the globus alliance
www.globus.org

National e-Science Centre

Data, Data everywhere with …

French N+N meeting, DTI, London

Prof. Malcolm Atkinson
Director

www.nesc.ac.uk

www.ogsadai.org.uk

3rd November 2003

epcc

e-Science dti

UNIVERSITY of GLASGOW

the globus alliance
www.globus.org

Contents

National e-Science Centre

- **Data:**
  The Lingua Franca of e-Science
- **Data:**
  The Challenge for e-Science
- **OGSA-DAI *Product*:**
  The First Steps in DAI
  - **An opportunity for collaboration**
- **OGSA-DAI *Product*:**
  The Next Steps
  - **More collaboration please**

ORACLE

IBM

National e-Science Centre

north east e-science centre

ESNW e Science North West

epcc

Three-way Alliance

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

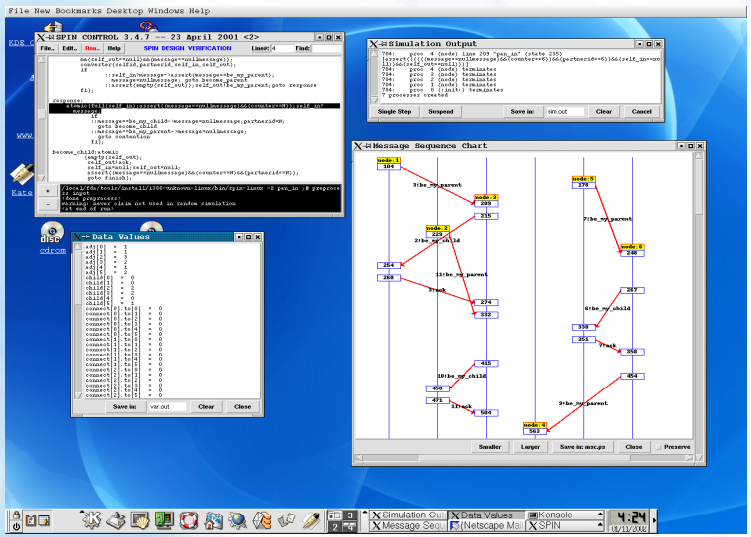
Requires Much Engineering, Much Innovation

Changes Culture, New Mores, New Behaviours

New Opportunities, New Results, New Rewards

Biochemical Pathway Simulator

(Computing Science, Bioinformatics, Beatson Cancer Research Labs)



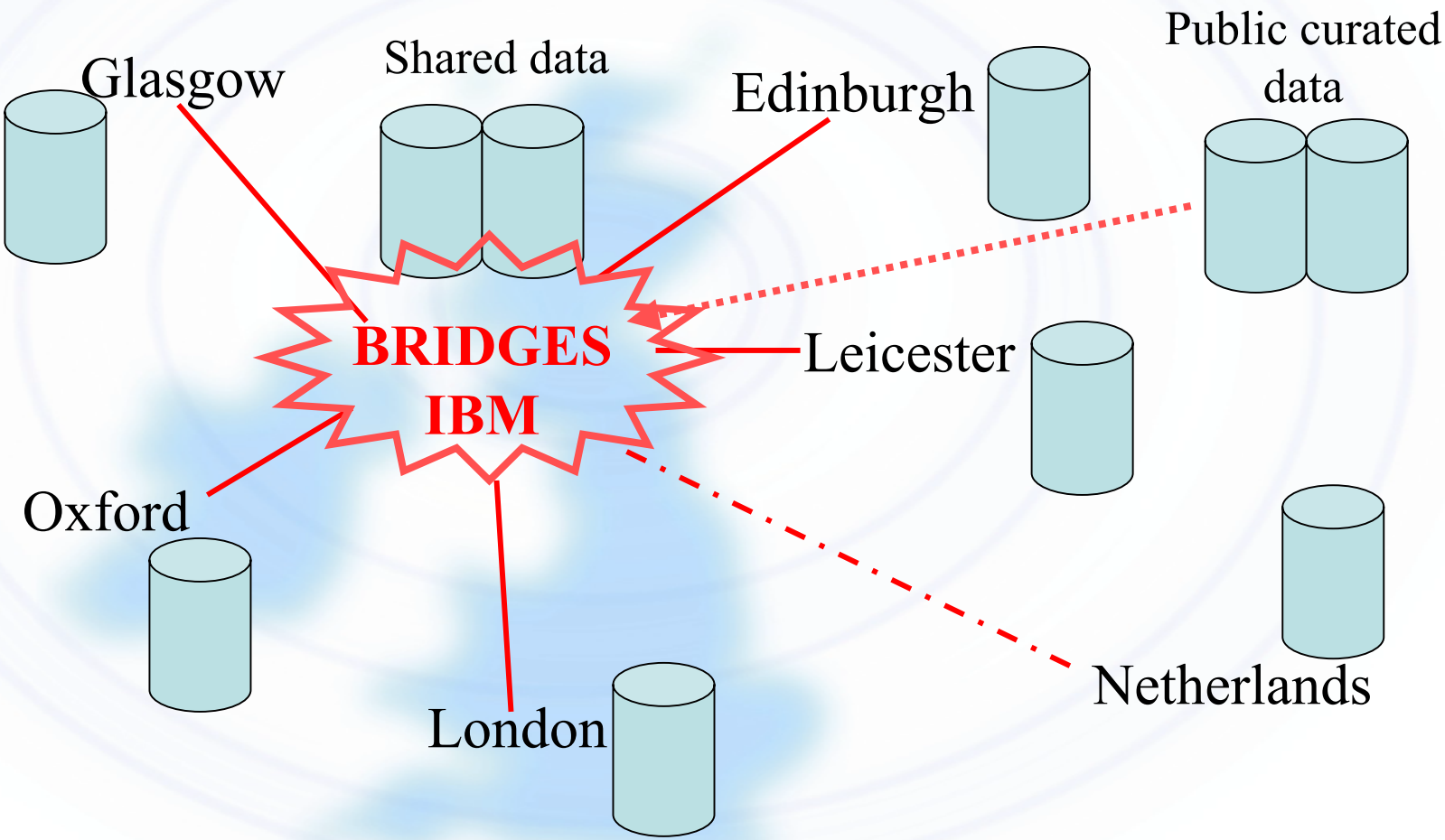
Closing the information loop - between lab and computational model.

DTI Bioscience *Beacon Project*

Harnessing Genomics Programme

Slide from Professor Muffy Calder, Glasgow

Wellcome Trust: Cardiovascular Functional Genomics

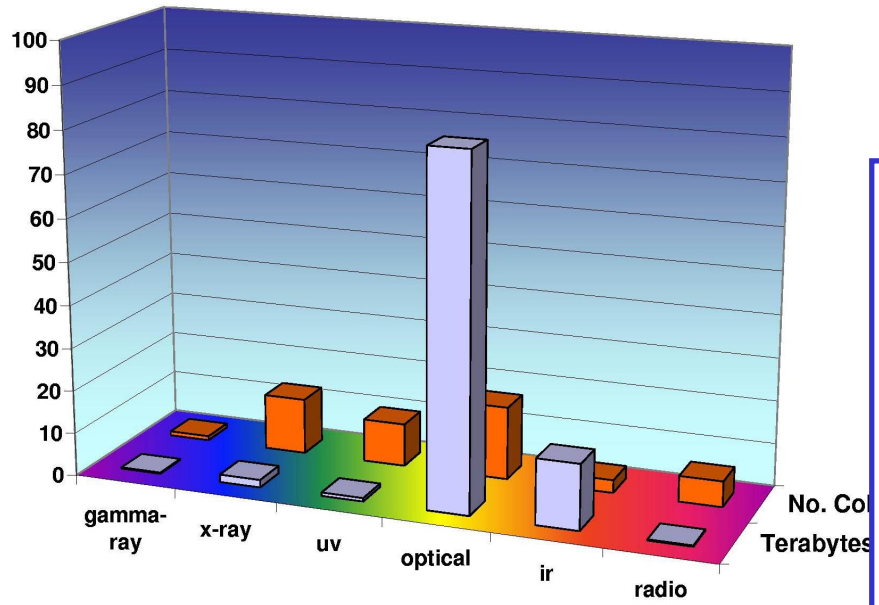


It's Easy to Forget How Different 2003 is From 1993

- Enormous quantities of data: Petabytes
  - For an increasing number of communities
  - gating step is not collection but analysis
- Ubiquitous Internet: >100 million hosts
  - Collaboration & resource sharing the norm
  - Security and Trust are crucial issues
- Ultra-high-speed networks: >10 Gb/s
  - Global optical networks
  - Bottlenecks: last kilometre & firewalls
- Huge quantities of computing: >100 Top/s
  - Moore's law gives us all supercomputers
  - Organising their effective use is the challenge
- Moore's law everywhere
  - Instruments, detectors, sensors, scanners, …
  - Organising their effective use is the challenge

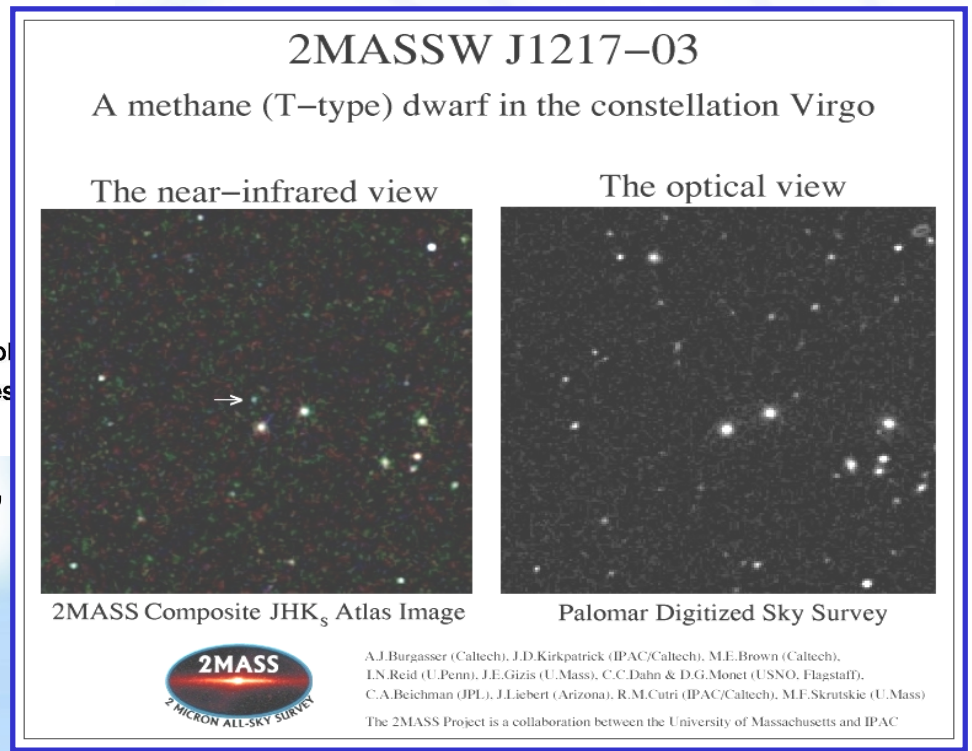
Derived from Ian Foster's slide at ssdbM July 03

Global Knowledge Communities driven by Data: e.g., 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




THE NEW YORK TIMES **NATIONAL** TUESDAY, JUNE 1, 199

Astronomers Detect New Category of Elusive 'Brown Dwarfs'

By JOHN NOBLE WILFORD
CHICAGO, May 31 — Ambitious

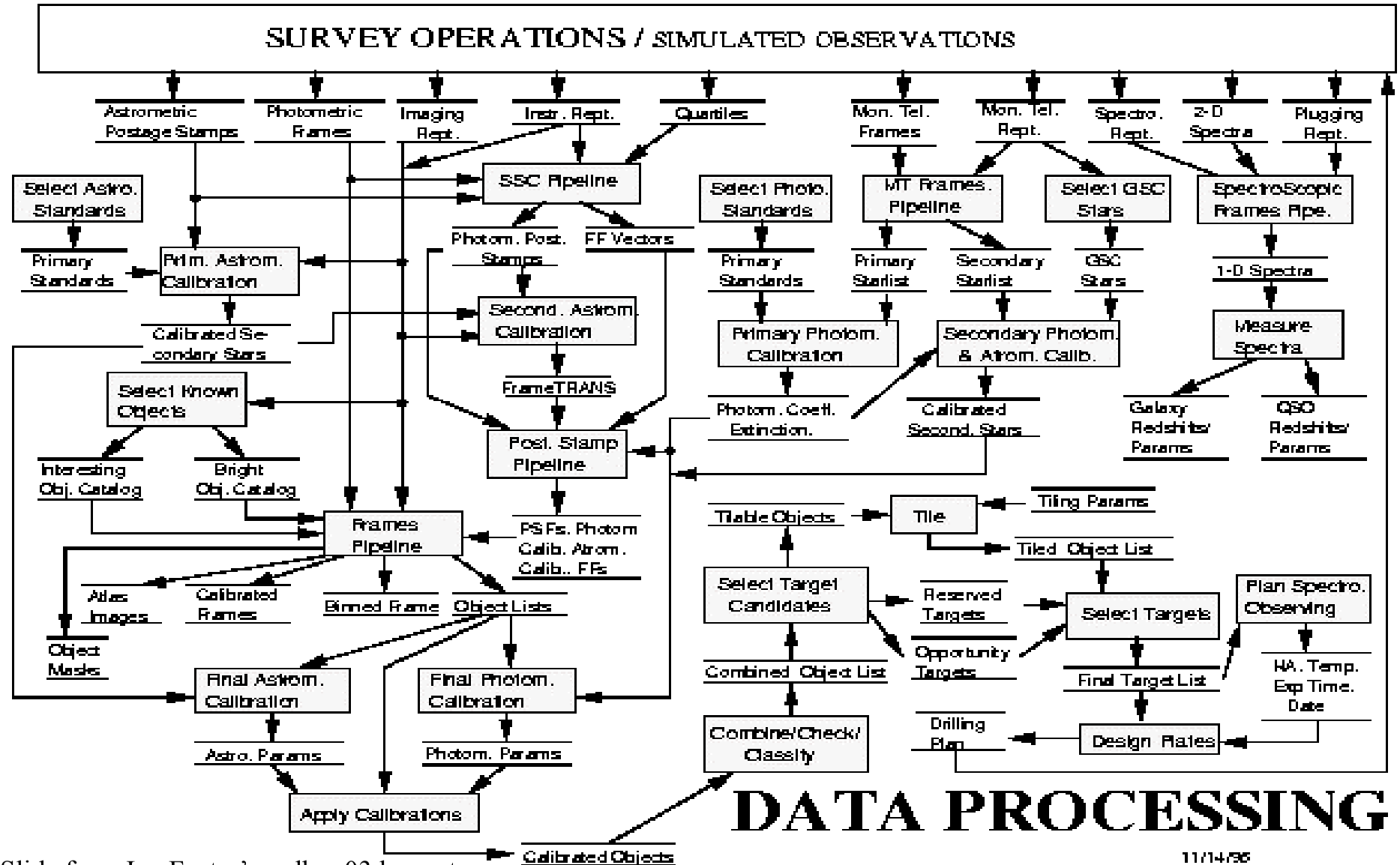
Apache Point, N.M. Dr. Michael Strauss and a graduate student,

was a brown dwarf, but was not associated with a star companion.

have been possible in hotter, younger objects. An estimate of their mass

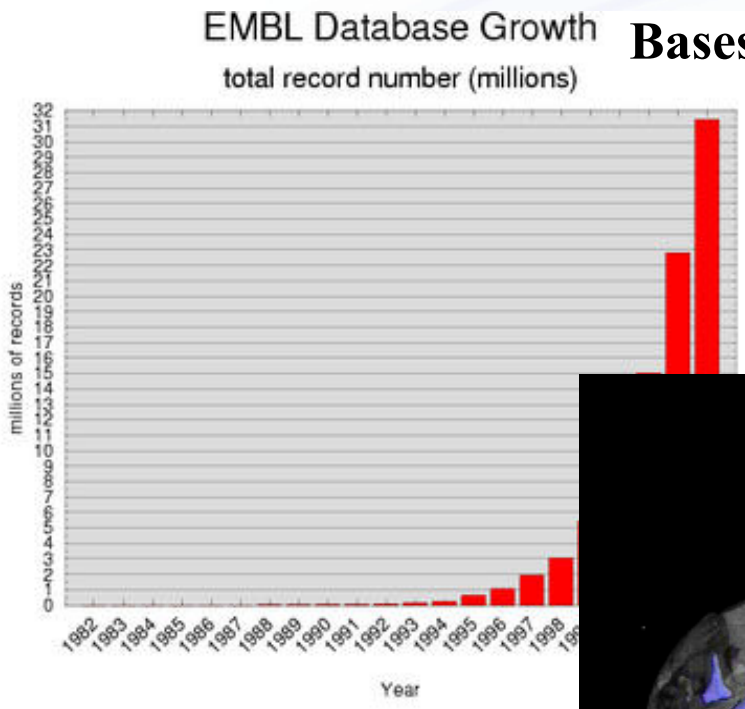
Data and images courtesy Alex Szalay, John Hopkins

Sloan Digital Sky Survey Production System



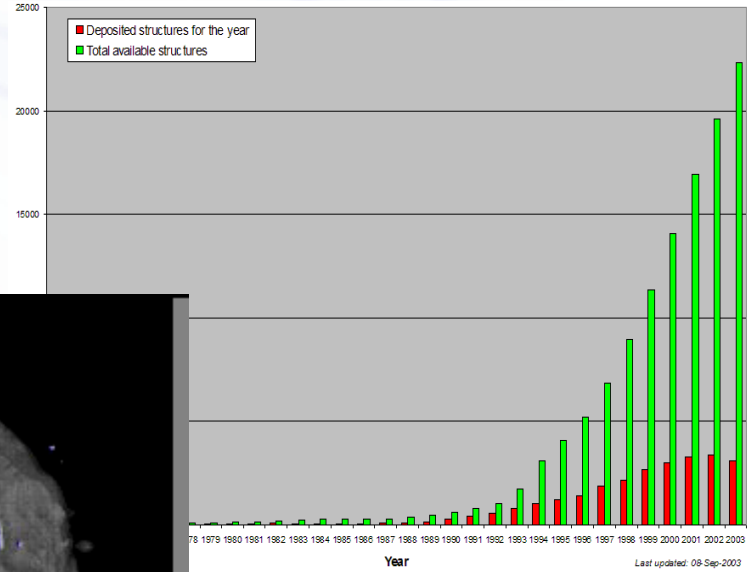
Slide from Ian Foster's ssdbm 03 keynote

Database Growth



Bases 45,356,382,990

PDB Content Growth



Medical Research Council

MRC HUMAN GENETICS UNIT · EDINBURGH ·

epcc

Tera → Peta Bytes

- RAM time to move
  - 15 minutes
- 1Gb WAN move tim
  - 10 hours ($1000)
- Disk Cost
  - 7 disks = $5000 (SCSI)
- Disk Power
  - 1
- Disk
  - 5
- Disk Footprint
  - Inside machine

- A
  - t
- WAN move time
  - 14 months ($1 million)
- Disk Cost
  - 680
  - r
- Disk Power
  - 100 Kilowatts
- Disk Weight
  - 33 Tonnes
- Disk Footprint
  - 60 m²

Modern Hospital: 1 Petabyte/year

How make it secure & reliable!

May 2003 Approximately Correct

See also *Distributed Computing Economics* Jim Gray, Microsoft Research, MSR-TR-2003-24

The Story so Far

- Technology enables Grids and MORE Data & …
- Information Grids will dominate
- Collaboration essential
  - Combining approaches
  - Combining skills
  - Sharing resources
- (Structured) Data is the language of Collaboration
  - Data Access & Integration a Ubiquitous Requirement
- Many hard technical challenges
  - Scale, heterogeneity, distribution, dynamic variation
- Intimate combinations of data and computation
  - Unpredictable (autonomous) development of both

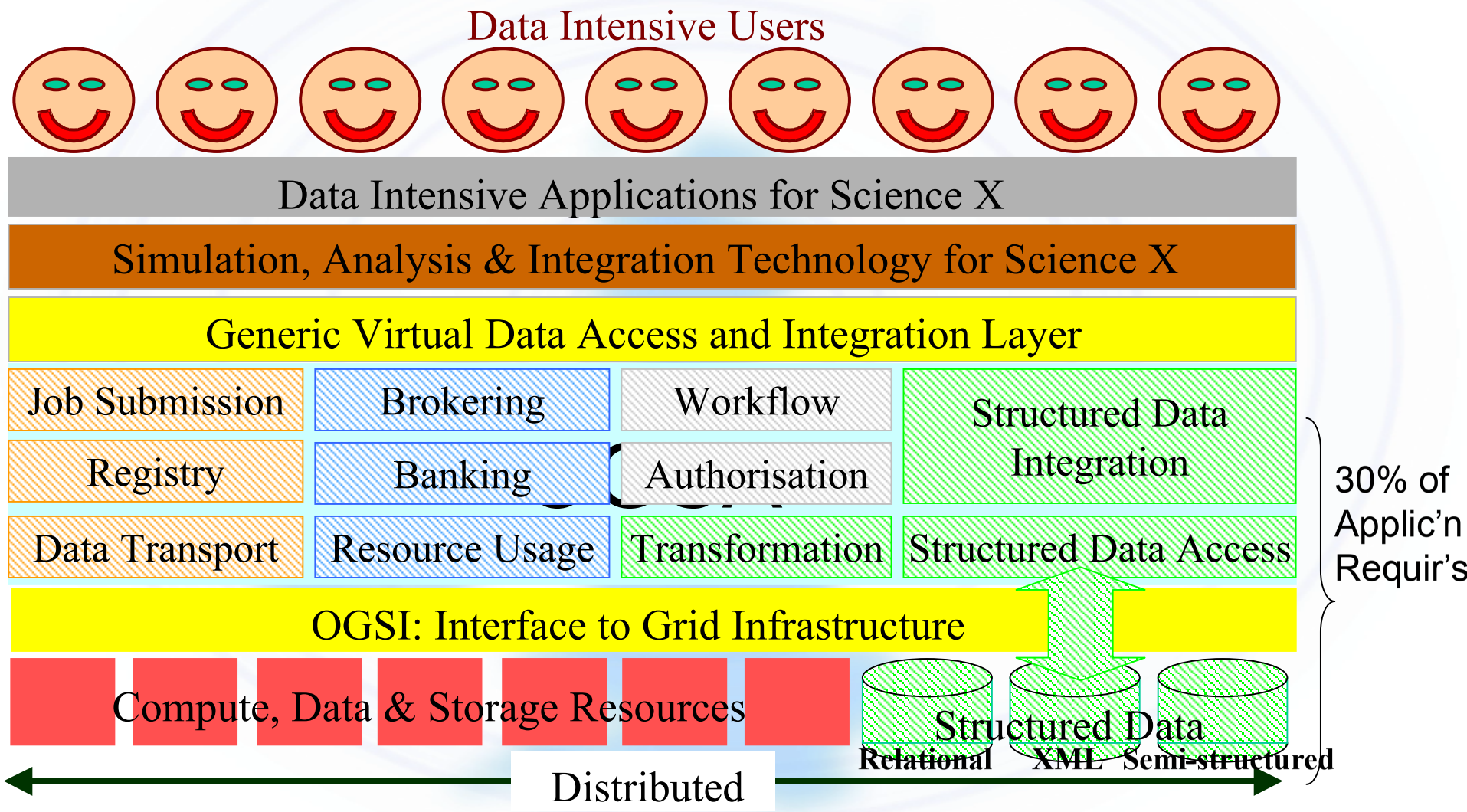
Scientific Data

- **Opportunities**
  - Global Production of *Published* Data
  - Volume↑ Diversity↑
  - Combination ⇒ Analysis ⇒ Discovery
- **Challenges**
  - Data Huggers
  - Meagre metadata
  - Ease of Use
  - Optimised integration
  - Dependability
- **Opportunities**
  - Specialised Indexing
  - New Data Organisation
  - New Algorithms
  - Varied Replication
  - Shared Annotation
  - Intensive Data & Computation
- **Challenges**
  - Fundamental Principles
  - Approximate Matching
  - Multi-scale optimisation
  - Autonomous Change
  - Legacy structures
  - Scale and Longevity
  - Privacy and Mobility

Contents

- Data:
  The Lingua Franca of e-Science
- Data:
  The Challenge for e-Science
- **OGSA-DAI *Product*:** ☜ **you are here**
  - **The First Steps in DAI**
  - **An opportunity for collaboration**
- **OGSA-DAI *Product*:**
  The Next Steps
  - **More collaboration please**

Infrastructure Architecture



Virtual Integration Architecture

Data Services

- **GGF Data Access and Integration Svcs (DAIS)**
  - **OGSI-compliant interfaces to access relational and XML databases**
  - **Will be generalized to encompass other data sources (see next slide…)**
- **Generalised DAIS is the foundation for:**
  - **Replication:**
    - Copies of data in multiple locations
  - **Federation:**
    - Composition of multiple sources
  - **Provenance: How was data generated?**

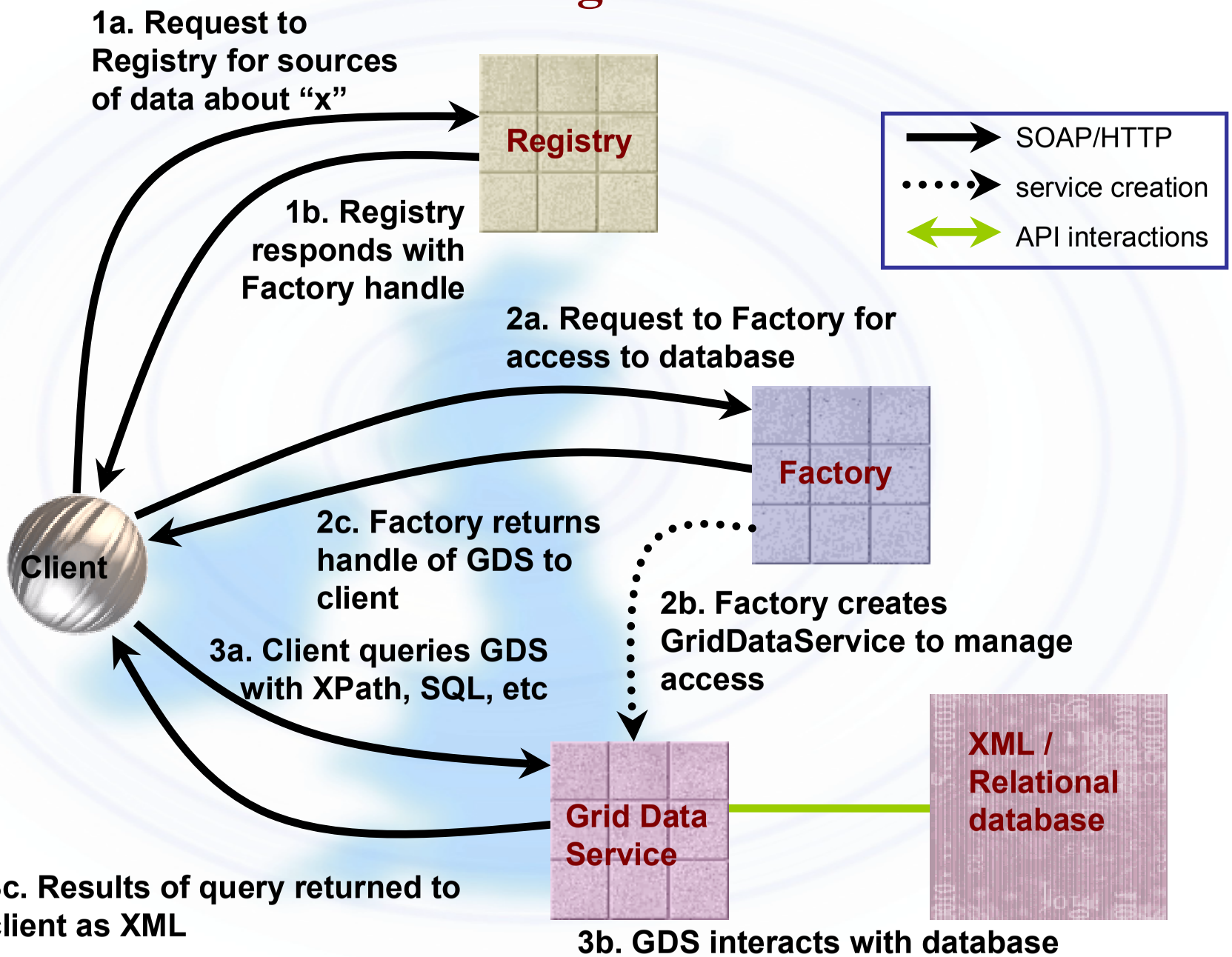
"OGSA Data Services" (Foster, Tuecke, Unger, eds.)

- **Conceptual model for representing all data sources as Web services**
  - **Database, filesystems, devices, programs, …**
  - **Integrates WS-Agreement**
- **Data service is an OGSI-compliant WS**
  - **implements ≥1 of base data interfaces:**
    - DataDescription, DataAccess, DataFactory, DataManagement
  - **Extended and combined for specific domains**
    - E.g. DAIS

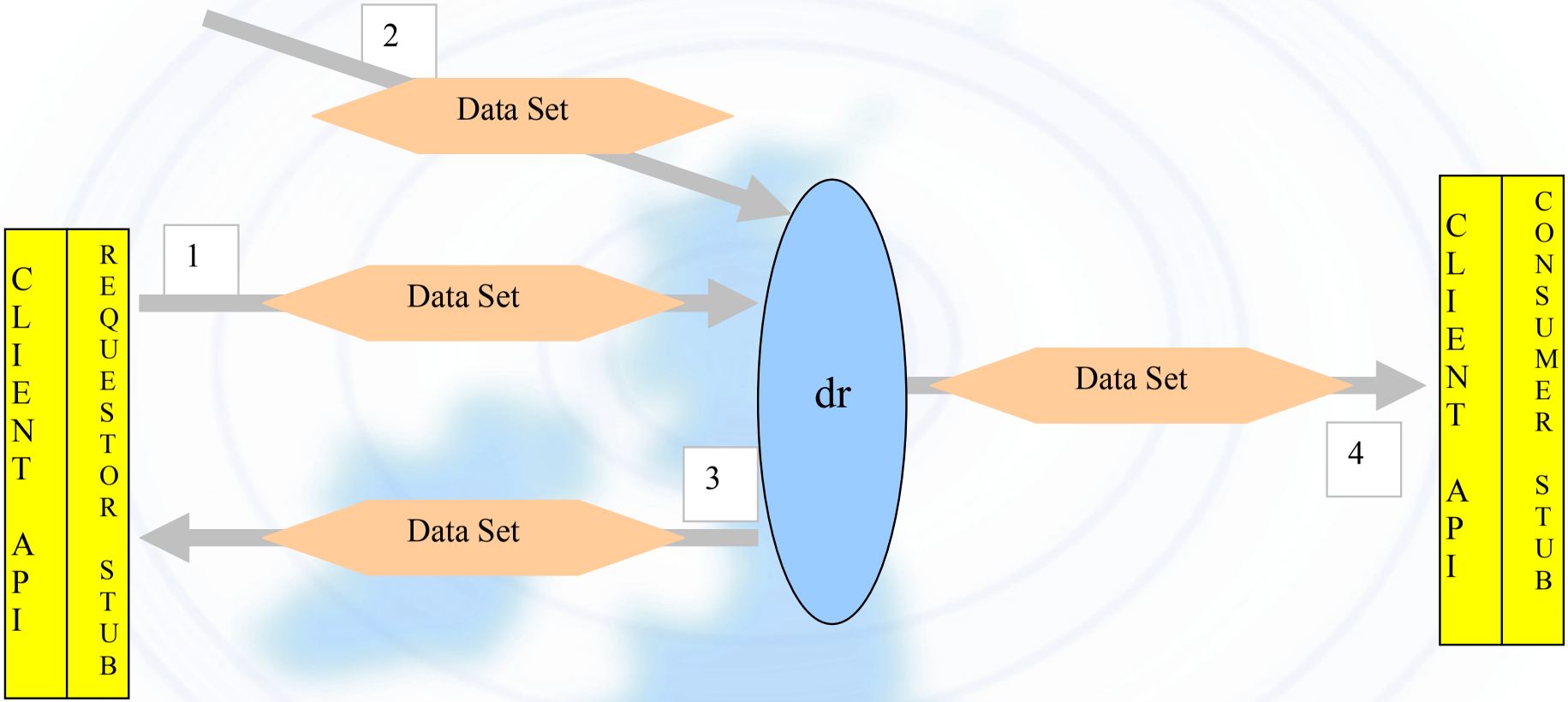
OGSA-DAI Approach

- **Reuse existing technologies and standards**
  - OGSA, Query languages, Java, data transport
- **Build portTypes and services that will enable:**
  - controlled exposure of heterogeneous data resources via an OGSI-compliant grid
  - access to these resource via common interfaces
  - using existing underlying query mechanisms
  - (ultimately) data integration across distributed data resources
- **OGSA-DAI *Product***
  - Reference implementation of GGF DAIS WG standard
  - Balance standard tracking & testing
  - With stability for application and product developers
- **See http://www.ogsadai.org.uk/ for details.**

Data Access & Integration Services



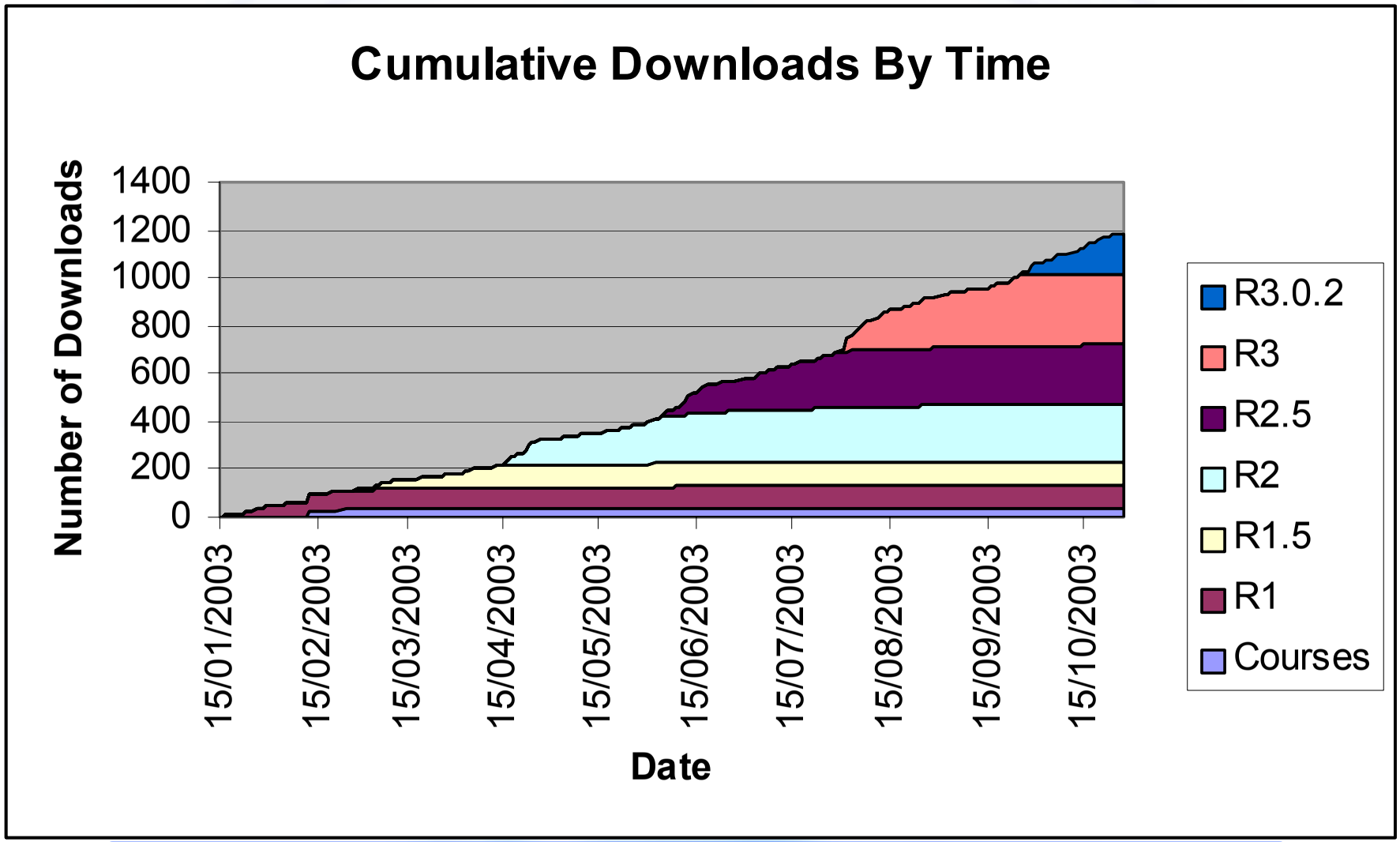
Third Party Delivery



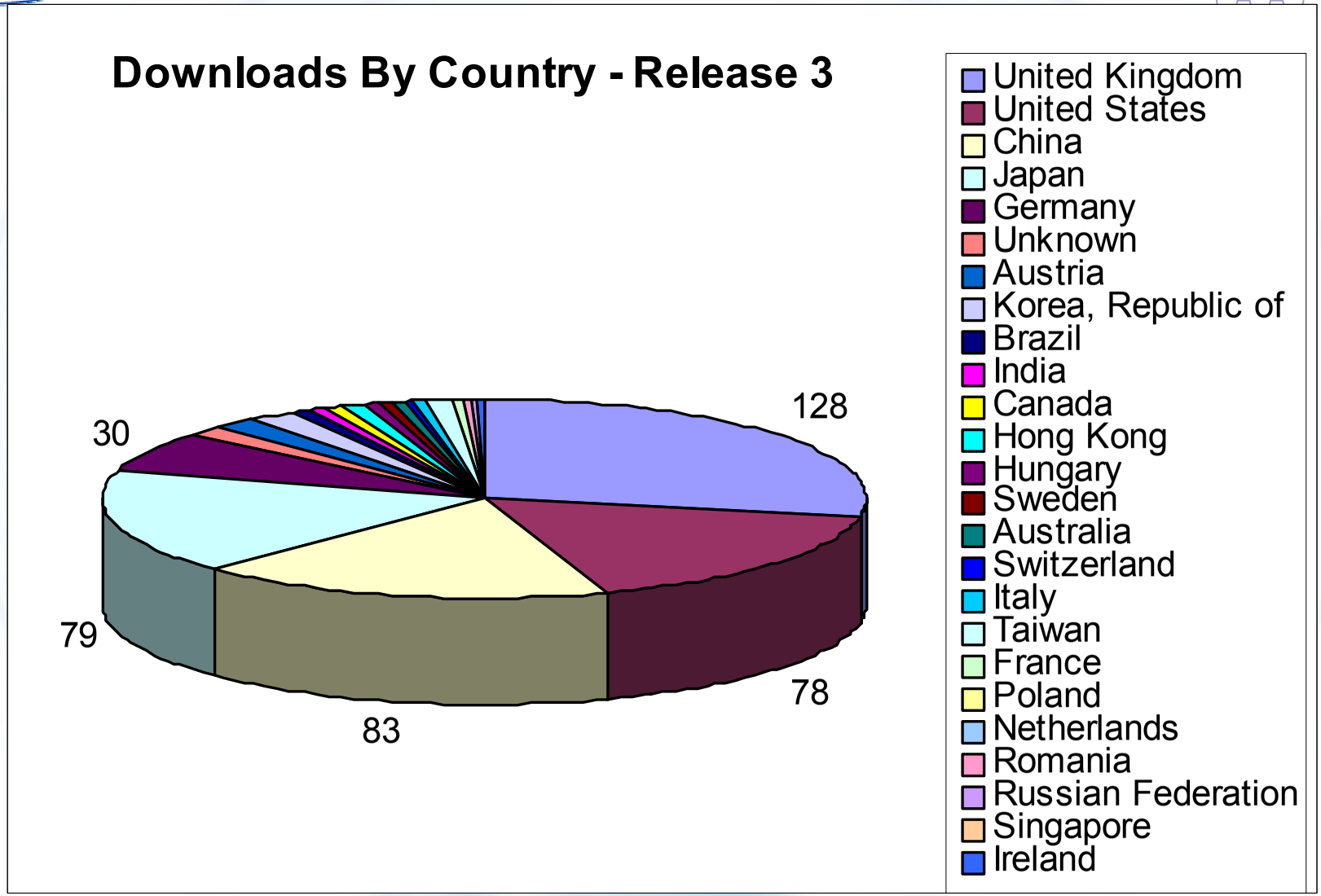
OGSA-DAI *Product*

- **Brand name: OGSA-DAI**
  - **Established**
- **Current release R3.0.2**
  - **OGSA-DAI: 1183 downloads**
    - 461 R3 & R3.0.2
    - >379 in UK
  - **50 downloads of R3.0.0 of R3.0.2 within a week**
  - **Recent performance analysis ⇒ R3.0.3 Nov 03**
  - **DQP prototype: 77 downloads**
    - Since 1st September 2003
- **Web site** **www.ogsadai.org.uk**
  - **471 registered users**

Releases & Downloads



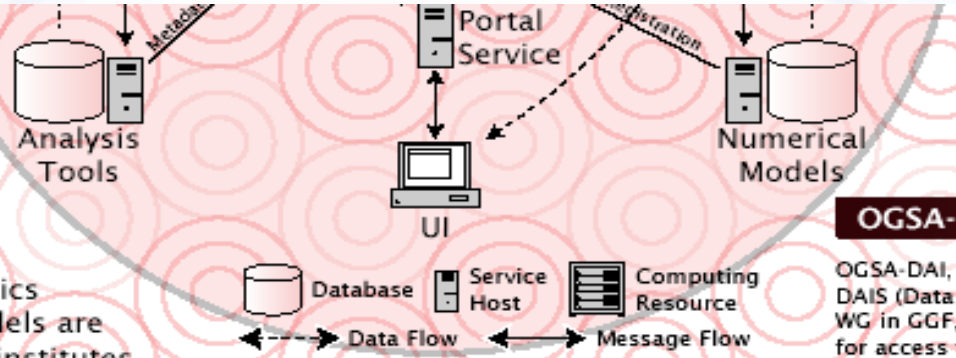
OGSA-DAI downloads

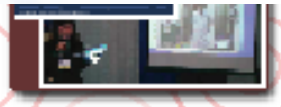


the globus alliance
www.globus.org

National e-Science Centre

automatically store data into database. MEG is a device which can detect the change in minute magnetic fields generated from the brain activity.





Design Strategy

Resources of neuroinformatics such as data, tools and models are located in various research institutes. We focus on metadata-driven access and service oriented grid environment as key technologies for organic linking of various resources.

Metadata-driven access

Metadata takes a central role in our framework. All the data, tools and models are explained with metadata formatted as XML document. Users' requests for computation are processed by Portal Service, which retrieves information about appropriate resources and organizes them.

Service-oriented grid environment

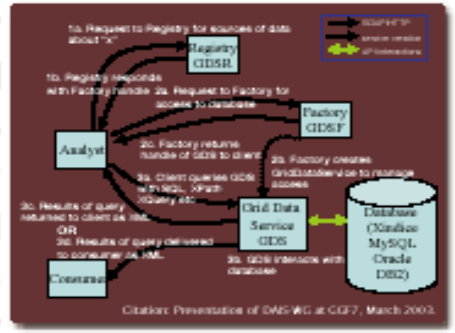
In widely distributed grid environment, data and metadata should be provided through commonly accepted interfaces.

Currently we are examining Globus Toolkit 3 and OGSA-DAI in terms of their functionality, working speed and so on.

OGSA-DAI

OGSA-DAI, which is under development at DAIS (Data Access and Integration Service) WG in GGF, provides a uniform framework for access to databases on the Grid. This tool enables us to use conventional SQL statement and XML search language such as XPath without any extra modification.

We envision that OGSA-DAI will be adopted in many domains involving grid and DBs.



Acknowledgement: This work was supported in part by a Grant-in-Aid for Scientific Research on the Priority Area, " Informatics Studies for the Foundation of IT Evolution" (13224059) by the Ministry of Education, Culture, Sports, Science and Technology of Japan, and the IT-program (Construction of Supercomputer Network) of the Ministry of Education, Culture, Sports, Science and Technology.

biogrid project
http://www.biogrid.jp

Researchers: Takahiro Kosaka (tak-k@ais.cmc.osaka-u.ac.jp)
Susumu Date (date@ais.cmc.osaka-u.ac.jp)
Yuko Mizuno-Matsumoto (yuko@ais.cmc.osaka-u.ac.jp)
Shinji Shimojo (shimojo@cmc.osaka-u.ac.jp)

Supported by:

epcc
e-Science
UNIVERSITY of GLASGOW

Contents

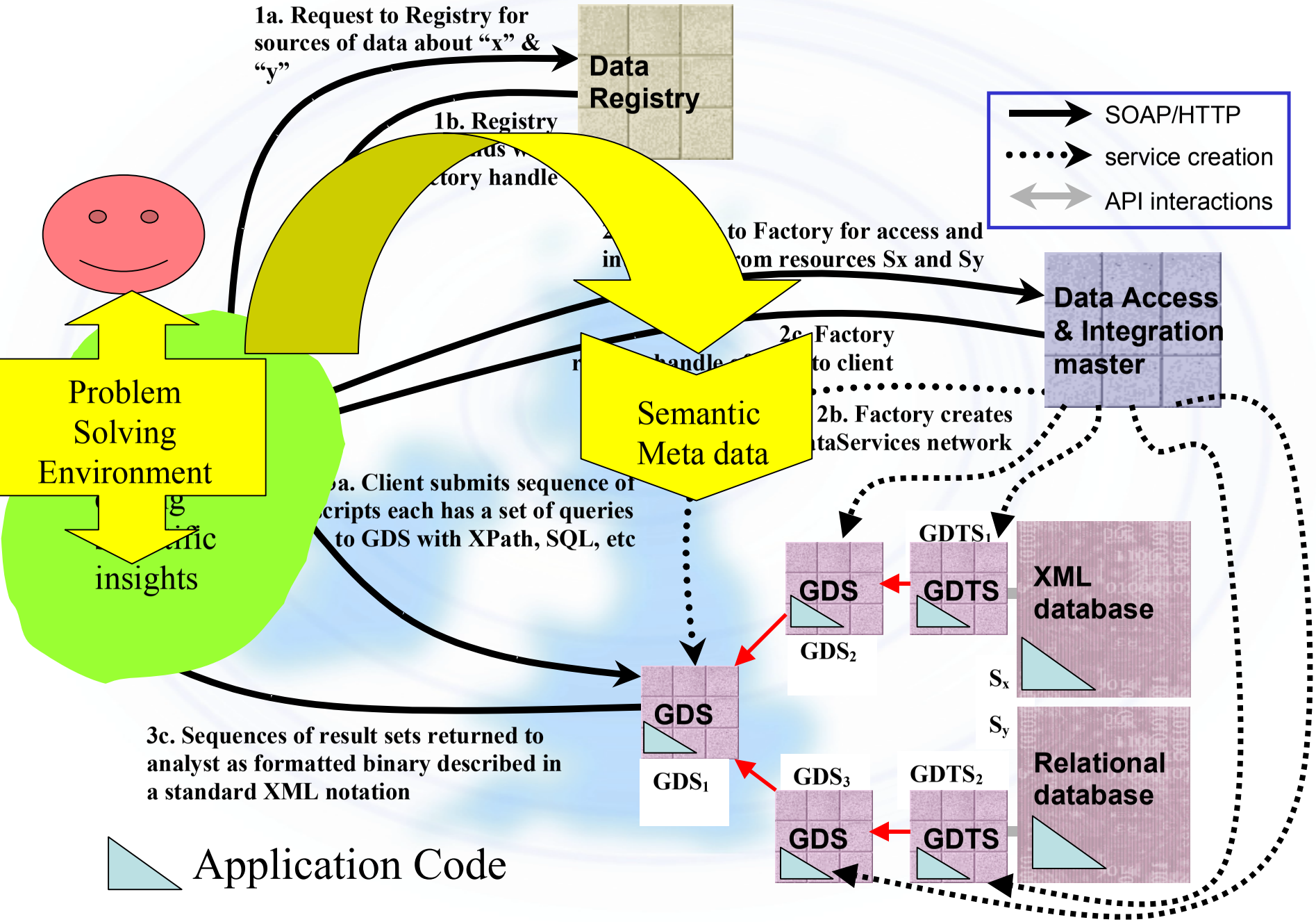
OGSA-DAI road map 1

- **R3.1.0 Jan 04 Tech. Preview part of R4**
- **User Group: inaugural meeting Q1 04**
- **R4.0.0 April 04**
  - **Performance & monitoring**
  - **Additional DBMS's supported**
  - **Additional SQL supported**
  - **DBMS management operations**
    - archive, restore, bulk load
  - **File access**
  - **Client libraries**
  - **Installation wizard**
  - **User support, courses, training material, performance report**

OGSA-DAI road map 2

- **R5 October 04**
  - Compliance with DAIS standards proposal
  - Distributed Relational Query Processing
  - Improved dependability and security integration
  - Extended & integrated XML and relational facilities
  - Distributed transaction participation
  - Coordinated OGSA-DAI contributor community
- **R6 April 05**
  - Integrated with GT3
  - New facilities depend on user priorities, context and research
  - OGSA-DAI components from contributor community
- **R7 October 05**
  - Maintainable release for the user community

Future DAI Services



Take Home Message

- **Data is a Major Source of Challenges**
  - **AND an Enabler of**
    - New Science, Engineering , Medicine, Planning, …
- **Information Grids**
  - **Support for collaboration**
  - **Support for computation and data grids**
  - **Structured data fundamental**
  - **Integrated strategies & technologies needed**
  - **Raise the level of discourse**
  - **Automate generation & use of semantic data**
- **OGSA-DAI is here now**
  - **Join in making DAI services & standards**

Many opportunities for International collaboration

Comments & Questions

www.ogsadai.org.uk