Developing a base data layer in the OGSA-DAI project

2nd Workshop on Scientific Data Mining, Integration and Visualisation

eSI, 14 - 15 December 2005

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

Overview

- The Data Deluge
  - challenges of increasing data availability
  - benefits of bringing data together
- OGSA-DAI
  - overview
  - use as a data integration base layer

Data Services: challenges to integration

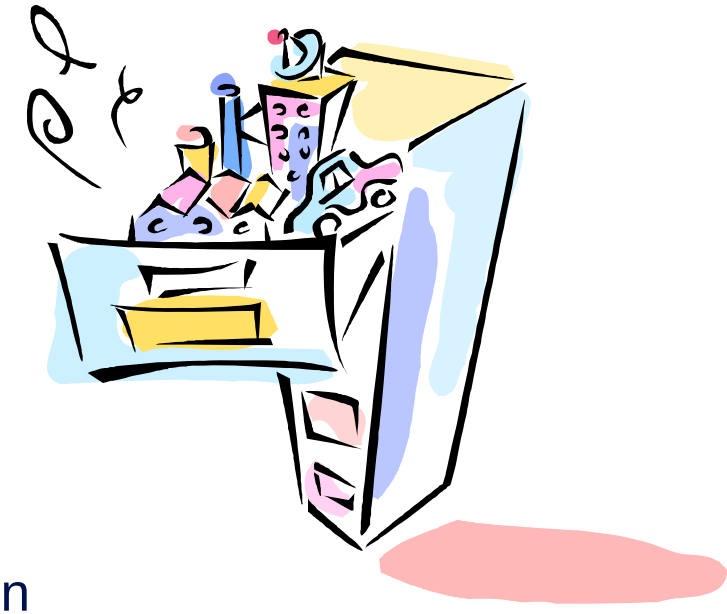
- Scale
  - Many sites, large collections, many uses
- Longevity
  - Research requirements outlive technical decisions
- Diversity
  - No “one size fits all” solutions will work
    - Primary Data, Data Products, Meta Data, Administrative data, …
- Many Data Resources
  - Independently owned & managed
    - 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!

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

Use Cases for Data Services

- Data Filtering:
  - Single source producing large amounts of data distributed to many sites downstream
- Data Discovery:
  - many sources, many query entry points in a linked system
- Data Translation:
  - source to sink, conversion of data model / structure
- Data Federation:
  - many sources, linked to provide view as a single source
- Data Replication
  - full or partial copies to improve throughput
- Data Integration (model aggregation)
  - e.g. integration of time variant data, streams, files
- Data Integration (knowledge expansion)
  - forming links between databases to increase knowledge

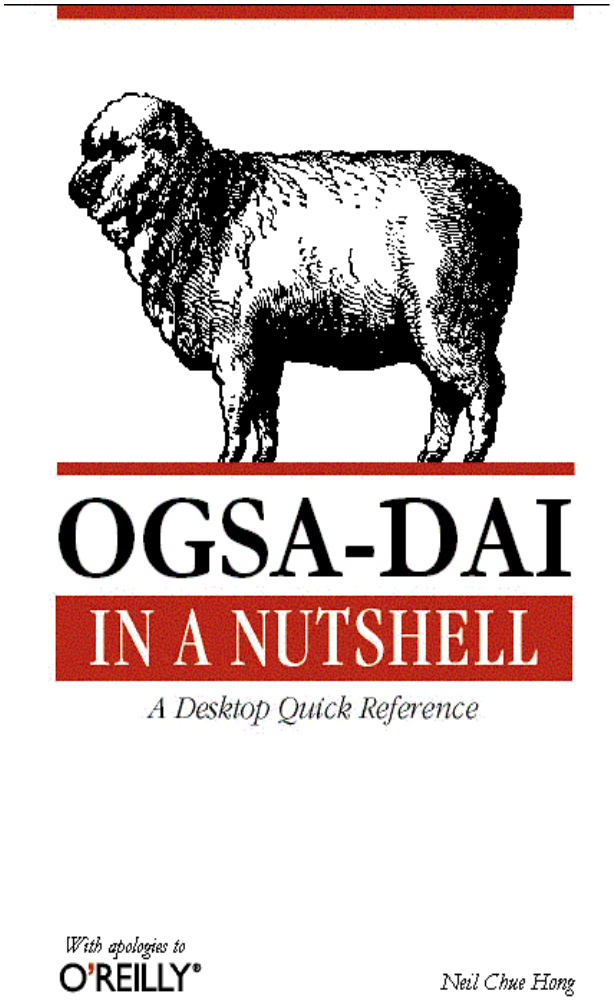
Trade Offs

- Speed vs completeness
  - do you require the exact answer or an answer?
- Application specific vs language specific queries
  - how will users interrogate a data service?
- Static system vs Dynamic Discovery
  - do you actually have dynamic resources?
- Static vs Dynamic data
  - READ only, READ/INSERT only, UPDATE permitted
- Static vs Dynamic queries
  - optimisation over flexibility
- Intranet vs Internet
  - speed over security
- Single data model versus mixed data models
  - ease/speed over integration
- Queries vs Questions
  - assume that we know the structure when we form the query

Requirements on Data Services?

- Common Data Model e.g. RowSet
- Common Query Language(s) e.g. XQuery, SQL
- Standard access to
  - data resource schema information for schema mapping
  - physical data resource information for optimisation purposes
  - data resource descriptive information for discovery / integration
- Single, seamless security model
- Dynamic publication and discovery
- Multiple, efficient delivery methods
- Move computation towards data
- Data aggregation functionality
- Provenance information
- Replication information

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,…

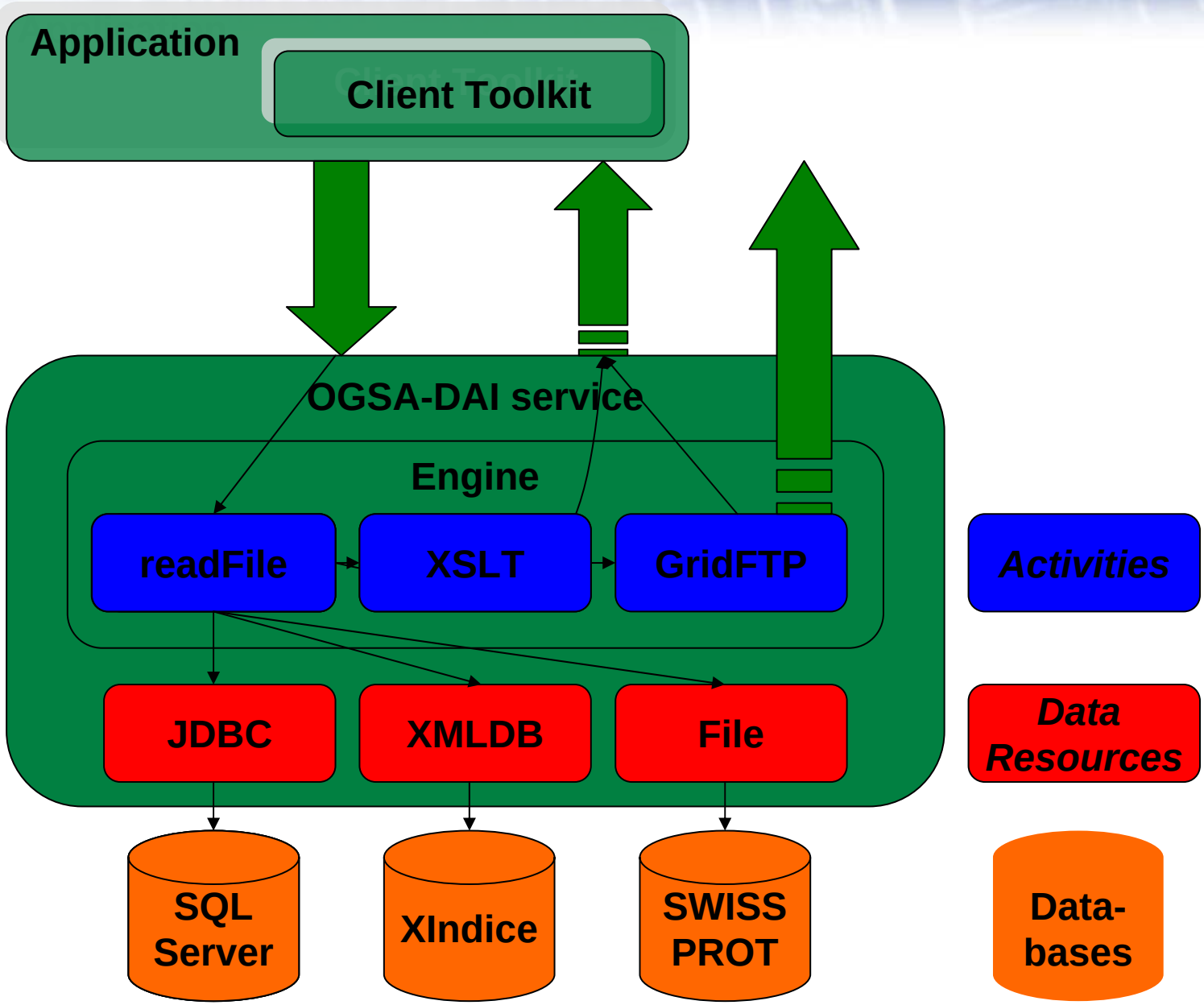
OGSA-DAI Design Principles – I

- 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
  - DAIS specifications from DAIS WG at GGF

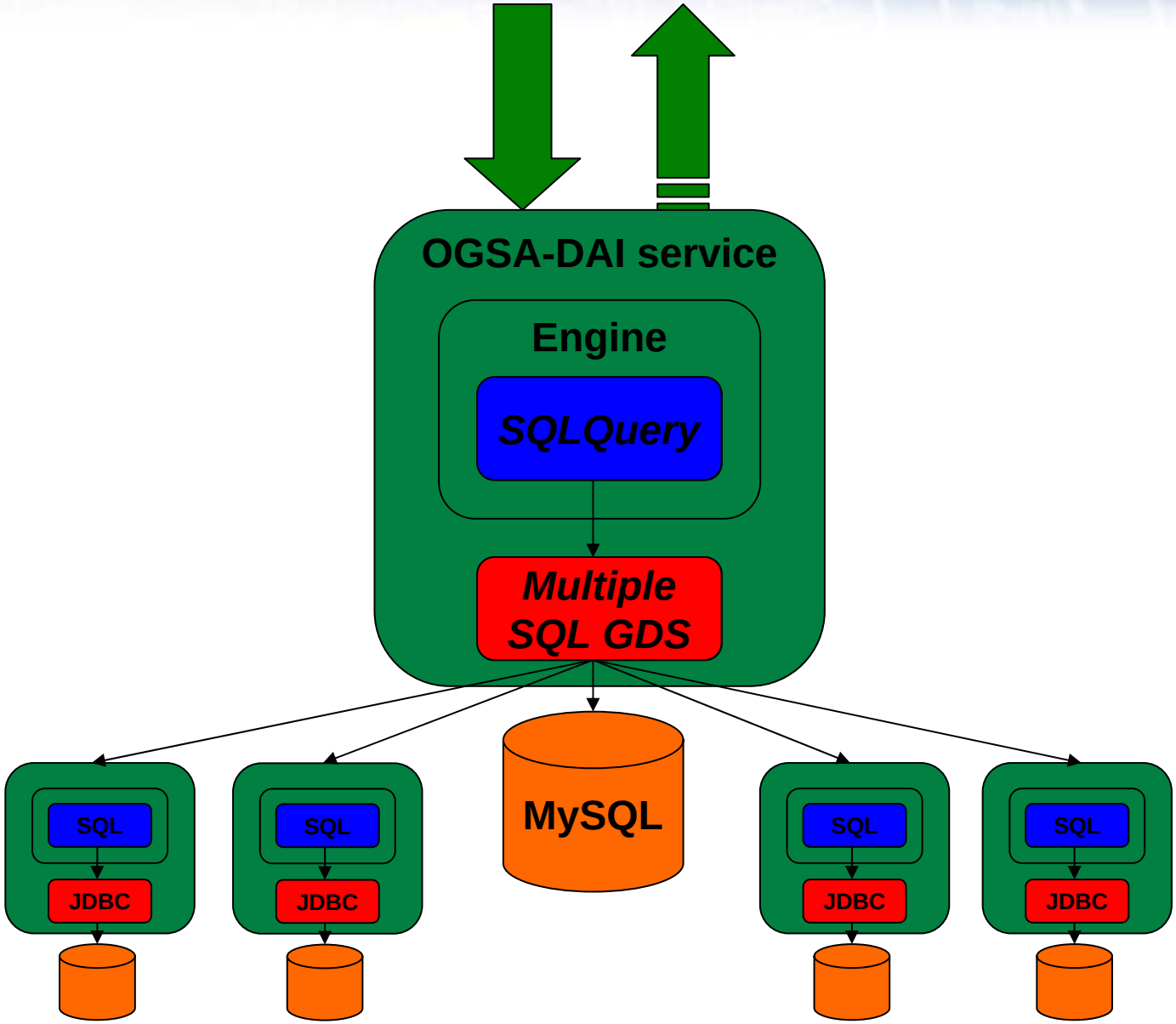
OGSA-DAI Design Principles – II

- 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

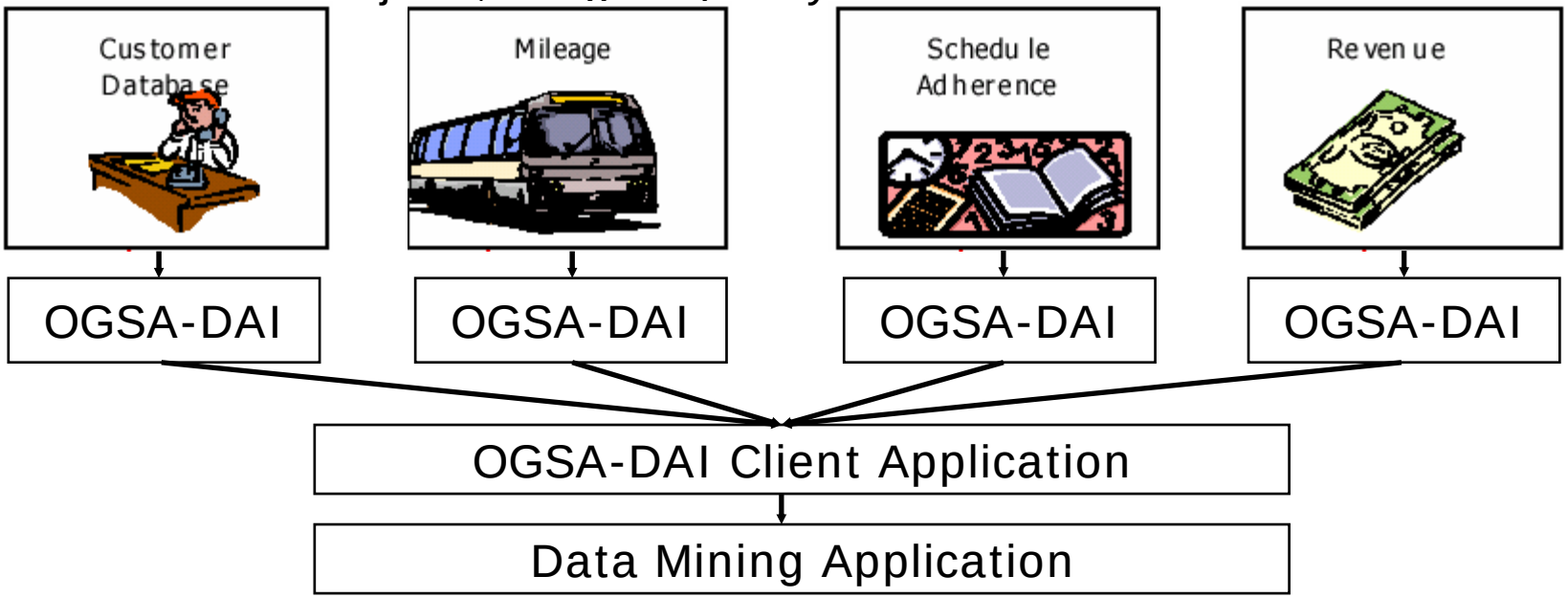


Extensibility Example



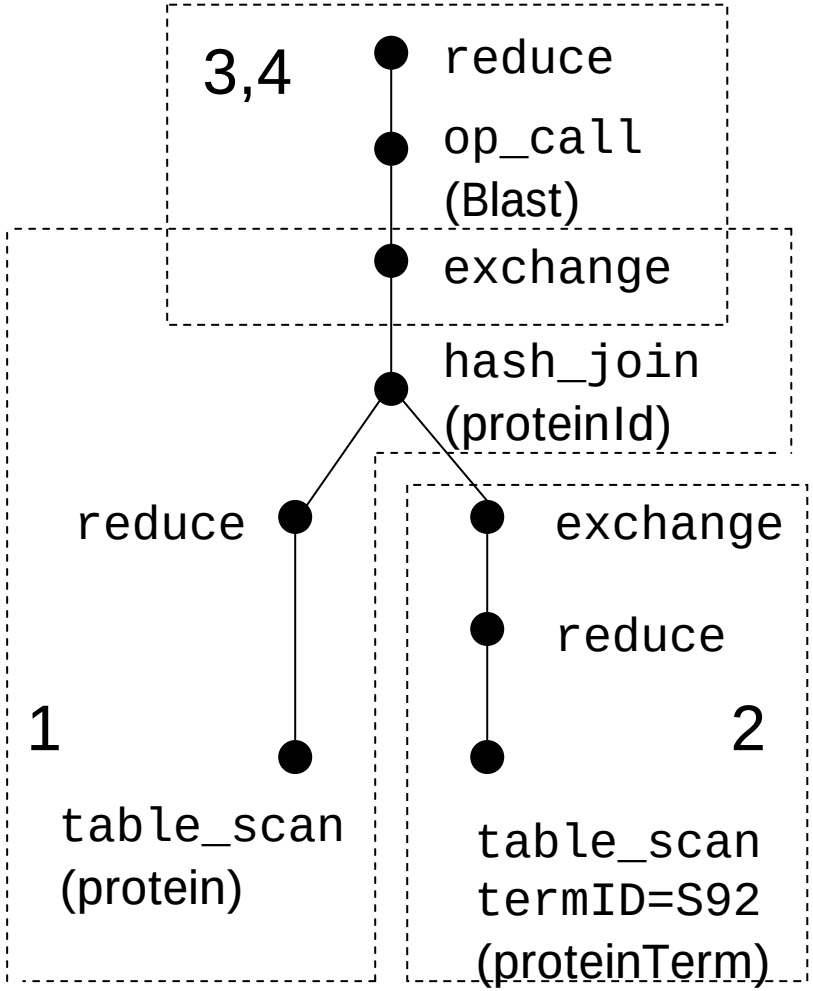
FirstDIG

- 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.
  - Client based joins, using temporary tables

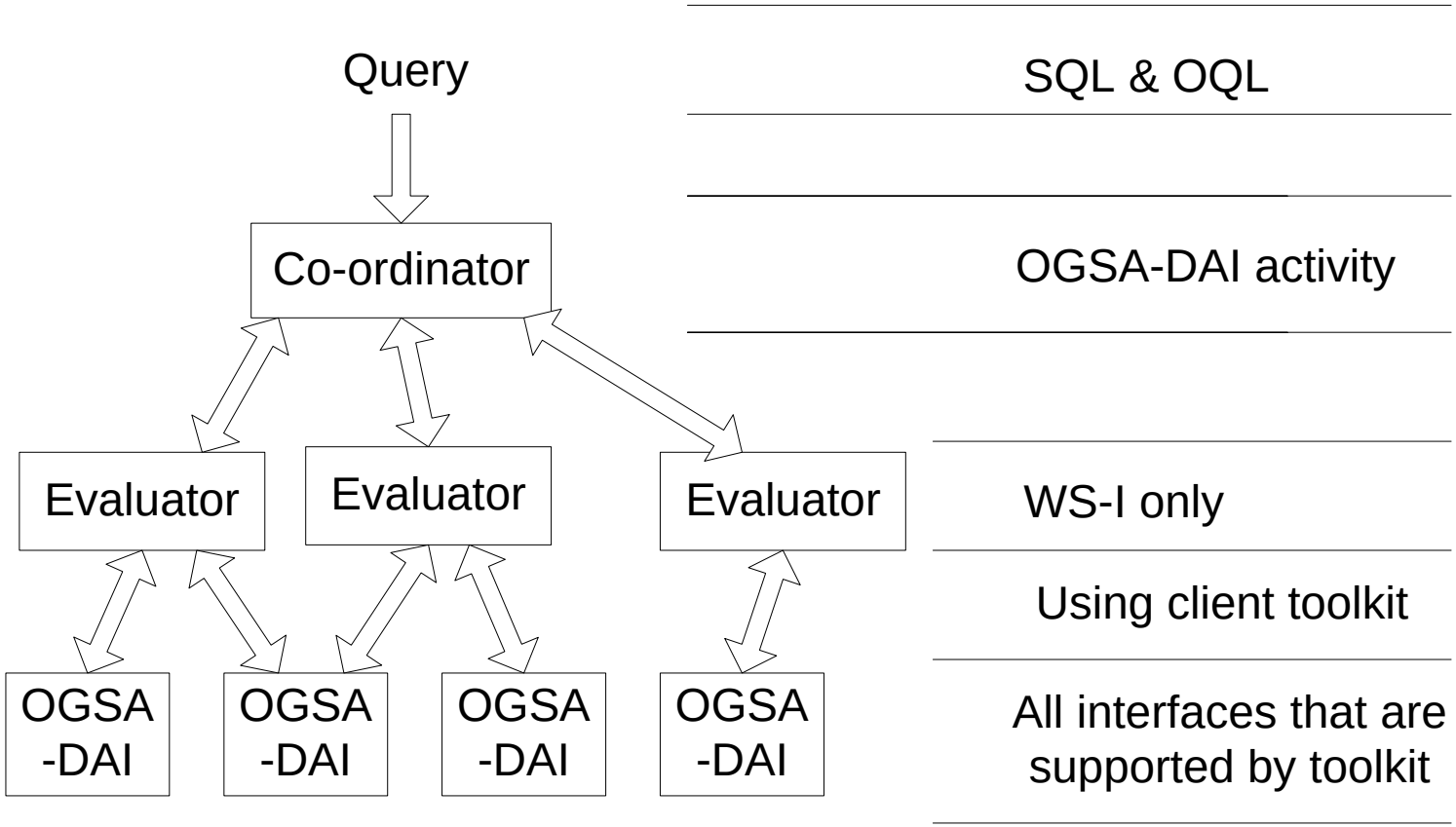


Distributed Query Processing

- Higher level services building on OGSA-DAI
  - specialised metadata extraction
- Execute queries in parallel over multiple data resources
- Queries mapped to algebraic expressions for evaluation
- Parallelism represented by partitioning queries
  - Use exchange operators
- Equality based joins in current release
  - supported types: long, integer, string, double and float



DQP architecture



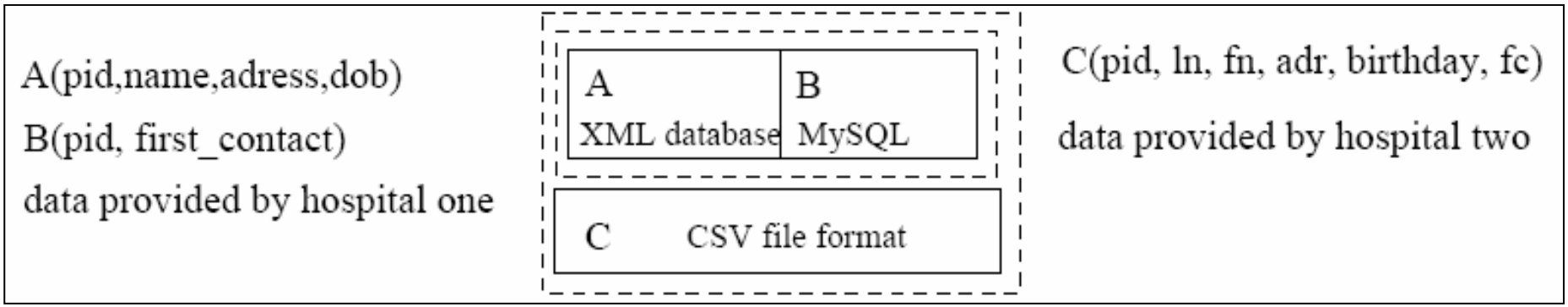
GridMiner: Data Mediation Service

Principles

- Tight Federation:
  - global (relational) schema
- Virtual integration:
  - leave the data where it is
  - always up-to-date data
- Build on data access from OGSA-DAI
- Not bound to special architecture

- **Supported data sources:**
  - RDBMS (via JDBC), XMLDB (Xindice), CSV files
- **Operators: “Union all” and “inner join”**
- **Operators are XQuery based (using SAXON)**

Data Integration Scenario



Heterogeneities:

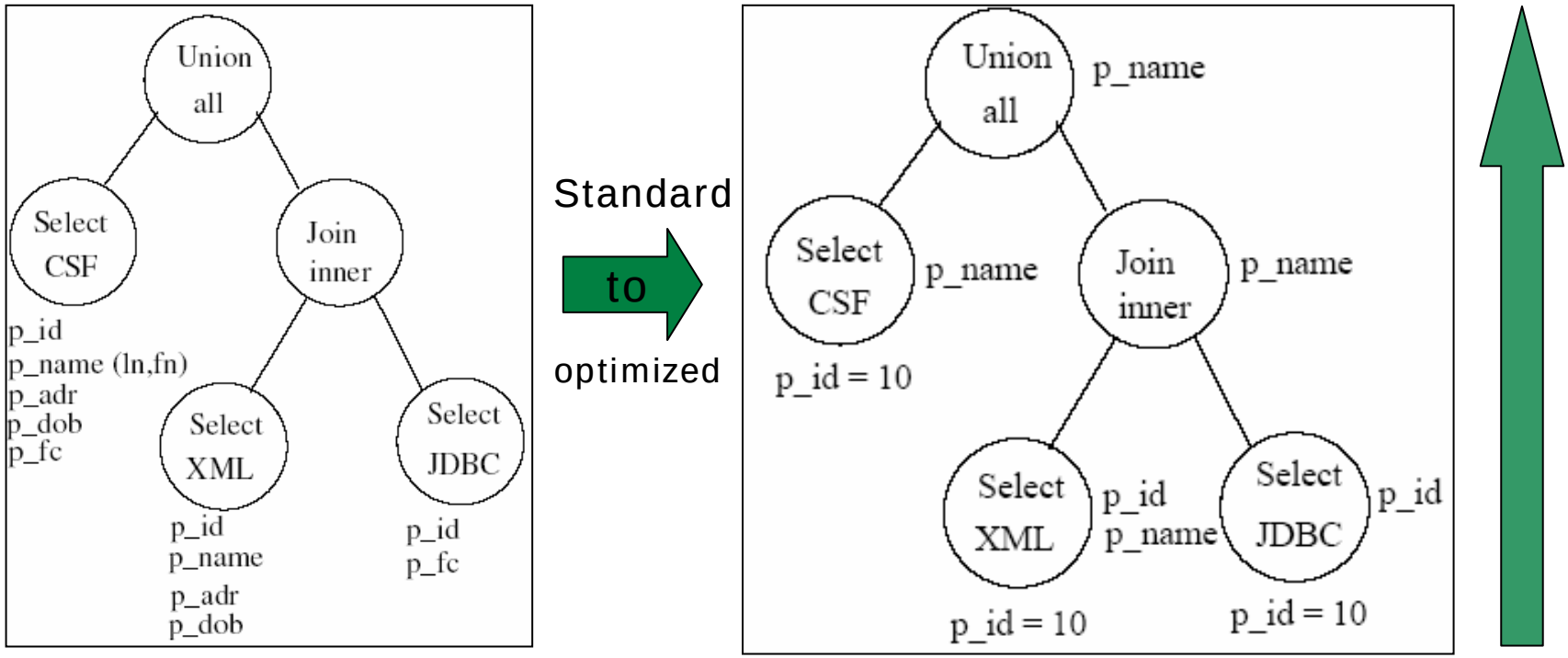
- Name in A is „First Last“ (as the target format)
- Name in C has to be combined

- Distribution:
  - 3 data sources
- Java based schema mapping to global schema
  - types limited by WebRowSet

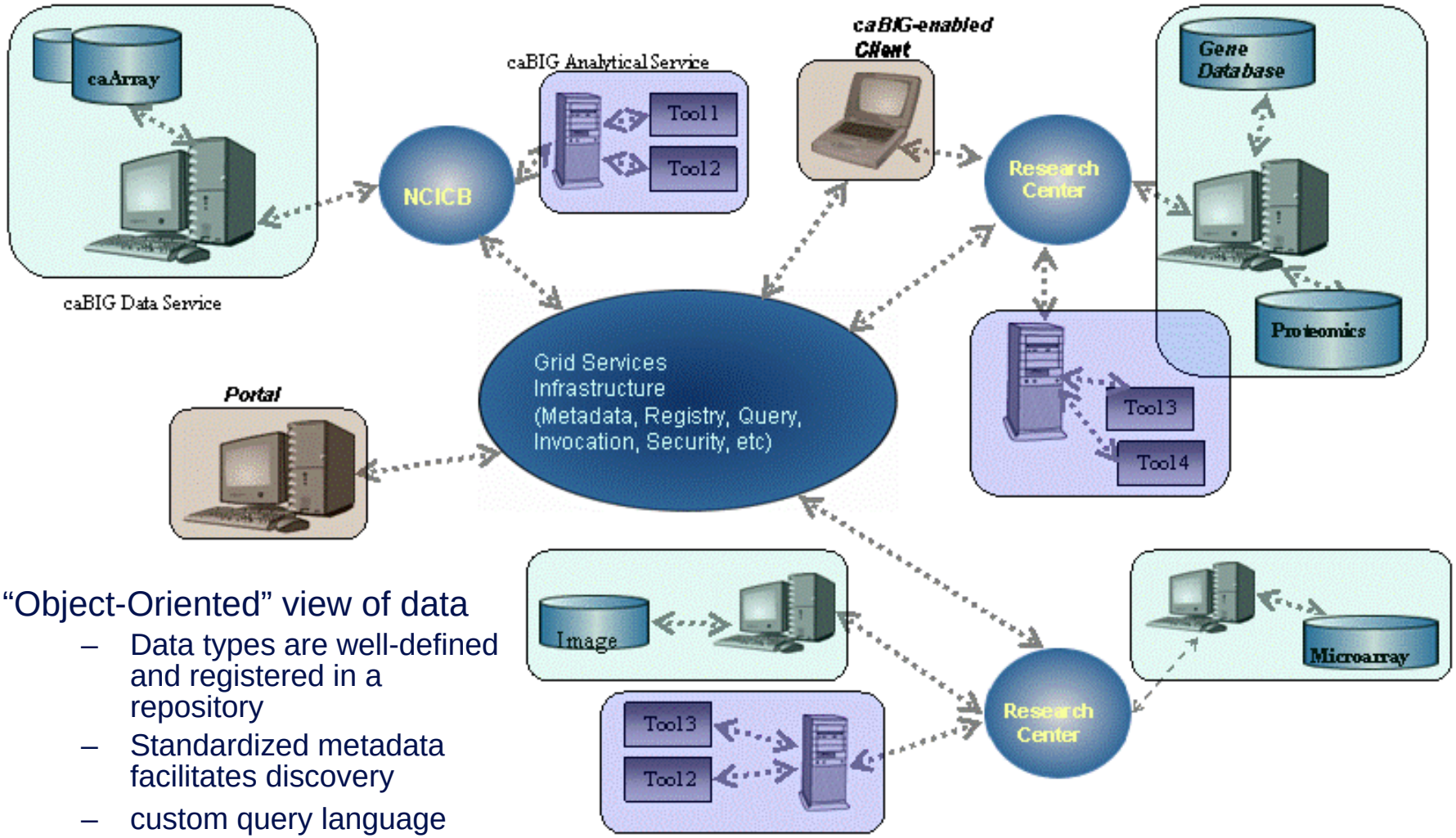
Data Integration Scenario (cont.)

- Query:

  SELECT p_name FROM patient WHERE id=10



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

OGSA-DAI 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
  - aggregation of results
- Data transport and workflow
  - workflow is typically compute driven
- Move computation to data
  - mobile code activities?
  - data services hosted on DBMS?

Contributing to OGSA-DAI

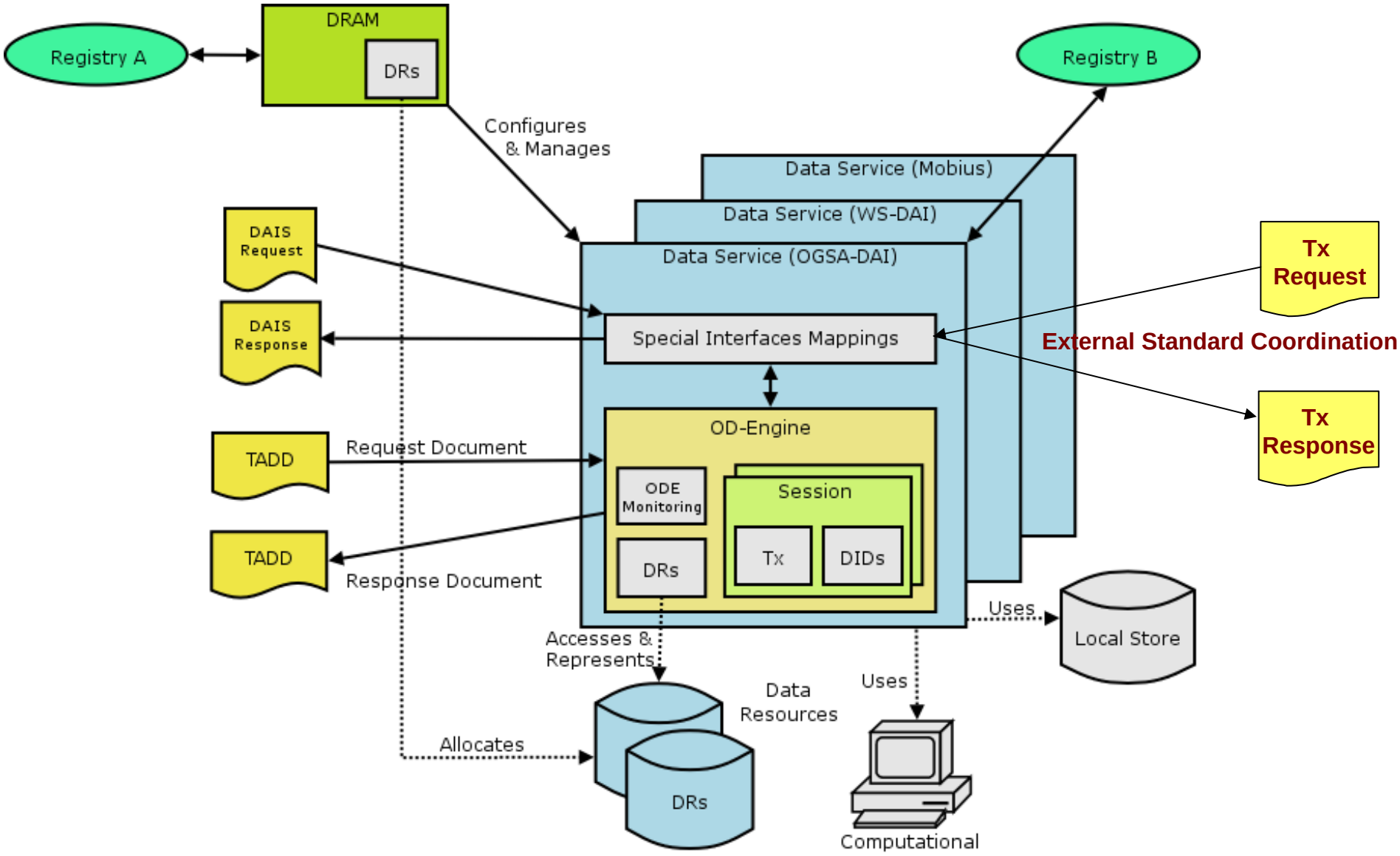
- Additional functionality:
  - Provide activities which implement specific functionality
  - Provide extra client functionality
  - Provide different security mechanisms
  - Provide higher level components and applications

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

Future plans

- A new version of the OGSA-DAI Engine
  - should look mostly the same externally
  - better support for concurrency, sessions and monitoring
- Implementing new versions of specifications
  - DAIS Specifications
- Key things that we will be addressing:
  - 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
- Research projects looking at:
  - schema mapping
  - extended data resources

New OGSA-DAI Architecture



Associated Meetings and Workshops

- DIALOGUE Workshops (http://www.datagrids.org)
  - Data Integration Applications: Linking Organisations to Gain Understanding and Experience
  - Bringing together Data Integration middleware and application providers with users
  - Next one at NeSC: 9-10th February 2006
    - http://www.nesc.ac.uk/esi/events/636/
- Next Generation Distributed Data Management (HPDC15, Paris)
- *Data Management on Grids (VLDB'06, Seoul)*

Conclusions

- The benefits of trying to integrate data are hindered by challenges such as heterogeneity, scale and distribution
- A common data service layer should make data integration easier
- OGSA-DAI provides an extensible, data service based framework which makes it easier to implement data integration
- Future work on OGSA-DAI is addressing some of the key challenges to data integration

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