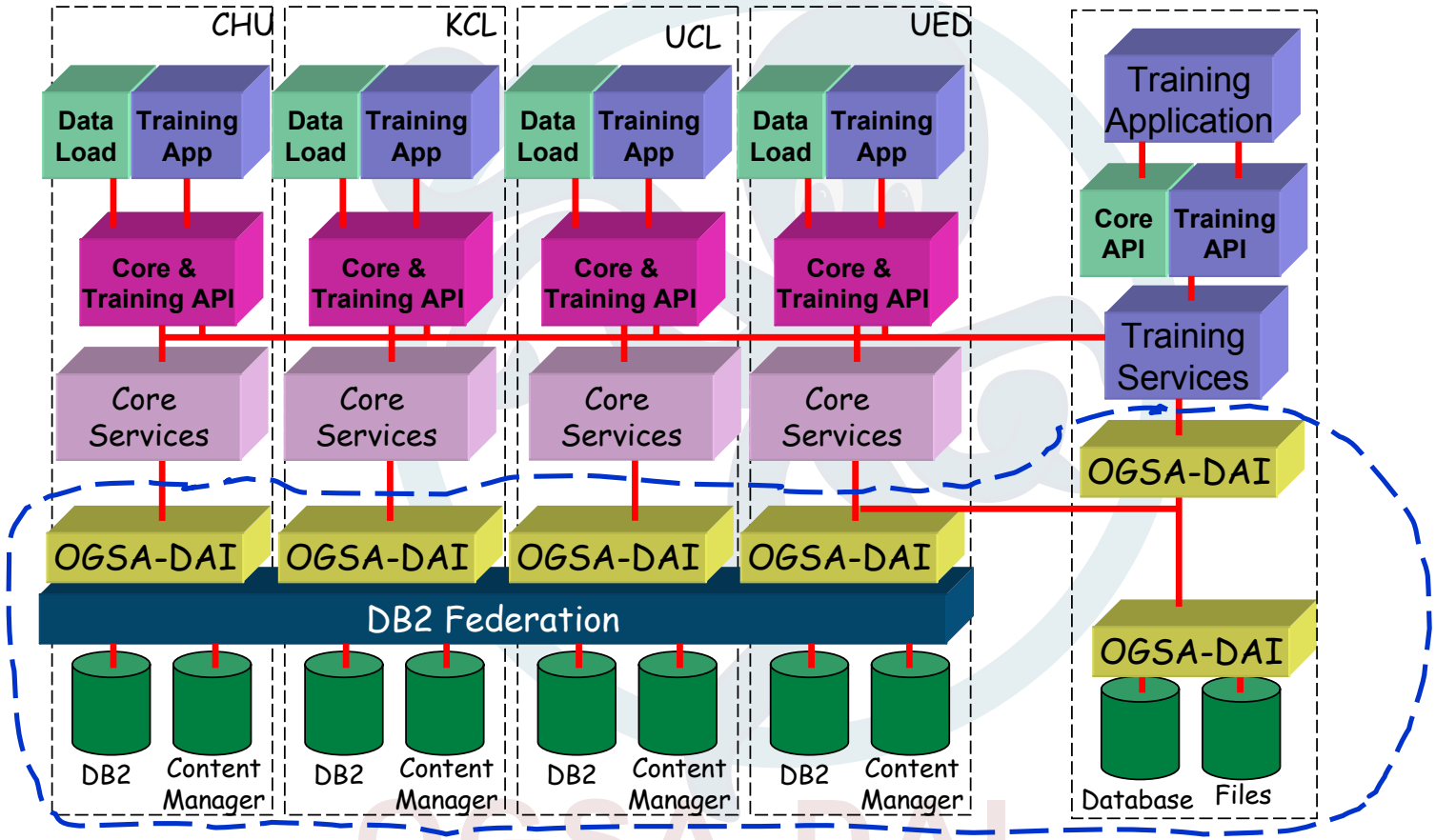


- e-Digital MammOgraphy National Database
  - –Mammogram - X-ray of the breast
- Built prototype of a national database of mammographic images
  - –In support of the UK Breast screening programme
- Employed Grid technologies to facilitate process



Thanks to eDiaMonND project and the Digital Database for Screening Mammography for this image.

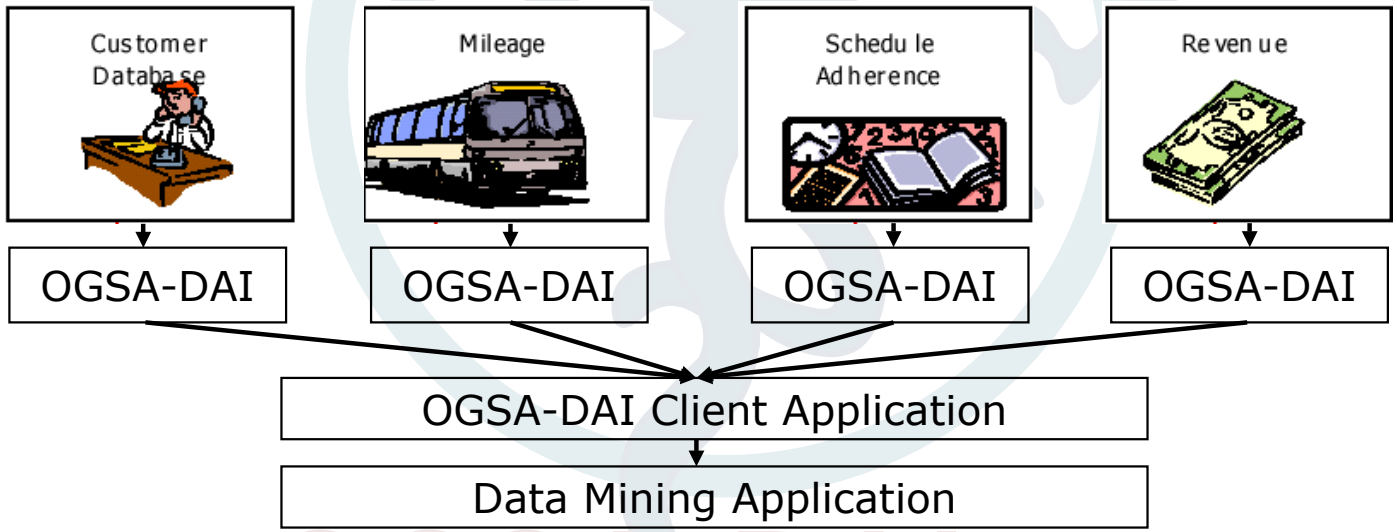
- Breast screening in the UK began in 1988
  - Women aged 50-64 screened every 3 Years
  - Women aged 50-70 from 2004
  - 1 View/Breast → 2 views by 2003
- UK has
  - Over 90 Breast screening units throughout the UK
  - Each one deals with about 45000 women on average p.a.
- Each centre sees 5000-20000 images/year
- In 2001-02 → 2002-03
  - Screened: 1.4M → 1.5M
  - Recalled for Assessment : 77911 → 79441
  - Cancers detected : 10003 → 10467
  - Lives per year Saved: 300 → 1250 (by 2010)
- Distributed team of doctors perform the analysis


CHU
KCL
UCL
UED
Data Load
Training App
Core & Training API
Core Services
OGSA-DAI
DB2 Federation
DB2
Content Manager
Training Application
Core API
Training API
Training Services
Database
Files

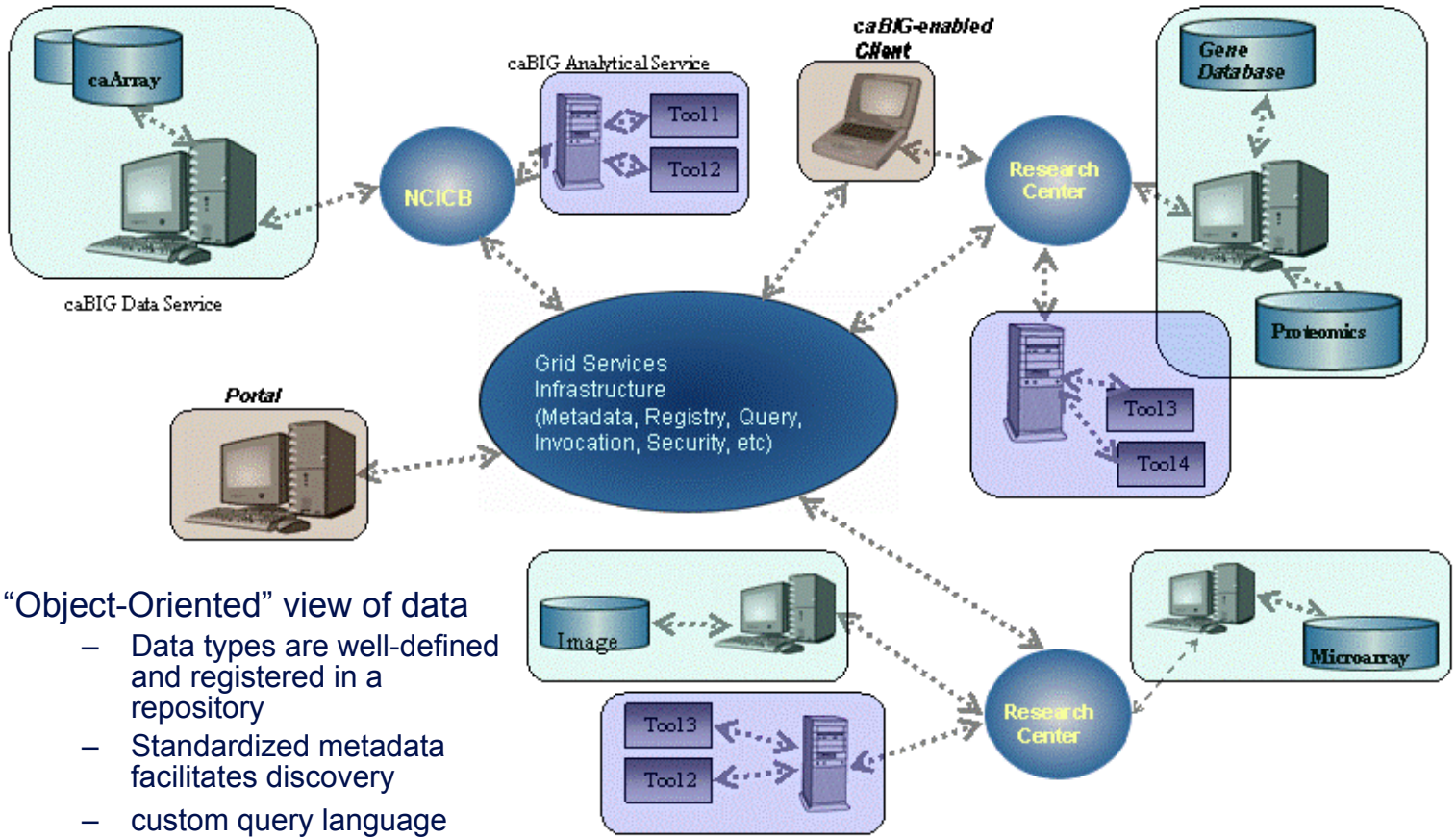
- eDiaMoND Findings:
  - OGSA-DAI provides a flexible framework
  - Dynamically configure the system through discovery
  - Activities can operate with different levels of granularity
  - Federation can be introduced at various levels
  - Good documentation on how to extend the framework
    - Extended Activities to access IBM DB2 Content Manager
- IBM have released II wrapper for OGSA-DAI

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%”
  - *"The results of this exercise will revolutionise the way we do things in the bus industry.“,* Darren Unwin, Divisional Manager, First South Yorkshire.



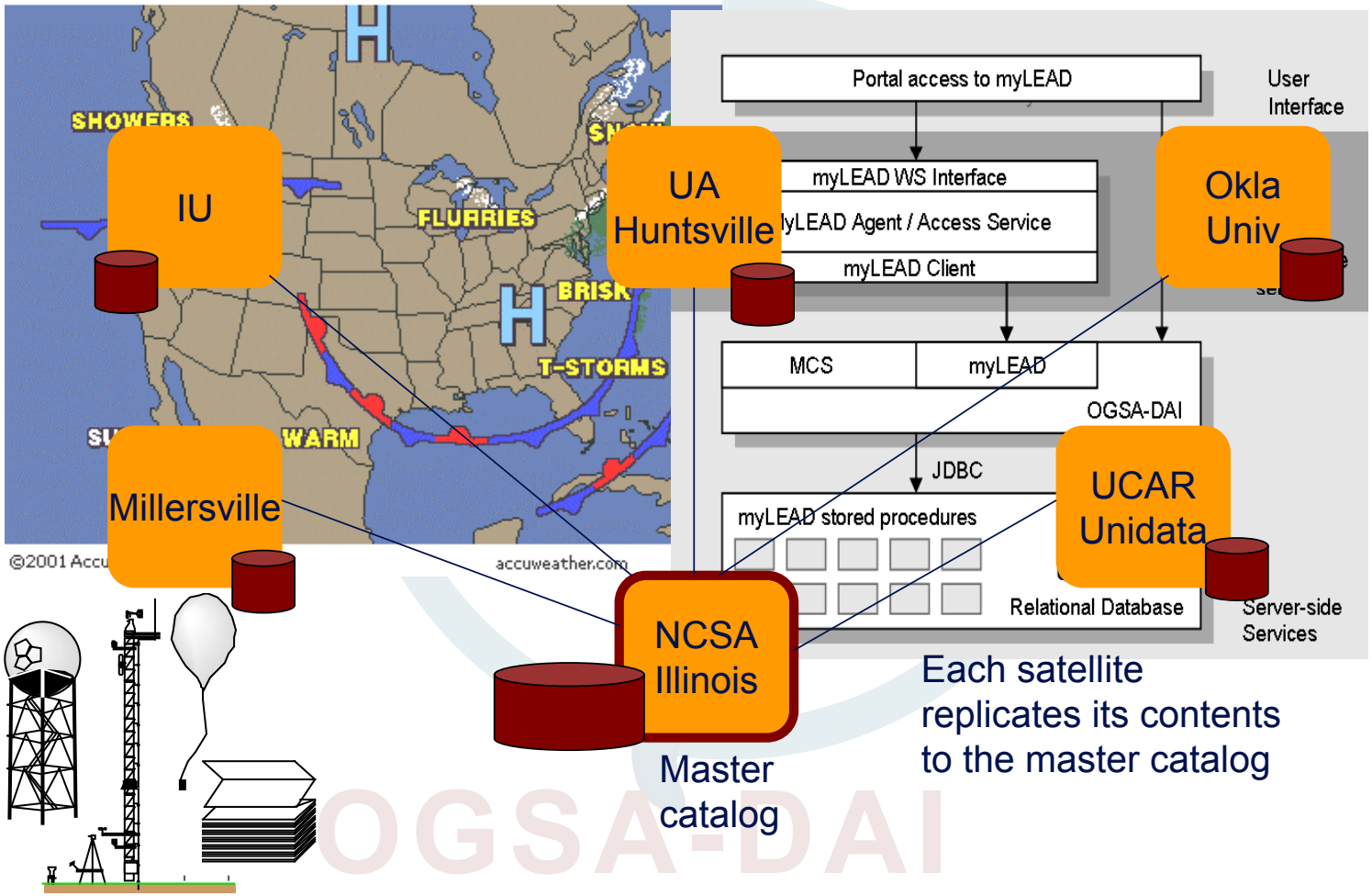
caBIG



"Object-Oriented" view of data

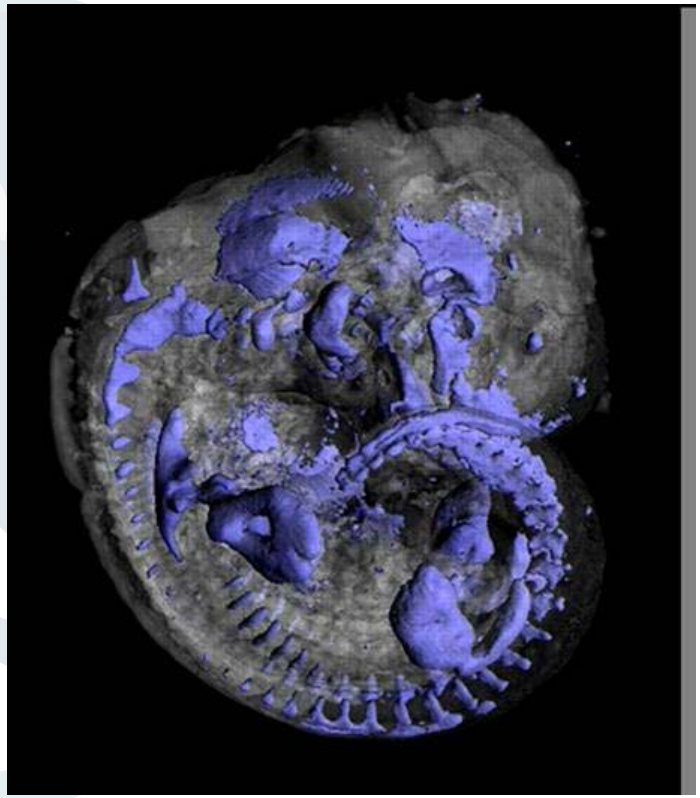
- Data types are well-defined and registered in a repository
- Standardized metadata facilitates discovery
- custom query language implemented as an activity

LEAD

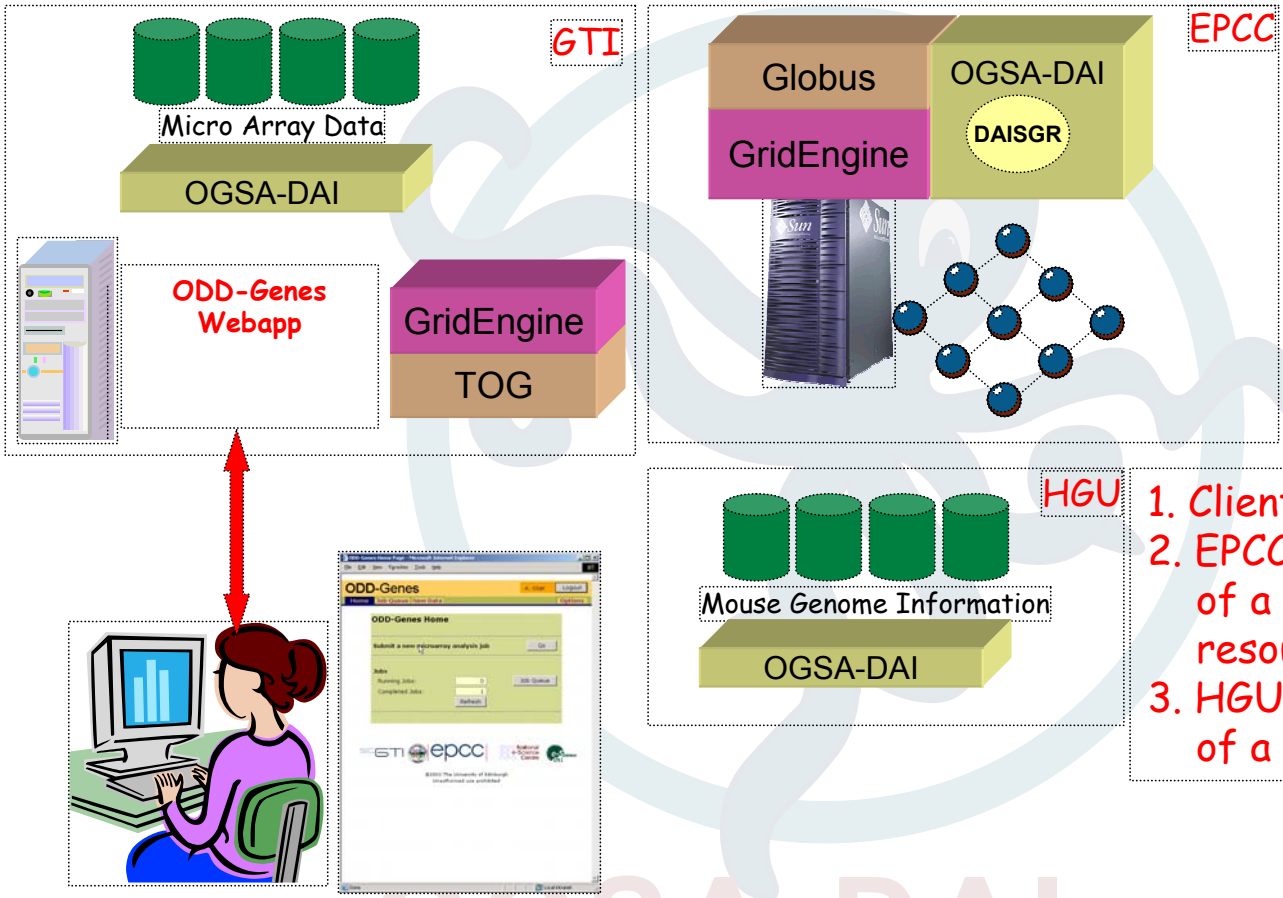

SHOWERS
FLURRIES
BRISK
T-STORMS
WARM
©2001 Accu
accuweather.com
IU
UA Huntsville
Okla Univ
Millersville
NCSA Illinois
UCAR Unidata
Portal access to myLEAD
User Interface
myLEAD WS Interface
lyLEAD Agent / Access Service
myLEAD Client
MCS
myLEAD
OGSA-DAI
JDBC
myLEAD stored procedures
Relational Database
Server-side Services
Master catalog
Each satellite replicates its contents to the master catalog
OGSA-DAI

ODD-Genes

- OGSA-DAI Demo for Genetics
- Collaboration between
  - –EPCC
  - –Scottish Centre for Genomic Technology and Informatics (GTI)
  - –Human Genetics Unit (HGU)
- ODD-Genes demonstrates:
  - –Perform high-speed batch analysis of microarray data on the Grid
  - –Browse the results of previous analyses stored in a database
  - –View data from arbitrary databases as HTML
  - –Discover related databases on the Grid
  - –Perform coupled queries on newly-discovered databases to provide a richer analysis of gene data



ODD-Genes Actors

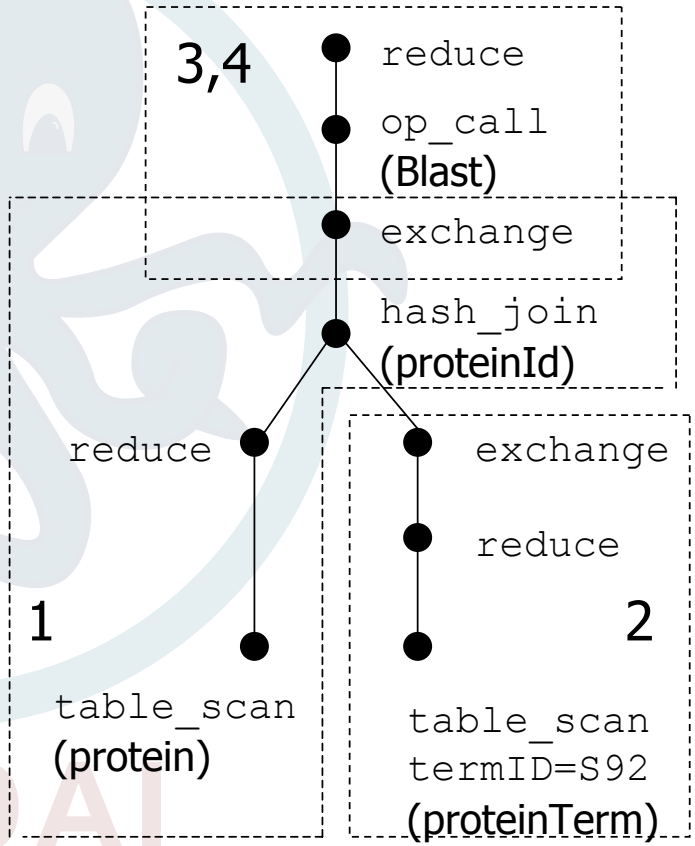


ODD-Genes Findings

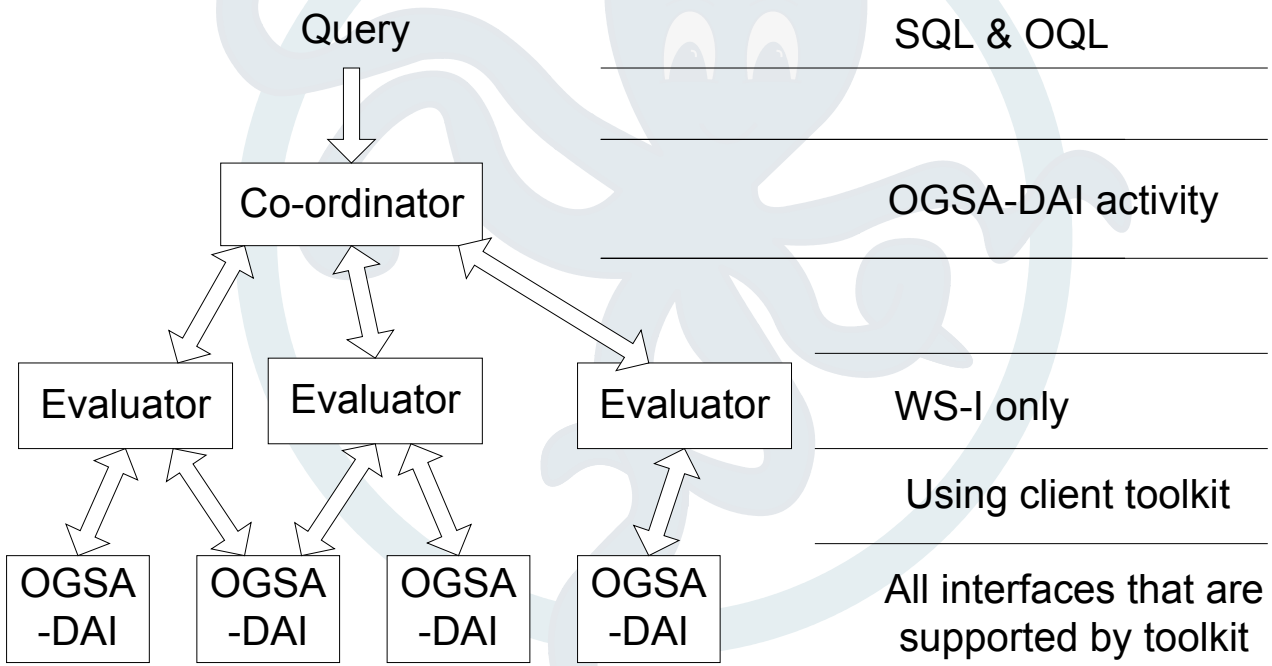
- Data discovery perceived to be very important
  - Map data views: time -> spatial locations
  - Discovery of new resources
- Transparency to data access
  - @HGU had an XML database
  - @GTI had a relational database
  - Deploy OGSA-DAI and not worry about databases
- Issues
  - Registry maintenance policy
  - Semantics of the discovery process
  - Groups working the same area but different schemas, no generic metadata (schemas were the effective metadata)
- Provides an additional tool for researchers

Distributed Query Processing

- Higher level services building on OGSA-DAI
- Queries mapped to algebraic expressions for evaluation
- Parallelism represented by partitioning queries
  - –Use exchange operators



DQP architecture



DQP Configuration

- Configured via operation specifying set of OGSA-DAI services to integrate
- Obtains the schema details of these services and generates global schema
- Global schema is used by client
- DQP also obtains meta-data such as table sizes etc. from remote services to use in query optimisation
- Clients interact with configured DQP services as if they were interacting with normal OGSA-DAI service

Distributed Query Processing

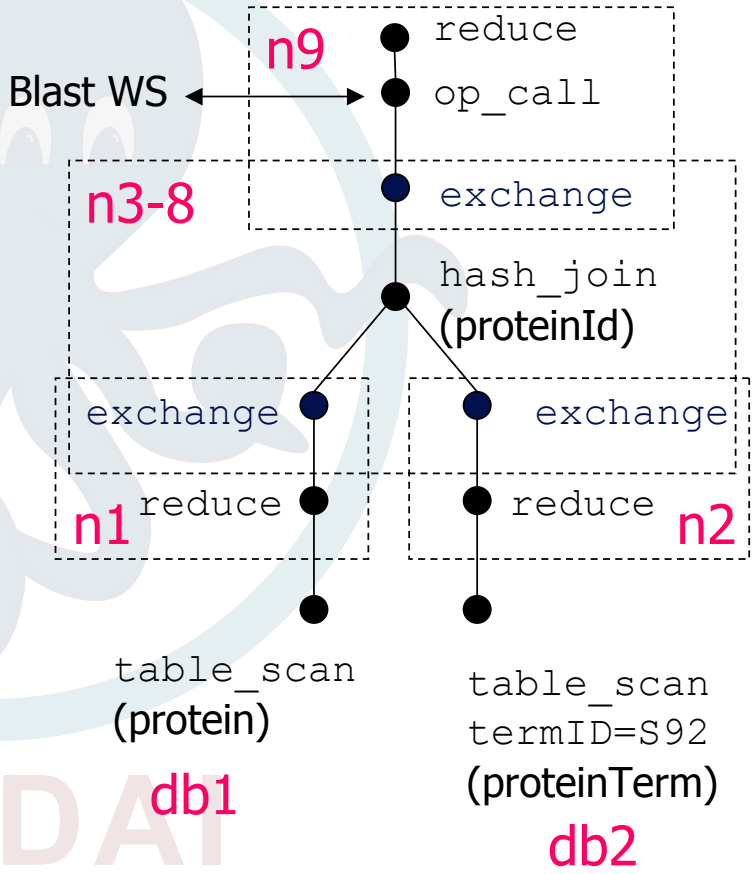
- Queries are expressed in OQL
  - allows computations to be included in the query
- A single query may reference data at multiple sites
  - the data locations may be transparent to the query author

```
select p.proteinId, Blast(p.sequence)
from   protein p, proteinTerm t
where  t.termId = 'S92' and
       p.proteinId = t.proteinId
```

Execution Plan

```
select p.proteinId,
   Blast(p.sequence)
from protein p,
      proteinTerm t
where
 t.termId = 'S92' and
 p.proteinId = t.proteinId
```

- The plan is split in to a set of partitions
- Compute resources are dynamically acquired to execute the partitions
  - –in parallel where possible, required and affordable



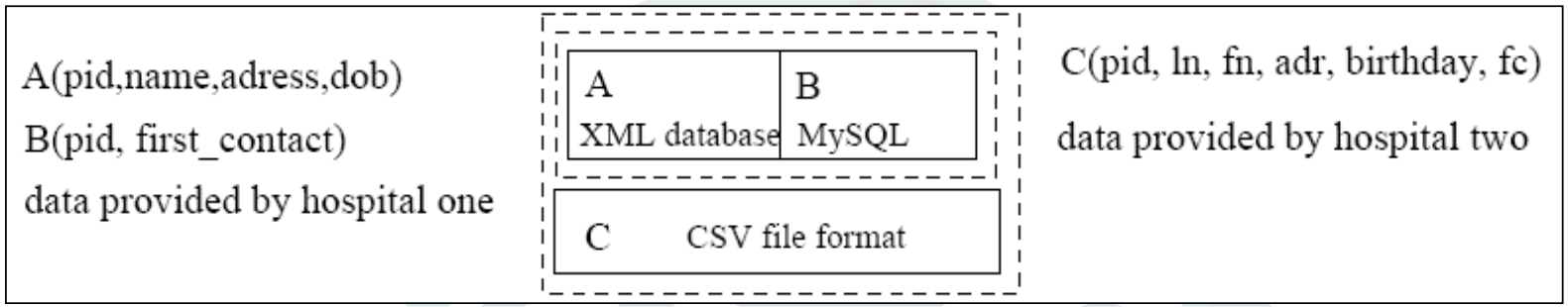
DQP Research issues

- Currently DQP makes decisions based on compute resources available at optimisation time
- What if they change?
  - New (faster, cheaper) node available
  - Node failure
- What if estimates made by optimiser prove to be significantly inaccurate
- Can we dynamically add, remove and replace nodes during execution?
- DQP teams at Manchester and Newcastle investigating these issues.

GridMiner

- Test application area: medical
  - traumatic brain injury treatment
  - Predicting the outcome of seriously ill patients
  - analytical part focuses on data mining and On-Line Analytical Processing (OLAP)
- Target:
  - provide tools to discover and access relevant knowledge and information from different distributed and heterogeneous data sources
  - building on and extending OGSA-DAI
- http://www.gridminer.org/

GridMiner Scenario



- Heterogeneities:
  - Name in A is „First Last“ (as the target format)
  - Name in C has to be combined
- Distribution:
  - 3 data sources

Summary

- New technology
  - Standardisation process still ongoing
  - Infrastructure still developing
- OGSA-DAI acting as an enabler
  - Showing people what can be done
  - Evolving and improving with each release
- Usage patterns are similar
  - Call for people to work together to solve similar problems
  - Try to implement in core OGSA-DAI
- Some problems are not OGSA-DAI specific
  - Metadata, time zones, security, …
- Data discovery opens up a window of integration opportunity
- Please try it out!
  - It's free and supported
  - Make suggestions, extend functionality, contribute to DAIS-WG

Summary

- Data Services help to provide solutions for common data scenarios
  - heterogeneity, location, integration, discovery
- OGSA-DAI is an implementation of data services
  - for GT4 / OMII_2 / Axis 1.2
- OGSA-DAI is highly extensible and flexible for developers
  - more on this in the afternoon
- Projects drive the requirements and roadmap for OGSA-DAI
  - please contribute!

Demo

Neil Chue Hong
Project Manager, EPCC
N.ChueHong@epcc.ed.ac.uk
+44 131 650 5957

Neil Chue Hong
Project Manager, EPCC
N.ChueHong@epcc.ed.ac.uk
+44 131 650 5957

Further information

- The OGSA-DAI Project Site:
  - http://www.ogsadai.org.uk
- The DAIS-WG site:
  - http://forge.gridforum.org/projects/dais-wg/
- OGSA-DAI Users Mailing list
  - users@ogsadai.org.uk
  - General discussion on grid DAI matters
- Formal support for OGSA-DAI releases
  - http://www.ogsadai.org.uk/support
  - support@ogsadai.org.uk
- OGSA-DAI training courses