



*International Summer School on Grid Computing
Vico Equense, 16th July 2005*

Enabling Grids for
E-science in Europe

www.eu-egee.org

Today's Wealth of Data: Are we ready for its challenges?

Malcolm Atkinson
Director
National e-Science Centre
www.nesc.ac.uk



What is e-Science?


- **Goal: to enable better research**
- **Method: Invention and exploitation of advanced computational methods**
 - to generate, curate and analyse research data
 - From experiments, observations and simulations
 - Quality management, preservation and reliable evidence
 - to develop and explore models and simulations
 - Computation and data at extreme scales
 - Trustworthy, economic, timely and relevant results
 - to enable *dynamic* distributed virtual organisations
 - Facilitating collaboration with information and resource sharing
 - Security, reliability, accountability, manageability and *agility*

Multiple, independently managed sources of data – each with own time-varying structure

Creative researchers discover new knowledge by combining data from multiple sources

DAI: What is needed?

Data Access and Integration: 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 independently collected
 - **Key to Large-scale Collaboration and Decisions!**
 - **Economies: data production, publication & management**
 - Sharing cost of storage, management and curation
 - Many researchers contributing increments of data
 - Pooling annotation \equiv rapid incremental publication
 - And criticism
 - **Accommodates global distribution**
 - Data & code travel faster and more cheaply
 - **Accommodates temporal distribution**
 - Researchers assemble data
 - Later (other) researchers access data
- 

Data Access and Integration: challenges

Petabyte of Digital
Data / Hospital / Year

- **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

Data Access and Integration: Scientific discovery

- **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**
 - The parts *you* care about for *your* research
- **Extracting nuggets from multiple sources**
 - Pieces of *your* jigsaw puzzle
- **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

You're an innovator

∴ Your model ≠ their model

⇒ Negotiation & patience
needed from *both* sides

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
- **Distributed & diverse data collections**
 - Discovery depends on insights
 - ⇒ Unpredictable sophisticated application code
 - Tested by combining data from many sources
 - Using *novel* sophisticated models & algorithms
- **What can you do?**

Scientific Data: Opportunities and Challenges

• Opportunities

- Global Production of *Published Data*
- Volume↑ Diversity↑
- Combination ⇒ Analysis ⇒ Discovery

• Opportunities

- Spreading
- New Data Organisation
- New Algorithms
- Varied Replication
- Shared Annotation
- Intensive Data & Computation

• Challenges

- Data Huggers
- Meagre metadata
- Ease of Use

A Cornucopia of Research Challenges

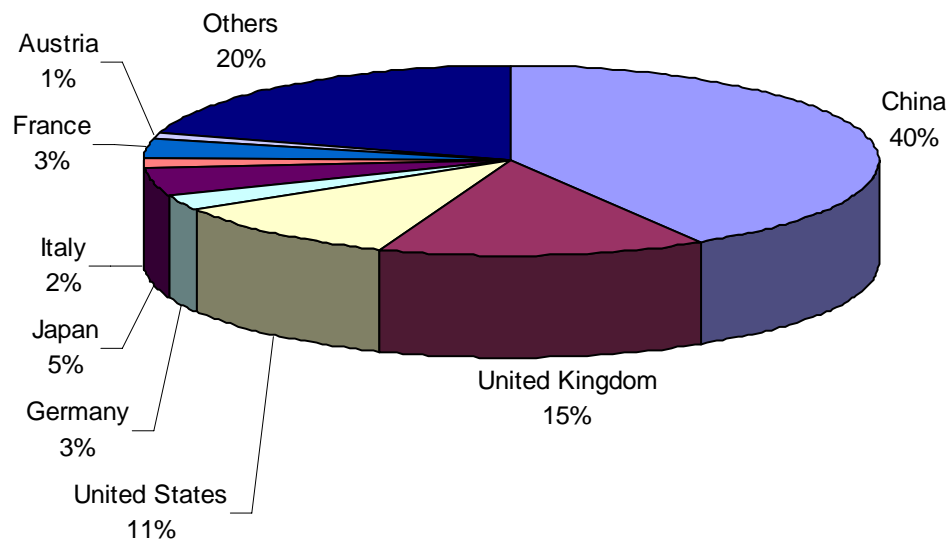
- Fundamental Principles
- Approximate Matching
- Multi-scale optimisation
- Autonomous Change
- Legacy structures
- Scale and Longevity
- Privacy and Mobility
- Sustained Support / Funding

The Story so Far

- **Technology enables Grids, 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 unpredictable (autonomous) development of both

Structure enables understanding, operations, management and interpretation

OGSA-DAI R5.0 downloads



Significant interest from China
 - two members of staff starting May 22nd

790 downloads since Dec 04
 -Actual user downloads not search engine crawlers
 -Does not include downloads as part of GT3.2 and GT4 releases

Total of 1212 registered users

R1.0 (Jan 03)	109
R1.5 (Feb 03)	110
R2.0 (Apr 03)	255
R2.5 (Jun 03)	294
R3.0 (Jul 03)	792
R3.1 (Feb 04)	686
R4.0 (May 04)	1083
Total	4119

at 17/5/2005

Goals for OGSA-DAI

- Aim to deliver application mechanisms that:
 - Meet the data requirements of Grid applications
 - Functionality, performance and reliability
 - Reduce development cost of data-centric Grid applications
 - Provide consistent interfaces to data resources
 - Acceptable and supportable by database providers
- A base for developing higher-level services
 - Data federation
 - Distributed query processing
 - Data mining
 - Data visualisation

Core features of OGSA-DAI

- A framework for building applications
 - Supports relational, xml and some files
 - MySQL, Oracle, DB2, SQL Server, Postgres, XIndice, CSV, EMBL
 - Supports various delivery options
 - SOAP, FTP, GridFTP, HTTP, files, email, inter-service
 - Supports various transforms
 - XSLT, ZIP, GZip
 - Supports message level security using X509
 - Client Toolkit library for application developers
 - Comprehensive documentation and tutorials
- Highly extensible
 - Strength is in customising out-of-box features

OGSA-DAI Design Principles - I

- Efficient client-server communication
 - Minimise number of messages exchanged
 - One request abstracts multiple interactions
- No unnecessary data movement
 - Move computation to the data
 - Utilise third-party delivery
 - Apply transforms (e.g., compression)
- Build on existing standards
 - Filling-in gaps where necessary

OGSA-DAI Design Principles -II

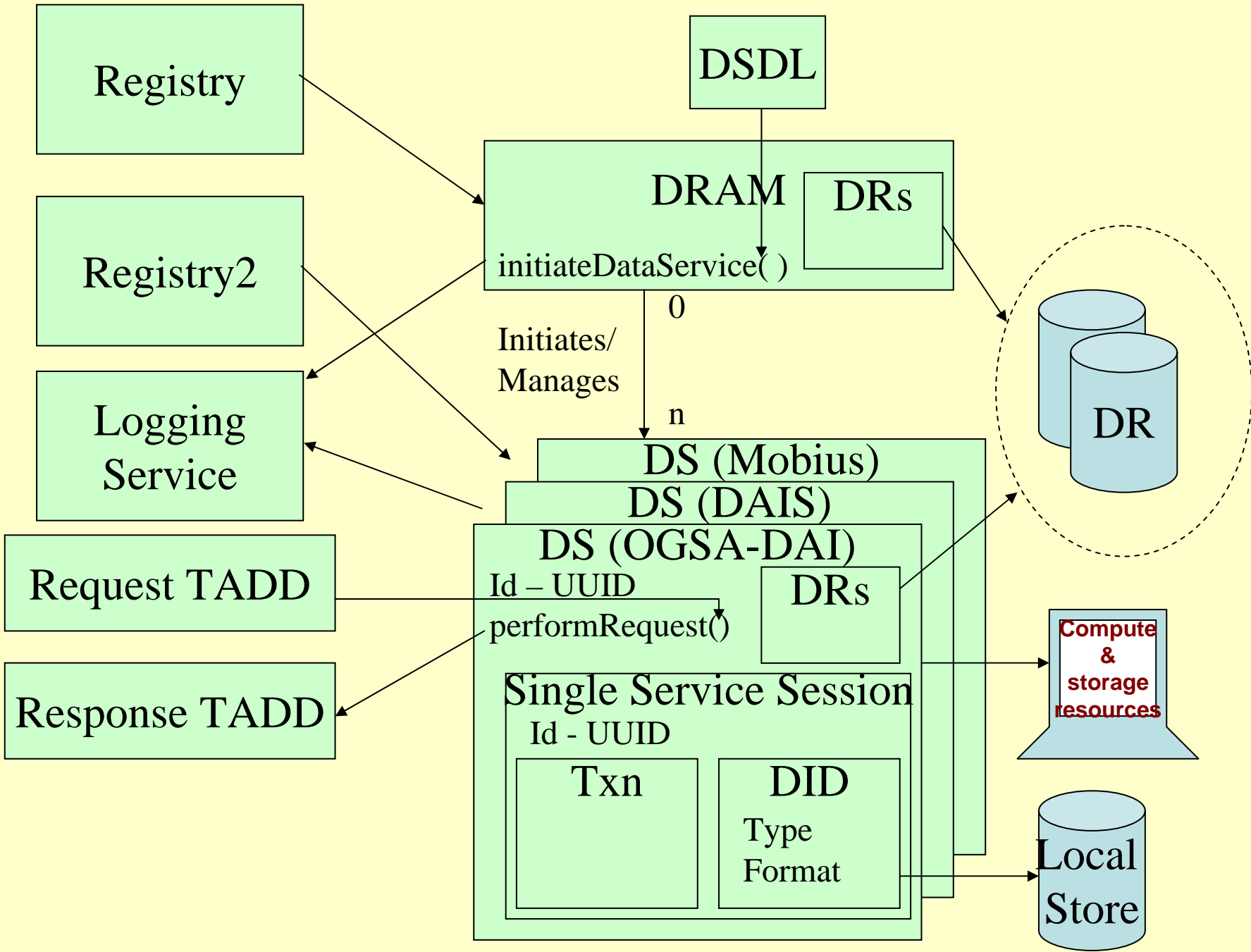
- Do not virtualise underlying data model
 - Users must know where to target queries
- Extensible architecture
 - Modular and customisable
 - E.g., to accommodate stronger security
- Extensible activity framework
 - Cannot anticipate all desired functionality
 - Activity = unit of functionality
 - Allow users to add their own

Why Use OGSA-DAI

- Provides common access view
 - Regardless of underlying infrastructure
 - “Everything looks like a database” metaphor
 - Access mechanism common to all clients
- Integrates well with other Grid software
 - OGSA, WSRF and OMII compliant
- Flexibility
 - Extensible activity framework
 - Won't tie you to storage infrastructure

Why You Might Not Want To Use OGSA-DAI

- OGSA-DAI slower than direct connection methods
 - E.g., compared to JDBC
 - This should improve with time
- Scalability issues
 - Mostly but not completely known
 - Depend on type of use (e.g. delivery mechanism)
- Only planning to use one type of data resource
 - and don't care about interoperability with other Grid software
 - OGSA-DAI an overkill in that case



Good Bye

- Thank you for coming to ISSGC'05
- Tell your friends to come next year