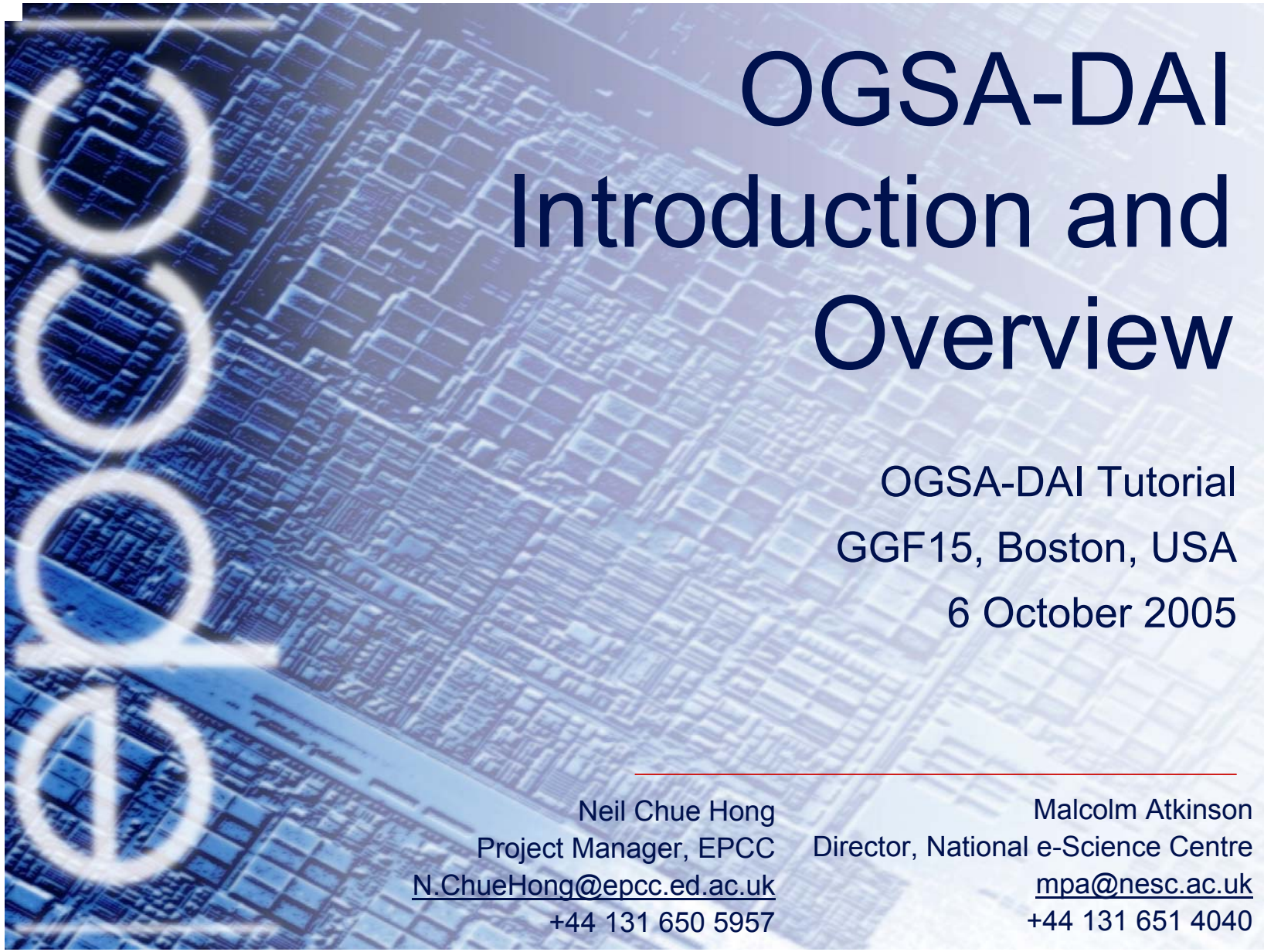


OGSA-DAI Introduction and Overview

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

Malcolm Atkinson
Director, National e-Science Centre
mpa@nesc.ac.uk
+44 131 651 4040

Data, Data Everywhere & …

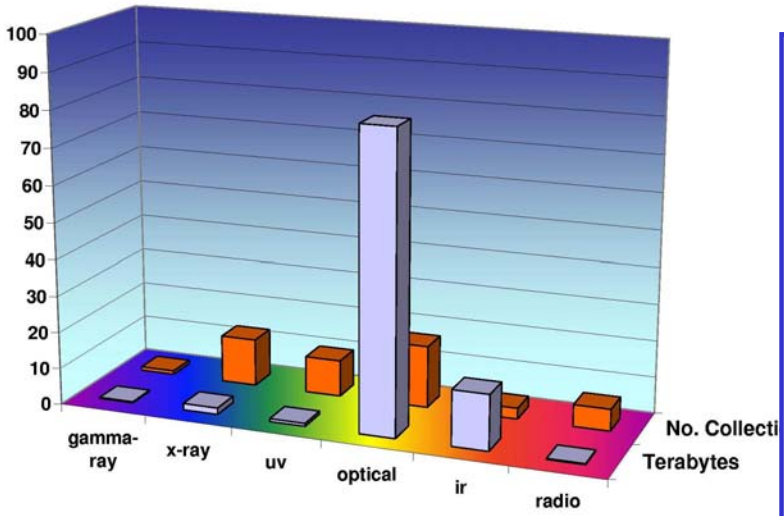
- Growing volumes
- Growing diversity
- Growing complexity

⇒ **Rich resource**

- How do we mine its riches for nuggets of information?
- Find & Access
- Understand
- Extract, Combine & Digest
- Test hypotheses
- Bingo!

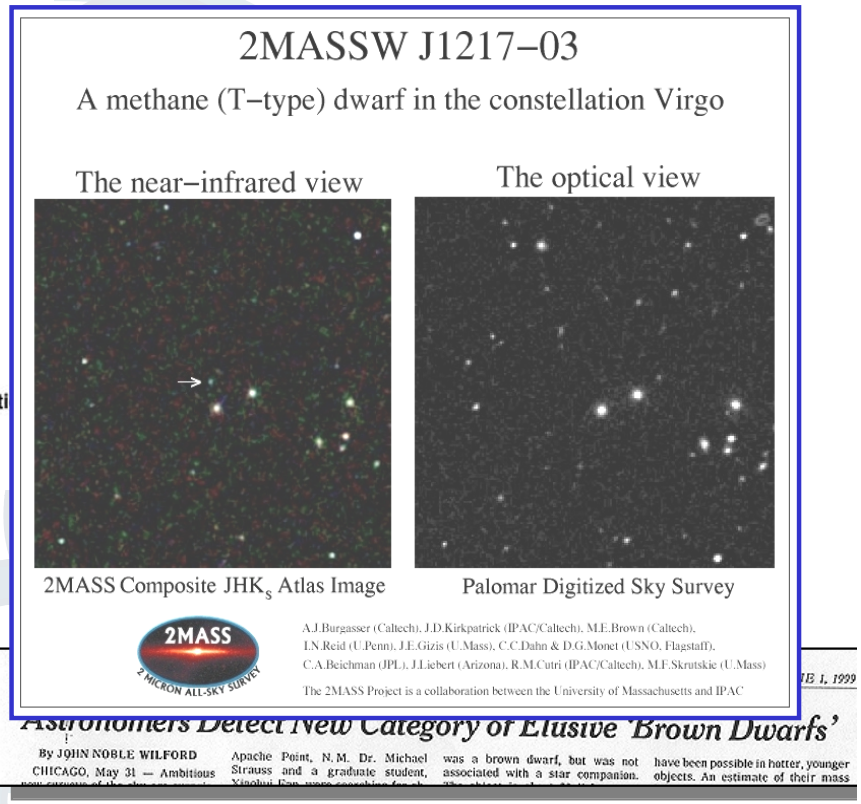
⇒ **Use OGSA-DAI**

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



Data and images courtesy Alex Szalay, John Hopkins

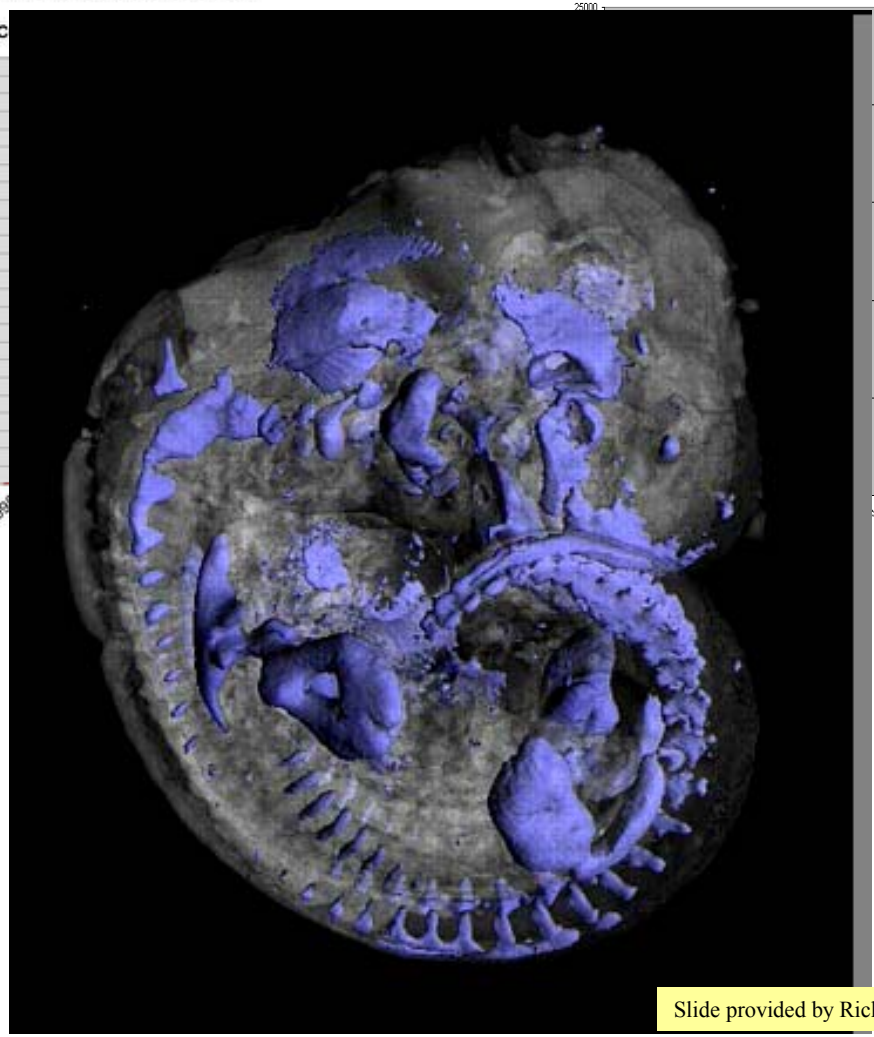
Biomedical data – making connections

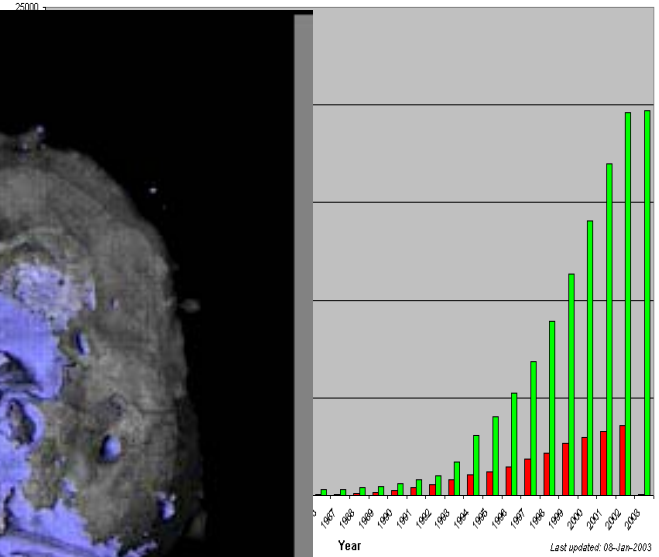
myGrid

Slide provided by Carole Goble: University of Manchester

ESNW e-Science North West

Database Growth

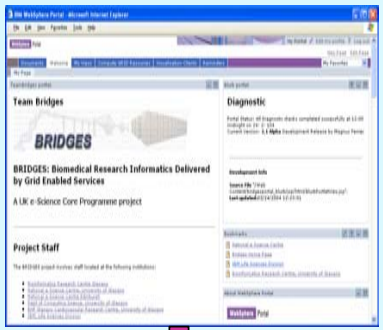




Slide provided by Richard Baldock: MRC HGU Edinburgh

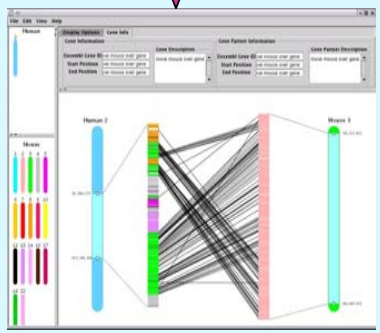
BRIDGES

VO Authorisation



Synteny Grid Service

blast



CFG Virtual Organisation

Information Integrator

DATA HUB

OGSA-DAI

Glasgow Private data

Edinburgh Private data

Oxford Private data

Leicester Private data

Netherlands Private data

London Private data

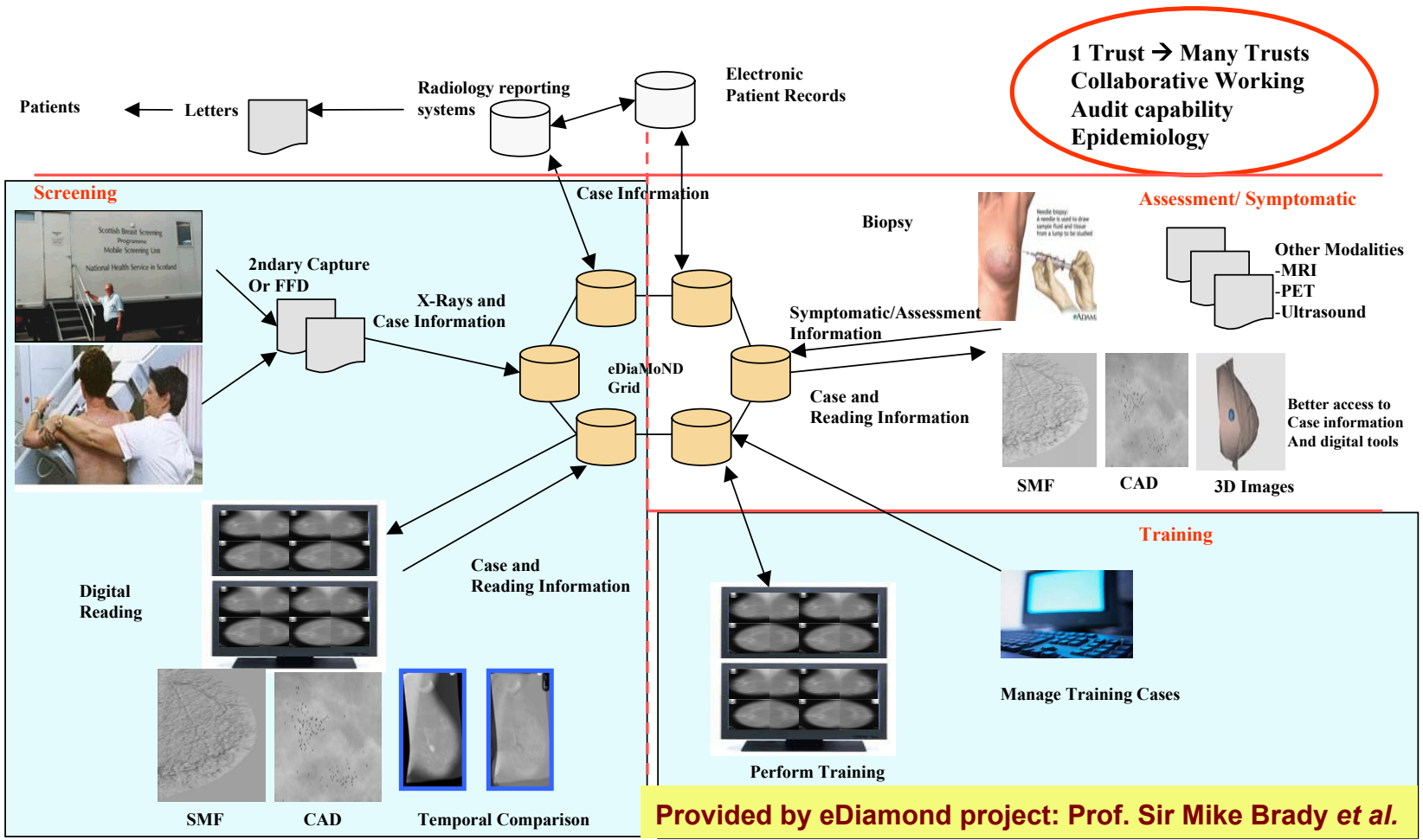
Publically Curated Data

Ensembl
OMIM
SWISS-PROT
MGI
HUGO
RGD
...

+

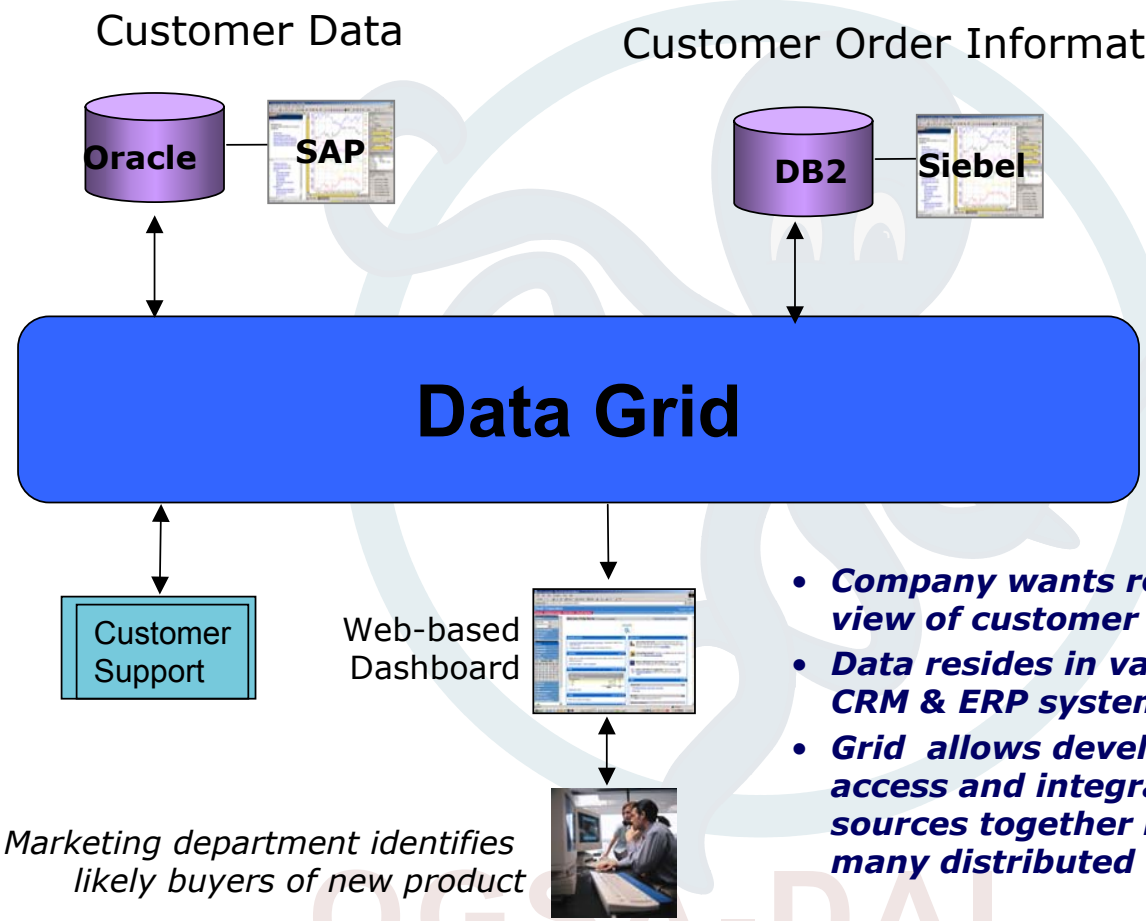
Slide provided by Richard Sinnott: University of Glasgow

eDiaMoND: Screening for Breast Cancer



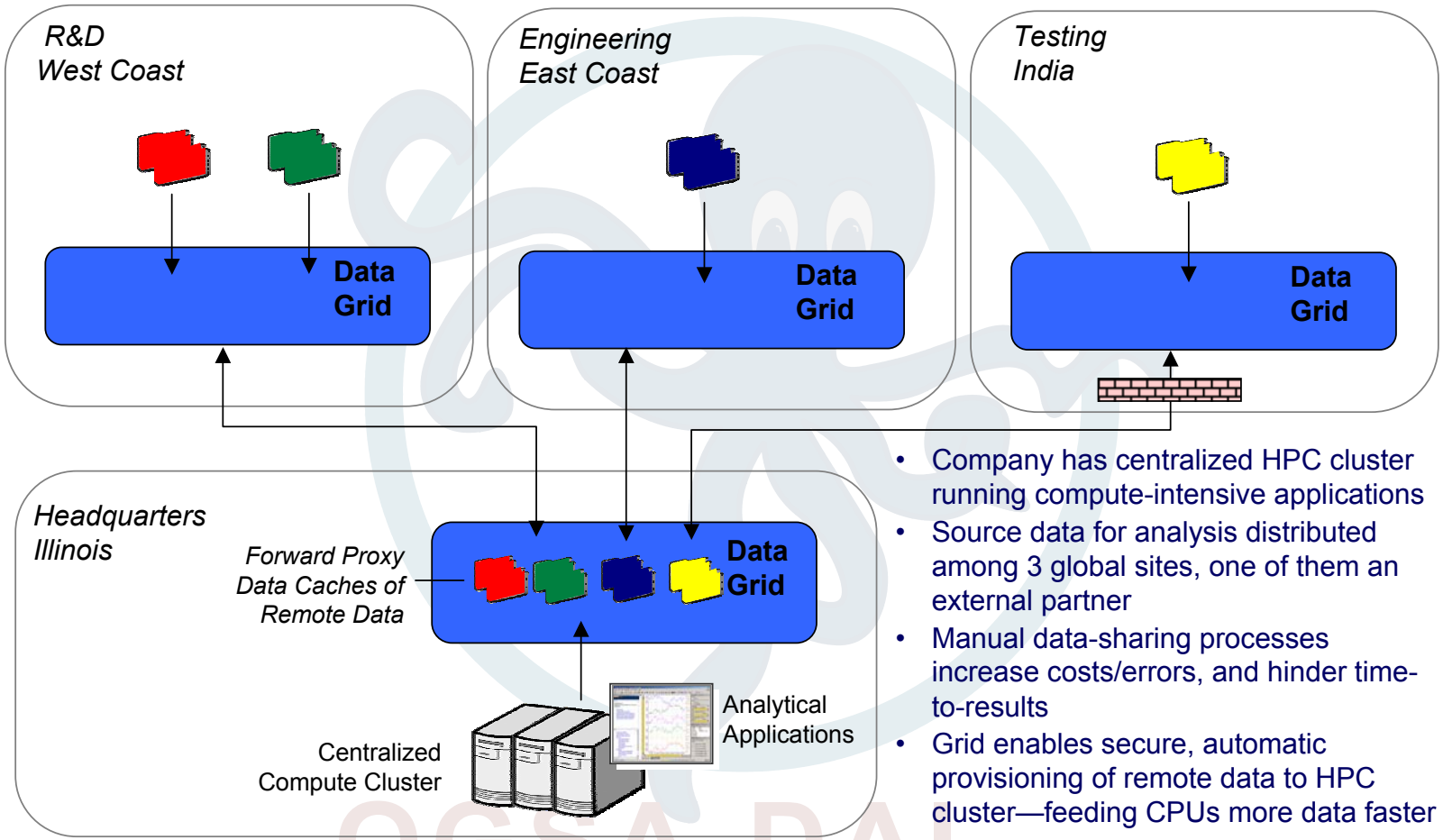
Provided by eDiamond project: Prof. Sir Mike Brady *et al.*

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***

Providing Data to Cluster-Based Analytical Application



- Company has centralized HPC cluster running compute-intensive applications
- Source data for analysis distributed among 3 global sites, one of them an external partner
- Manual data-sharing processes increase costs/errors, and hinder time-to-results
- Grid enables secure, automatic provisioning of remote data to HPC cluster—feeding CPUs more data faster

Story so far

- There is a lot of data
  - growing in every dimension
  - Distributed
  - Many different producers & owners
  - Heterogeneous
  - High value resource
  - Combined it is more valuable
- There are many requirements for [illegible]ntegration
  - Takes many forms
  - Driven by insights
  - Enable conversion of in[illegible]o tested hypothesis
- There are many d[illegible]ners
  - Their autonomy and policies must be respected

Generic Repeatable Solutions Required

Changing the way we manage Data

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

Malcolm Atkinson
Director, National e-Science Centre
mpa@nesc.ac.uk
+44 131 651 4040

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

- 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?
- Develop the storage architecture for this?
- Develop experiment, sensor & simulation architectures
  - That take code to select and digest data as an output control
  - That attach the provenance & metadata
- 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
  - Content
  - Provenance
  - Structure
  - Types, Formats & Ontologies
  - Operations available
  - Access requirements
  - Quality of service
- No established standards for heterogeneous data sources

Cultural Challenges

- Changing the way we work?
- Publication and sharing of result data
  - 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

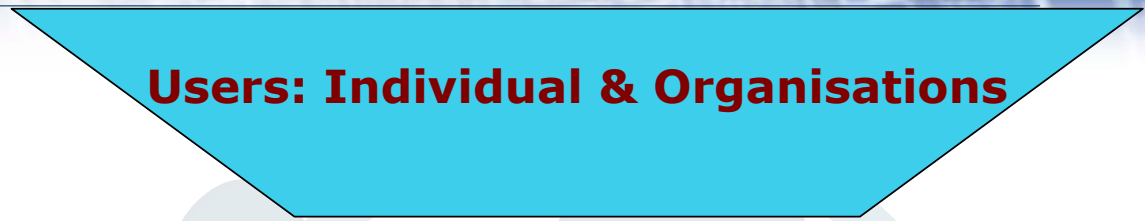
Economic Challenges

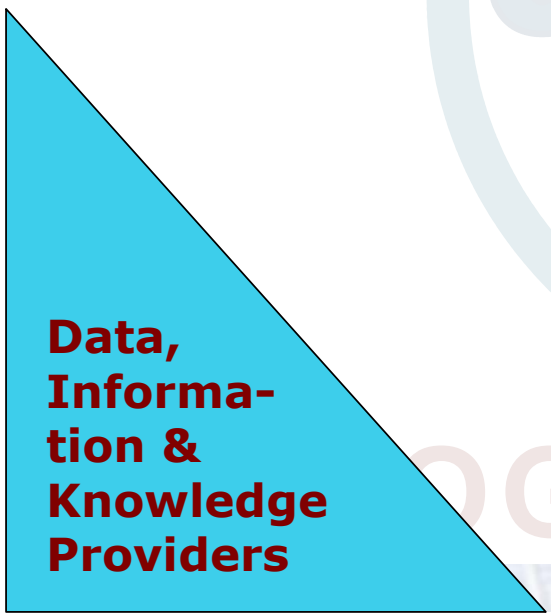
- Data production, publication & management
  - Many researchers contributing increments of data
  - Who pays for storage, transport, management and curation?
- Data longevity
  - Research requirements may outlive technical decisions
  - Data does not preserve itself indefinitely!
- When a community is dependent on a data resource
  - Who pays or decides to switch it off?

Summary: Scientific Discovery

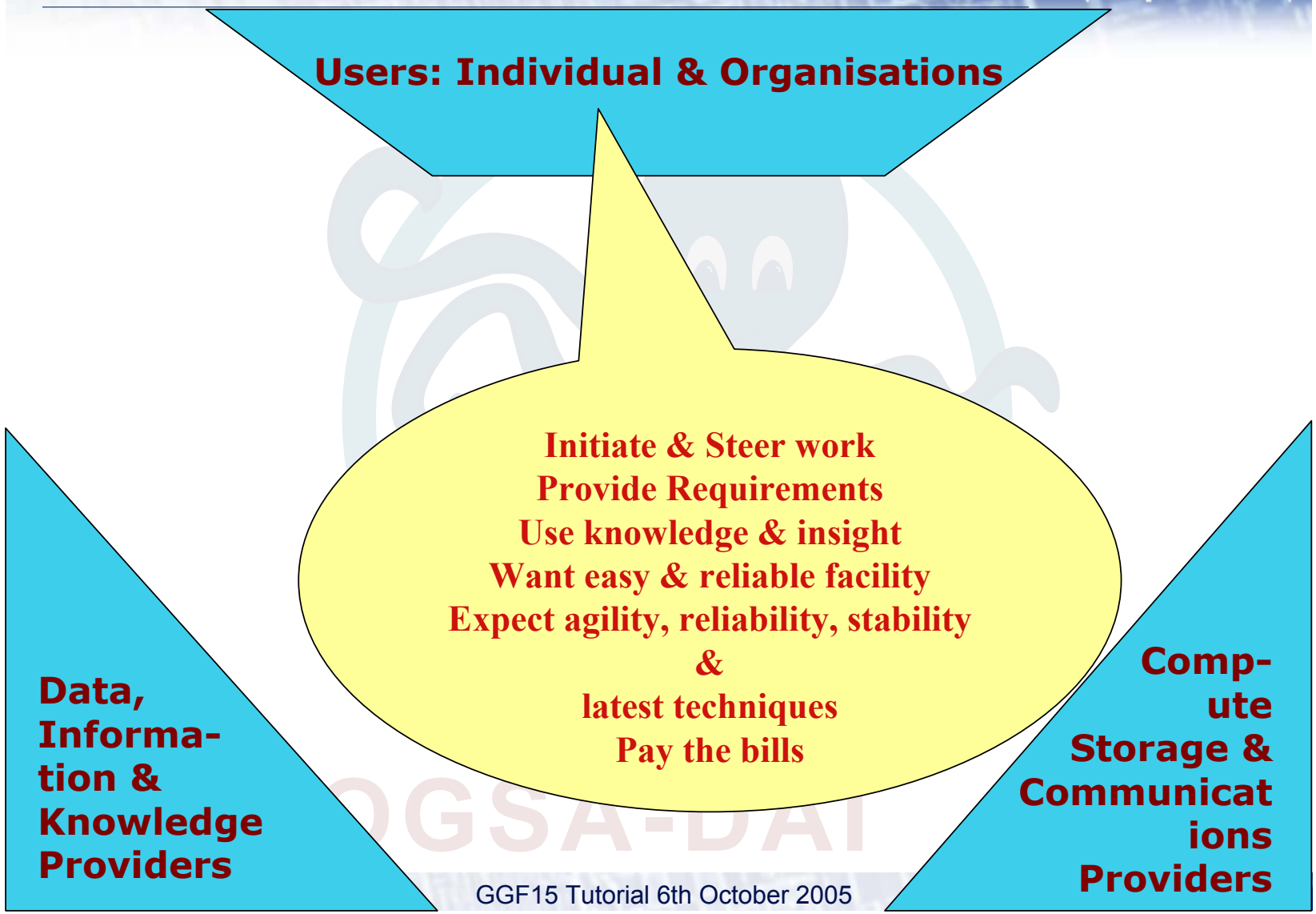
- Obtaining access to that data
  - Overcoming administrative barriers
  - Overcoming technical barriers
- Understanding that data
  - The parts you care about for your research
- Combing them using sophisticated models
  - The picture of reality in your head
- Analysis on scales required by statistics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

Three communities

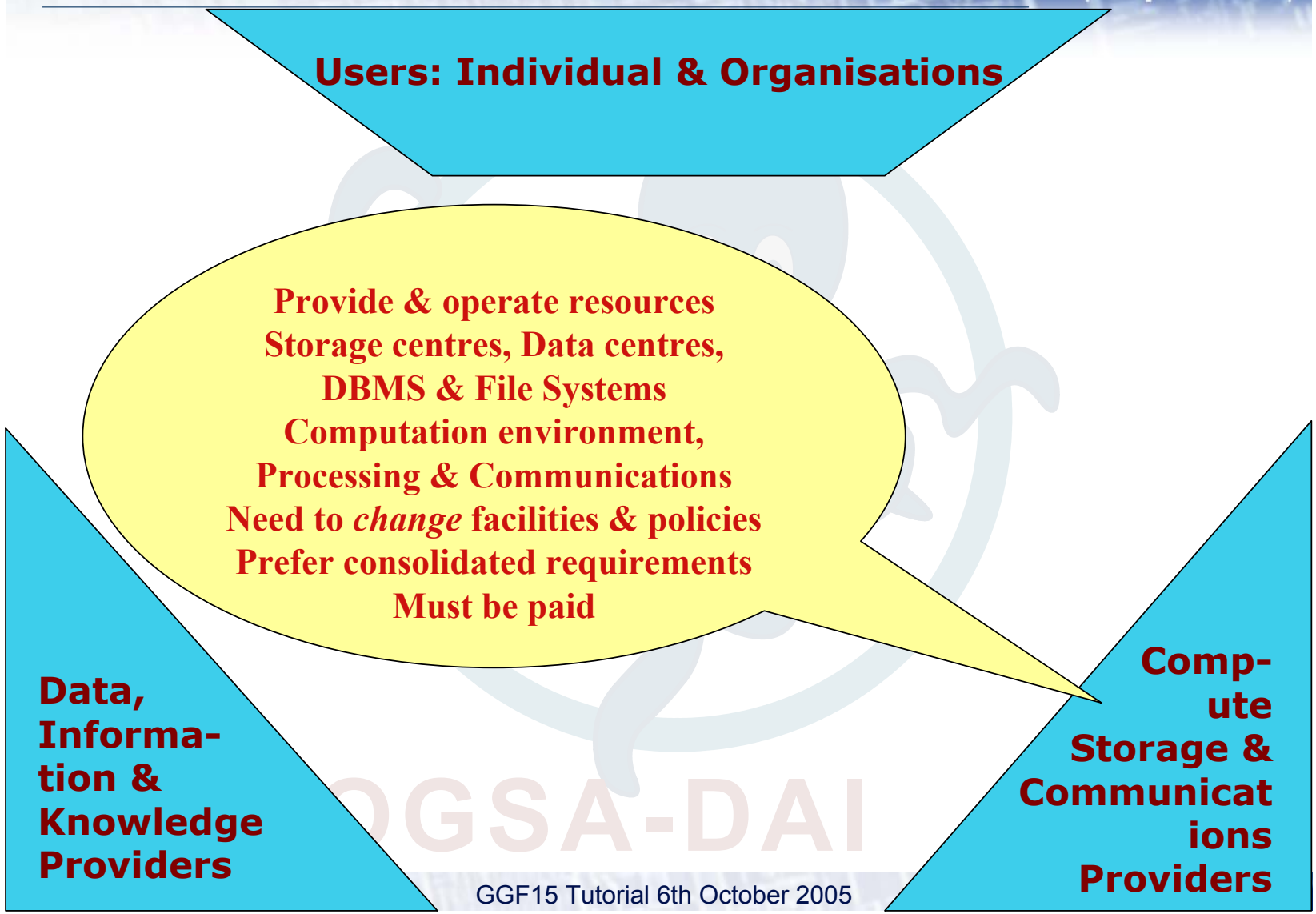




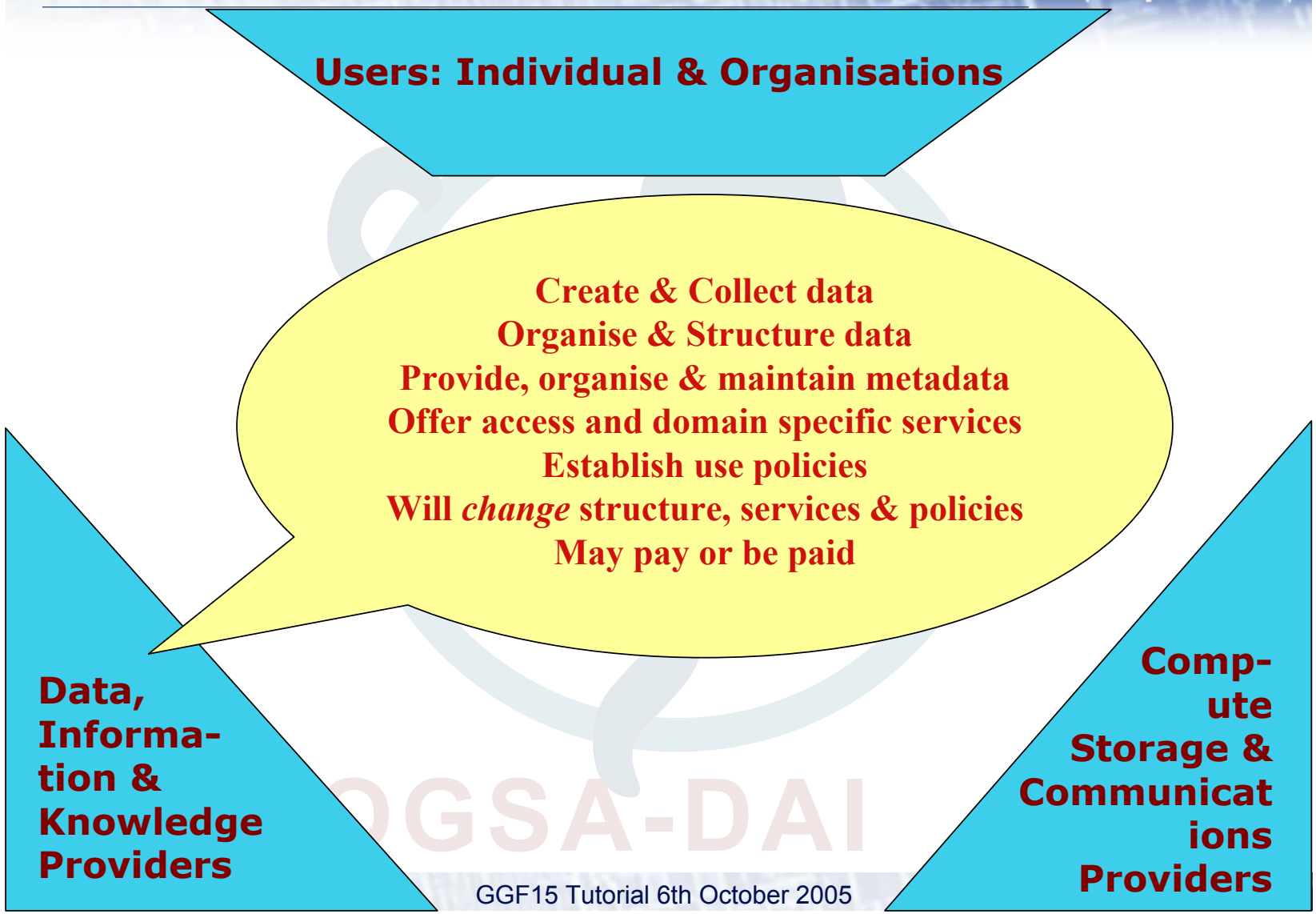
Three communities



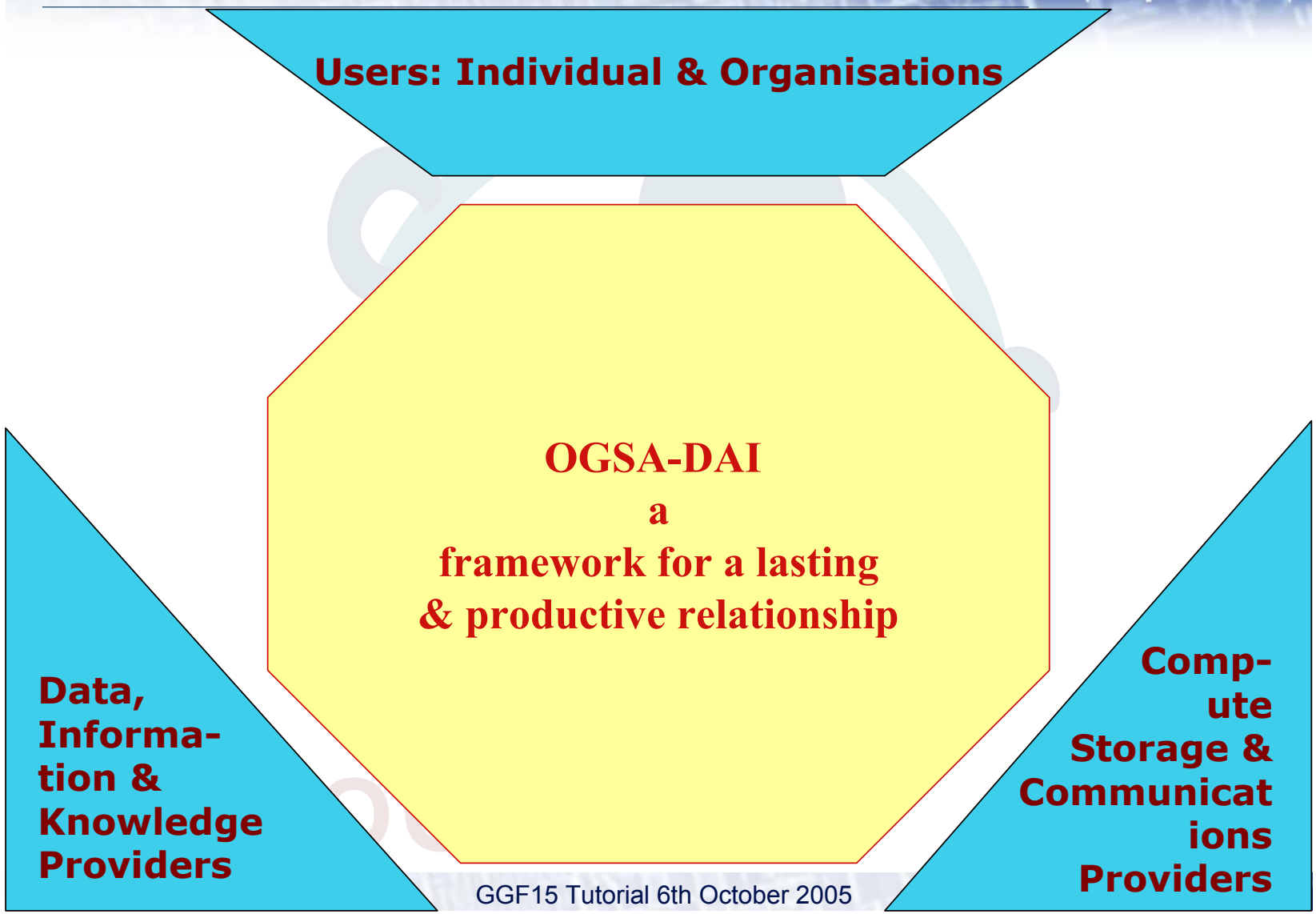
Three communities



Three communities



Three communities

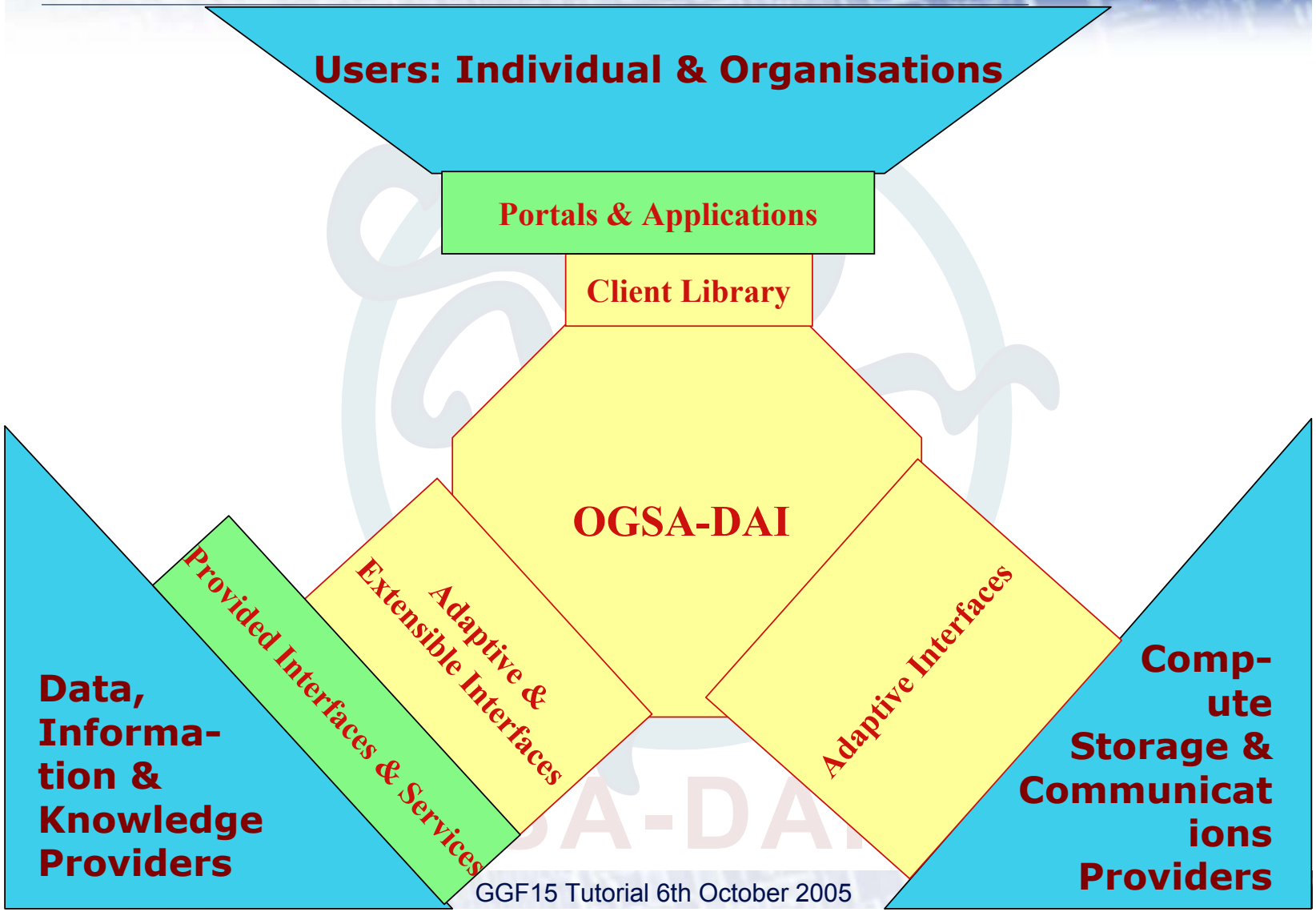


An Architecture

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

Malcolm Atkinson
Director, National e-Science Centre
mpa@nesc.ac.uk
+44 131 651 4040

Three communities



DAIS Working Group

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

Malcolm Atkinson
Director, National e-Science Centre
mpa@nesc.ac.uk
+44 131 651 4040

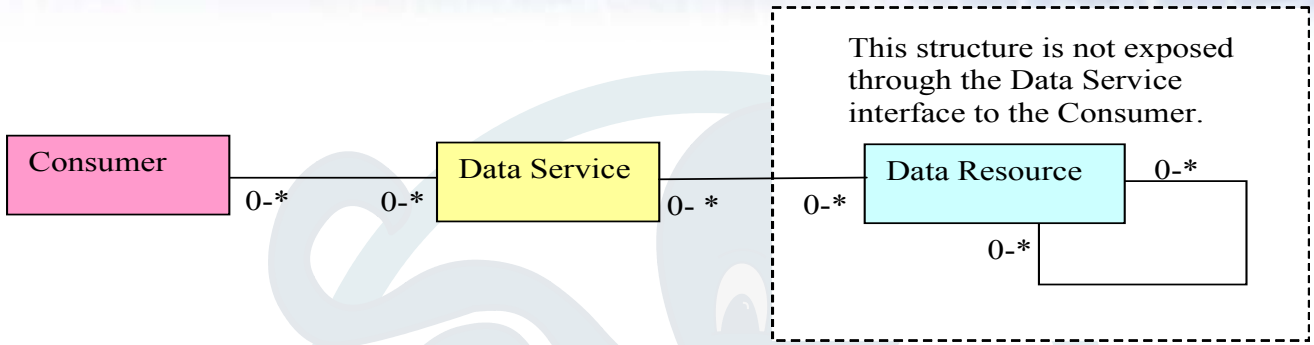
DAIS WG Goals

- 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

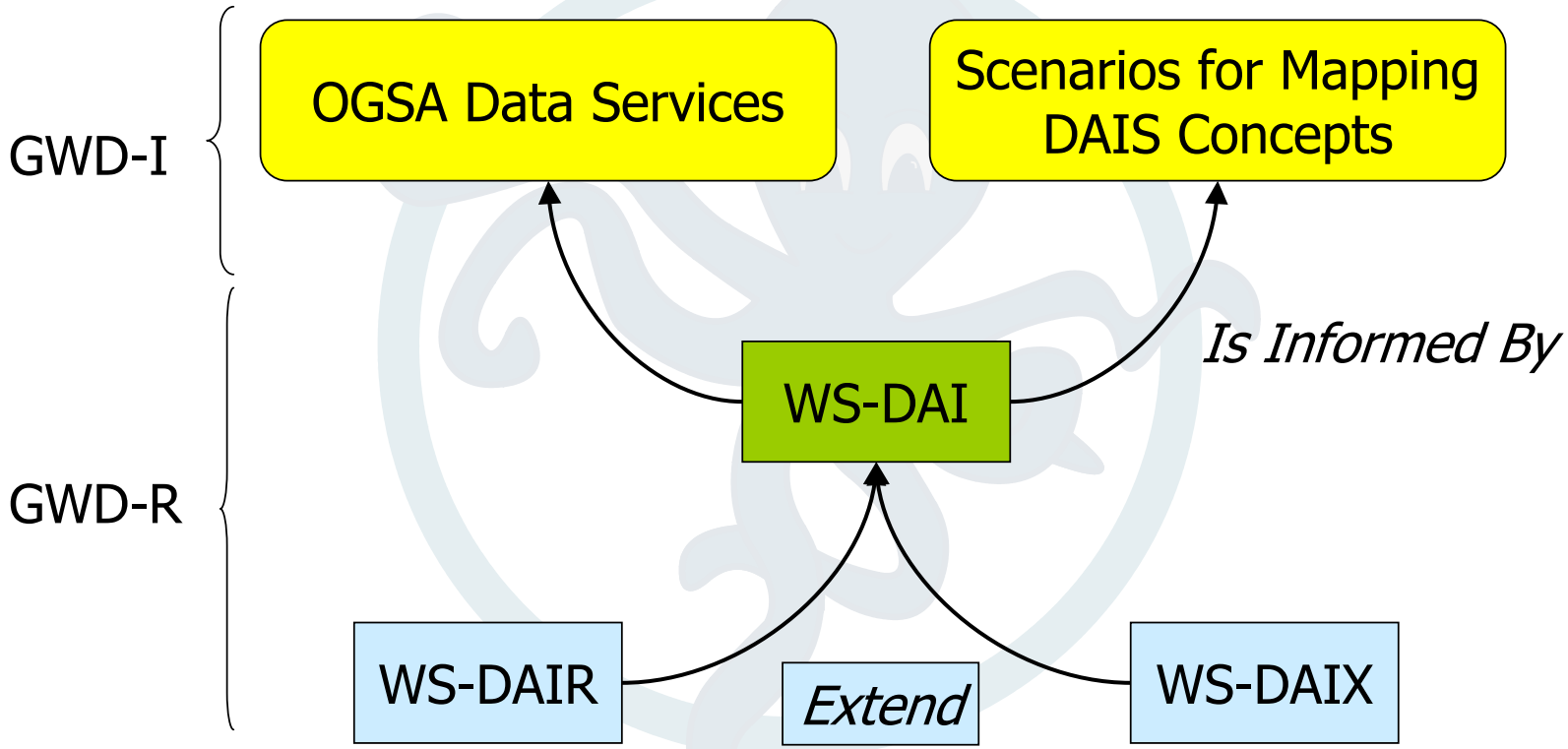
- No new common query language
- No schema integration or common data model
- No common namespace or naming scheme
- No data resource management
  - e.g. starting/stopping database managers
- No push based delivery
  - Information Dissemination WG?

DAIS View Of Data Services Model

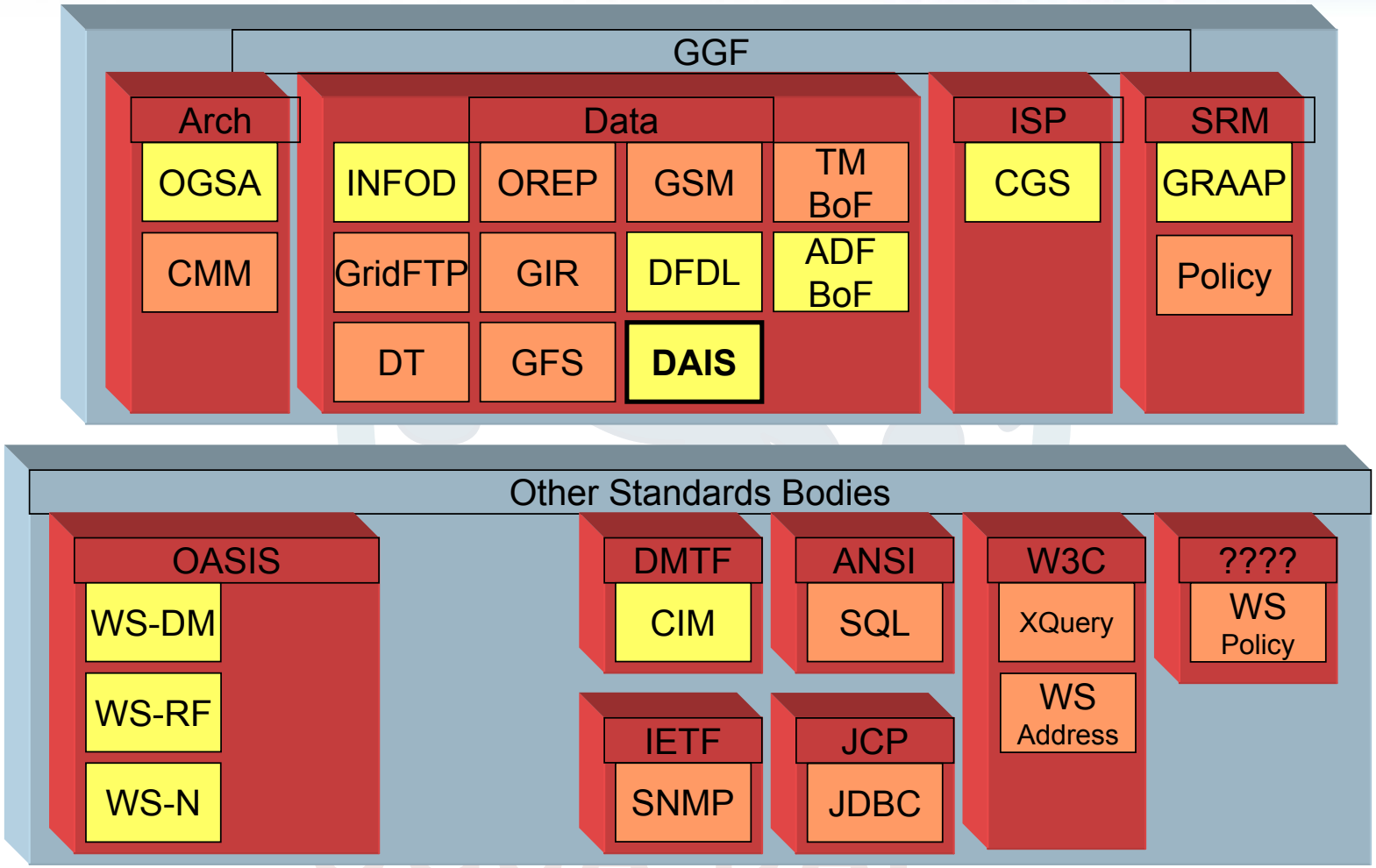


- A Data Service presents a Consumer with an interface to a Data Resource.
- A Data Resource can have arbitrary complexity, for example, a file on an NFS mounted file system or a federation of relational databases.
- A Consumer is not typically exposed to this complexity and operates within the bounds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

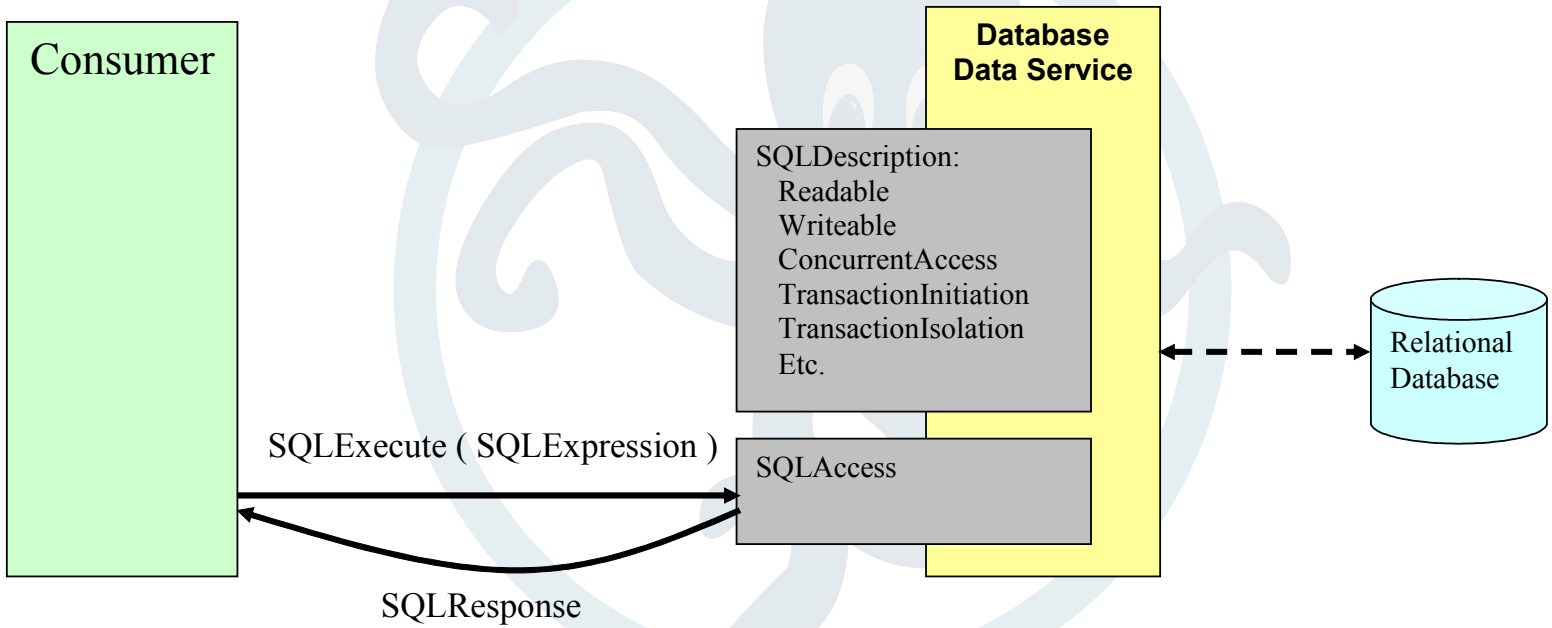
DAIS Specification Landscape



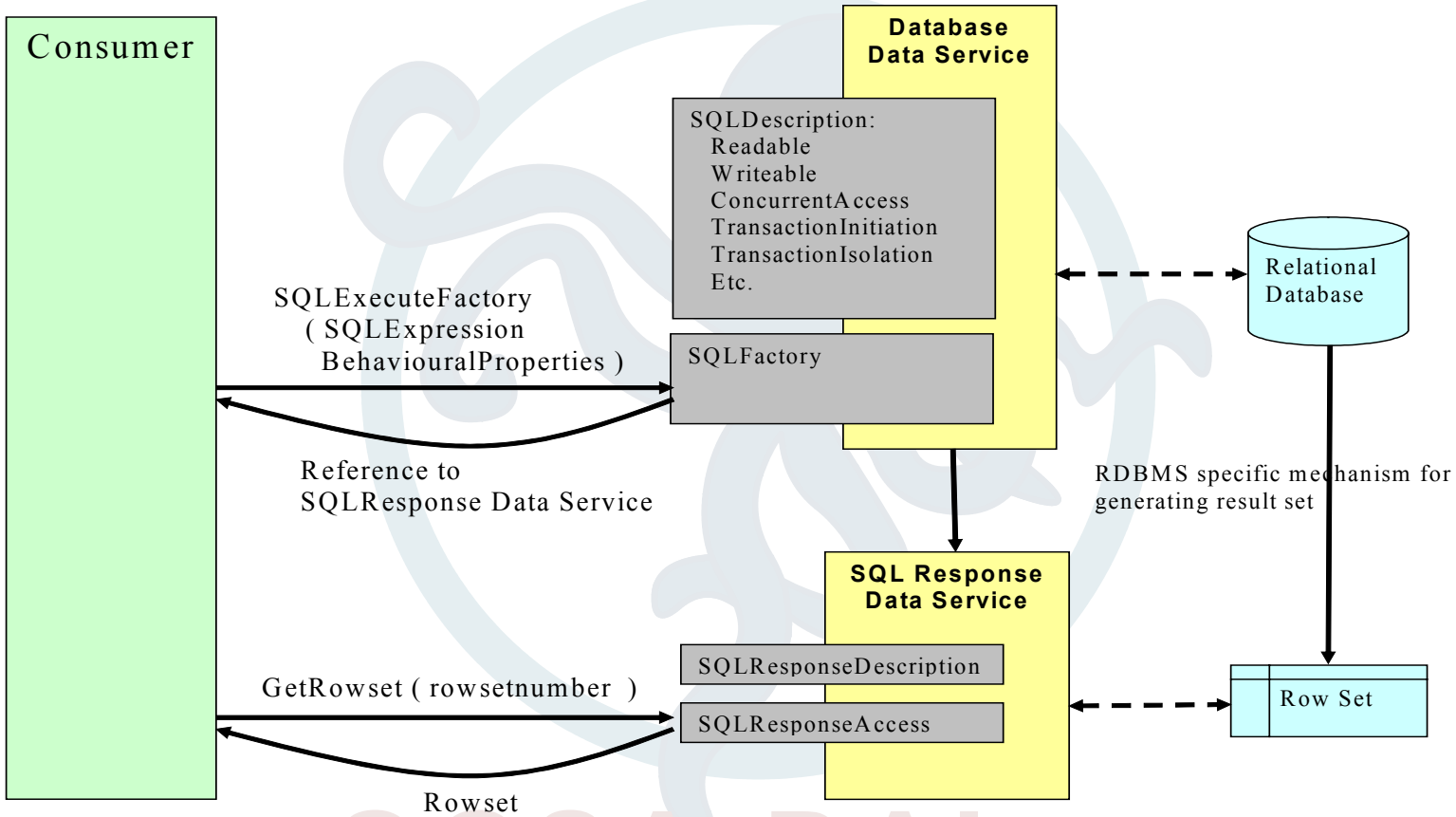
DAIS and Other Standards/Specifications



DAIS Data Access



DAIS Derived Data Access



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

And now OGSA-DAI

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

Malcolm Atkinson
Director, National e-Science Centre
mpa@nesc.ac.uk
+44 131 651 4040

OGSA-DAI Projects

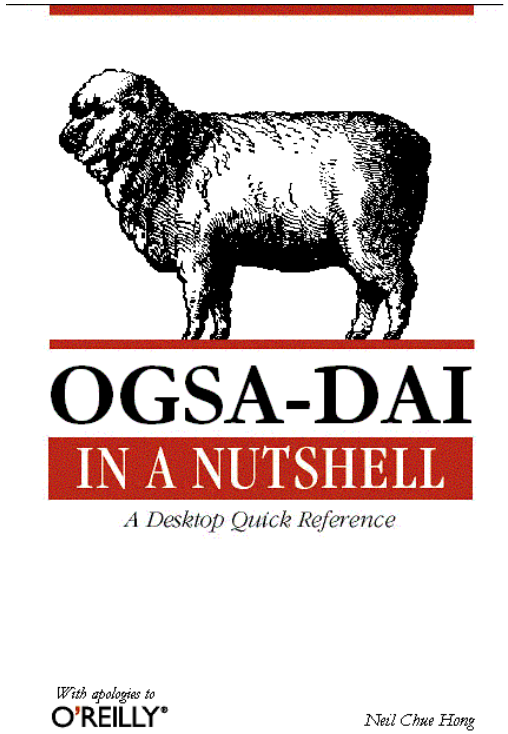
- Develop a component library
  - Access and manipulate data in a grid
  - Serve UK and International e-Science communities
- Provide an *extensible* framework
- Provide
  - Common interface to data resources
  - Simple integration of distributed queries to multiple data resources
- Contribute to standardisation efforts
  - Input into GGF DAIS WG and other groups
  - Provide a reference implementation of DAIS spec
- Based on Open Grid Services Architecture (OGSA)
  - WSRF (GT4) & WS-I (OMII_2) versions
- Support Application Developers & Contributors

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!

Slide from Neil Chue Hong

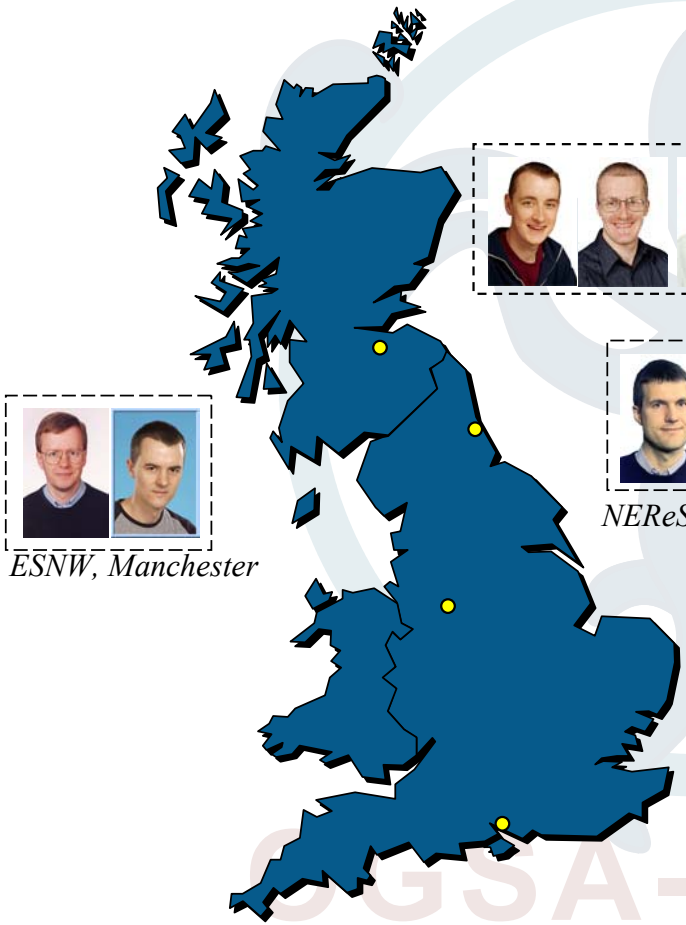
OGSA-DAI In One Slide



- 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 you project using
  - Additional Activities
  - Client Toolkit APIs
  - Data Resource handlers
- A base for higher-level services
  - federation, mining, visualisation,…

Slide from Neil Chue Hong

OGSA-DAI team



EPCC Team, Edinburgh

NeSC, Edinburgh

NEReSC, Newcastle

ESNW, Manchester

IBM Development Team, Hursley

IBM Dissemination Team

International Collaboration & Use

USA:
- o Globus Alliance
- o IBM Corporation
- o caBIG
- o BIRN
- o Indiana University
- o GridSphere
- o GEON
- o LEAD
- o MCS
- o NCSA
- o Secure Data Grid
- o UNC

UK:
- o OMII
- o OMII-UK
- o NGS
- o NCeSS
- o NIeeS
- o AstroGrid
- o BioSimGrid
- o BRIDGES
- o CancerGrid
- o ConvertGrid
- o eDiaMonD
- o EDINA
- o First Group plc
- o Fujitsu Labs Europe
- o GEDDM
- o GeneGrid
- o Genomic Technology and Informatics
- o GOLD
- o Human Genetics Unit
- o IBM UK
- o myGrid
- o Oracle UK

Europe:
- o CERN
- o DataMiningGrid
- o GridMiner
- o GridSphere
- o inteligrid
- o N2Grid
- o OntoGrid
- o Provenance
- o SIMDAT
- o OMII-EU

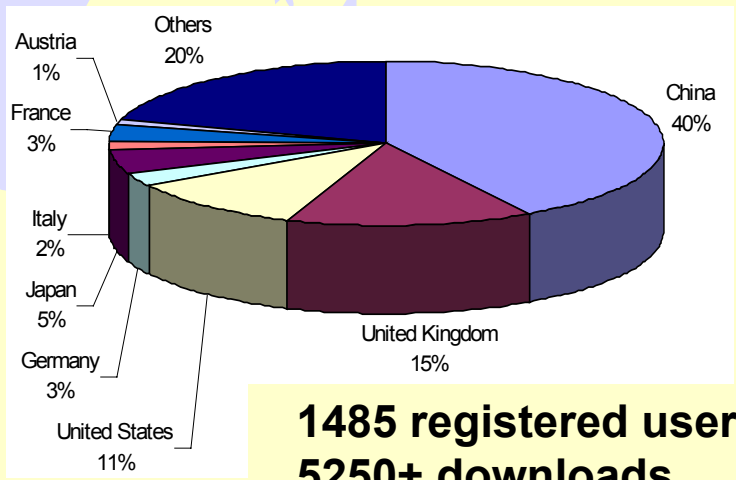
China:
- o CAS
- o ChinaGrid
- o cnGrid
- o INWA
- o OMII-China

Japan:
- o AIST
- o BioGrid
- o NAREGI

South Korea:
- o KISTI

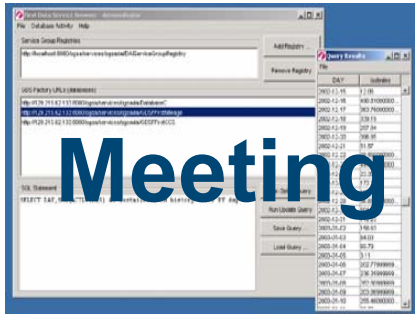
Tutorials

Boston	Cambridge
CERN	Chicago
Edinburgh	London
San Francisco	Seattle
Seoul	Singapore
Tokyo	ISSGC 03 to 05

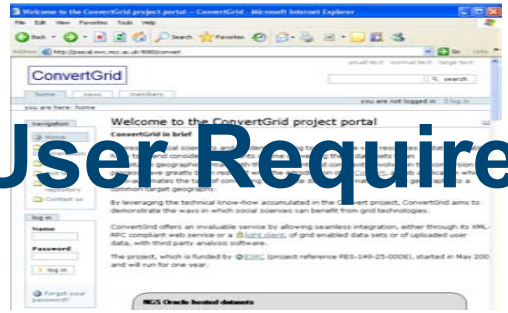


1485 registered users
5250+ downloads

Meeting User Requirements

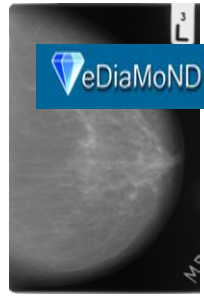


FirstDIG



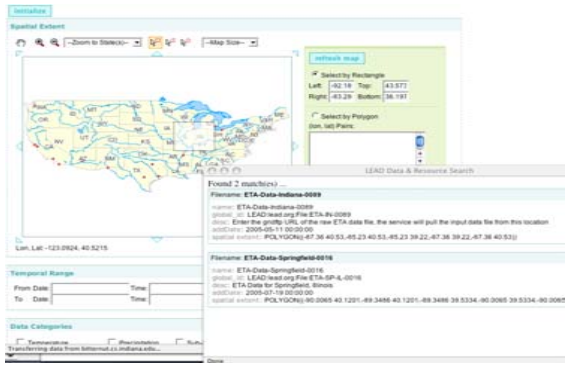
ConvertGrid

GeneGrid

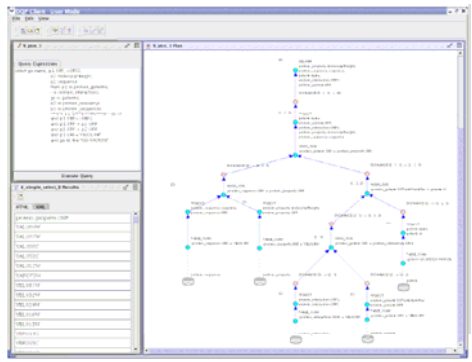


eDiaMoND

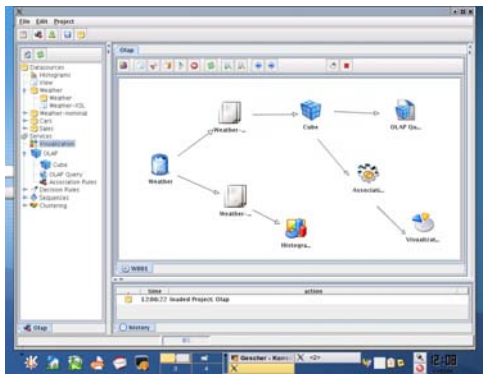
BRIDGES



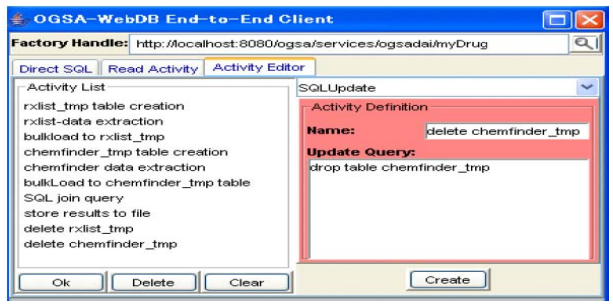
LEAD



OGSA-DQP



Grid Miner



OGSA WebDB

caBIG

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)

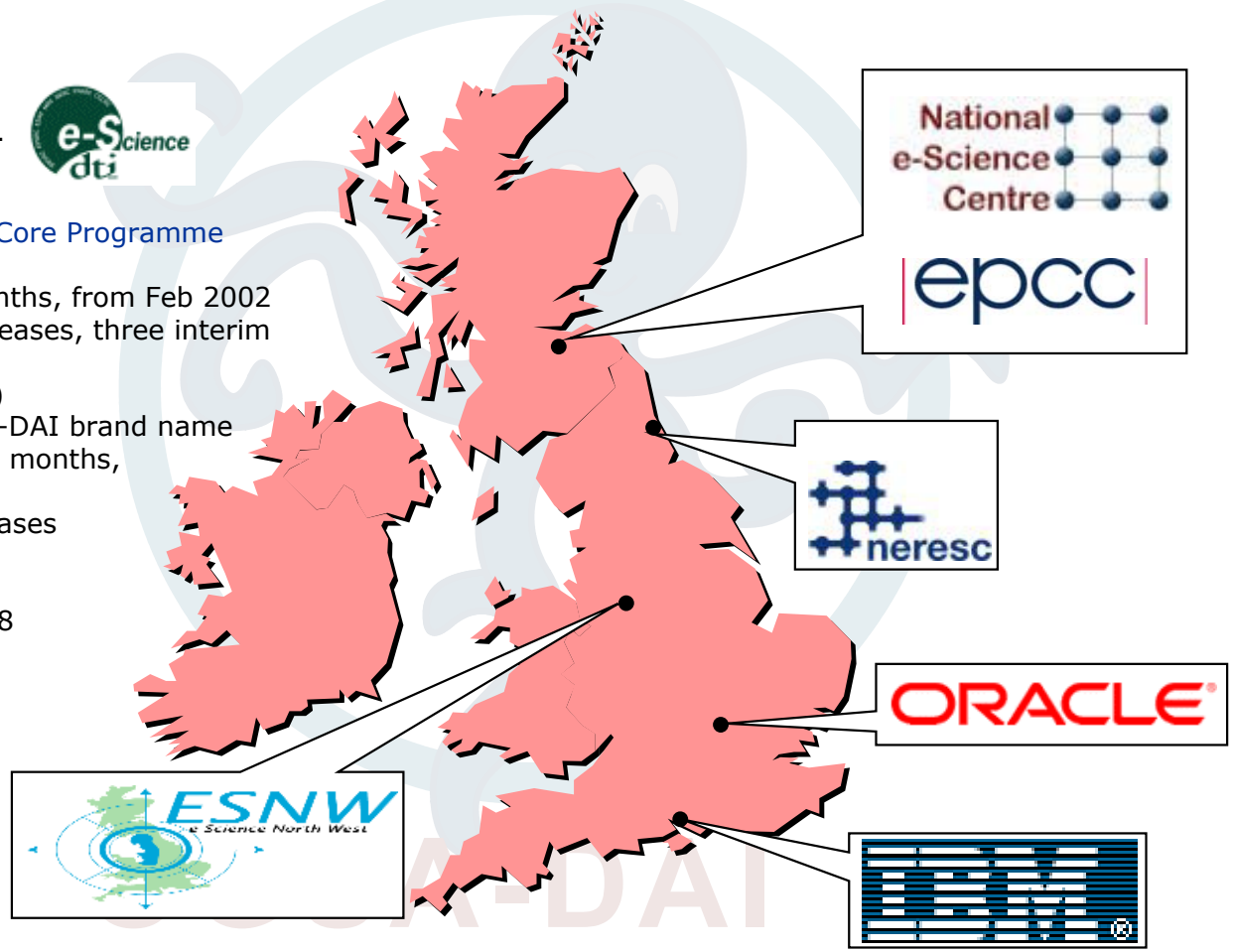
Keep the OGSA-DAI brand name

£1.5 million, 24 months, from Oct 2003

Four major releases

OMII-UK

To October 2008



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

Malcolm Atkinson
Director, National e-Science Centre
mpa@nesc.ac.uk
+44 131 651 4040