

ISPIDER: Grid-based Integration of Biological Data

N. Martin – A. Poulouvassilis – L. Zamboulis
{nigel,ap,lucas}@dcsl.bbk.ac.uk

Overview

- Project overview
- Data Integration: the AutoMed project
- The OGSA-DAI project
- Project Implementation

Project Details

- 3 year BBSRC-funded project
- Members
 - Birkbeck College
 - European Bioinformatics Institute
 - University of Manchester
 - U.C.L.

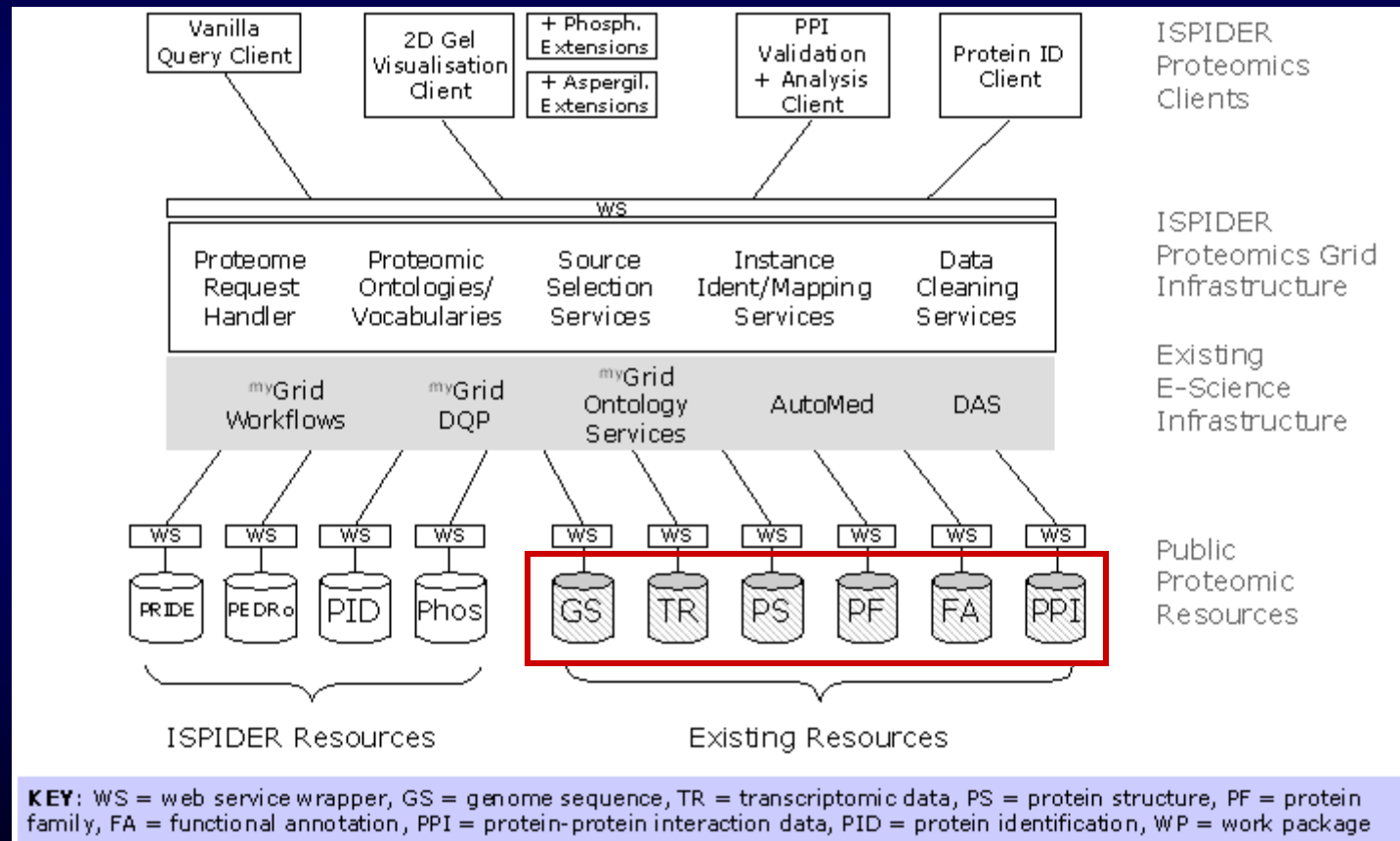
Project Goal

- Produce an integrated platform for biologists
 - Human Genome Project completed in 2003
 - Laboratories across the world produce vast amounts of experimental data
 - Combining efforts will result in added value

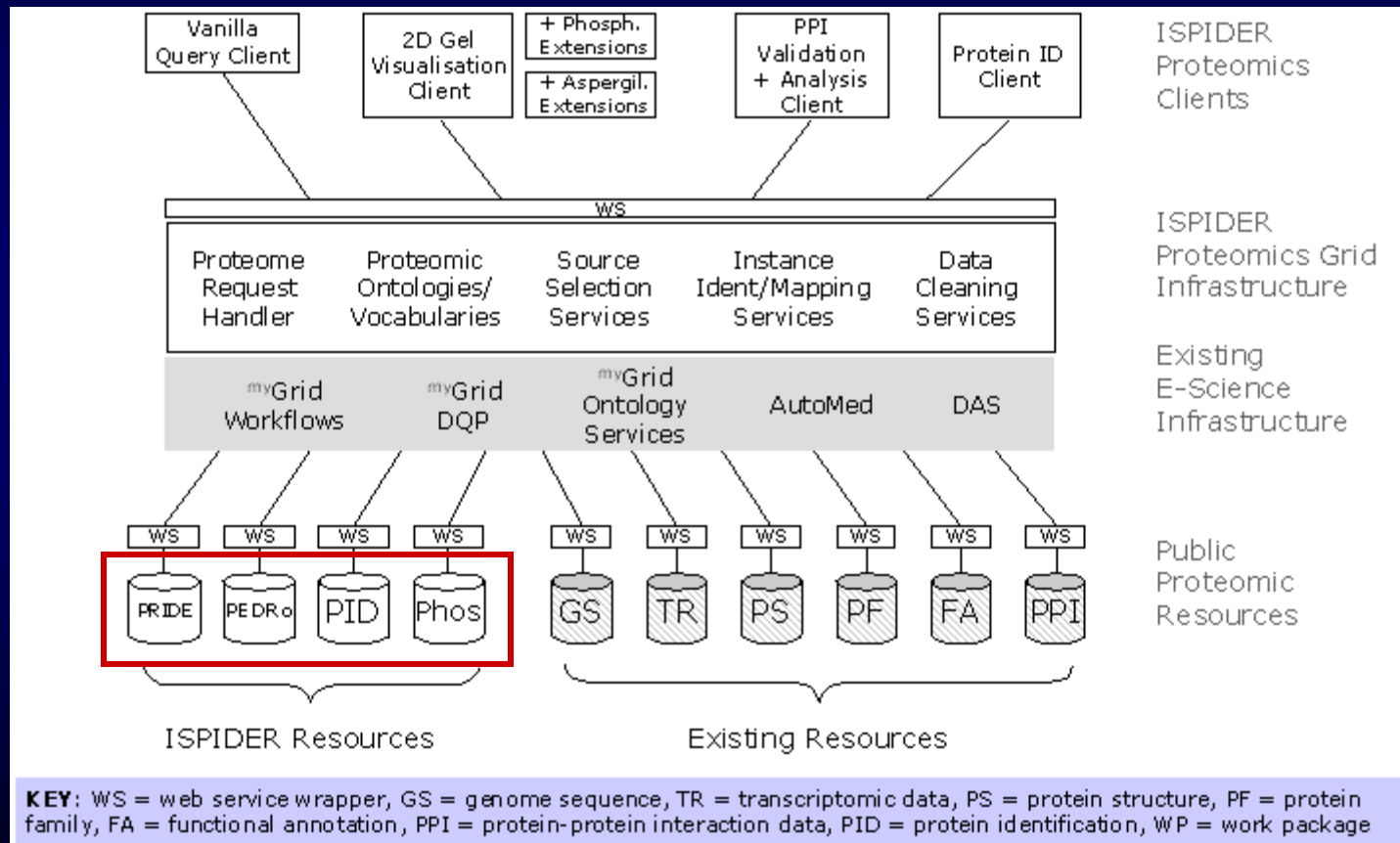
Project Challenges

- Data are overlapping and heterogeneous
- Data rapidly updated/modified/evolved
- Physical distance between repositories
- Need for processing power

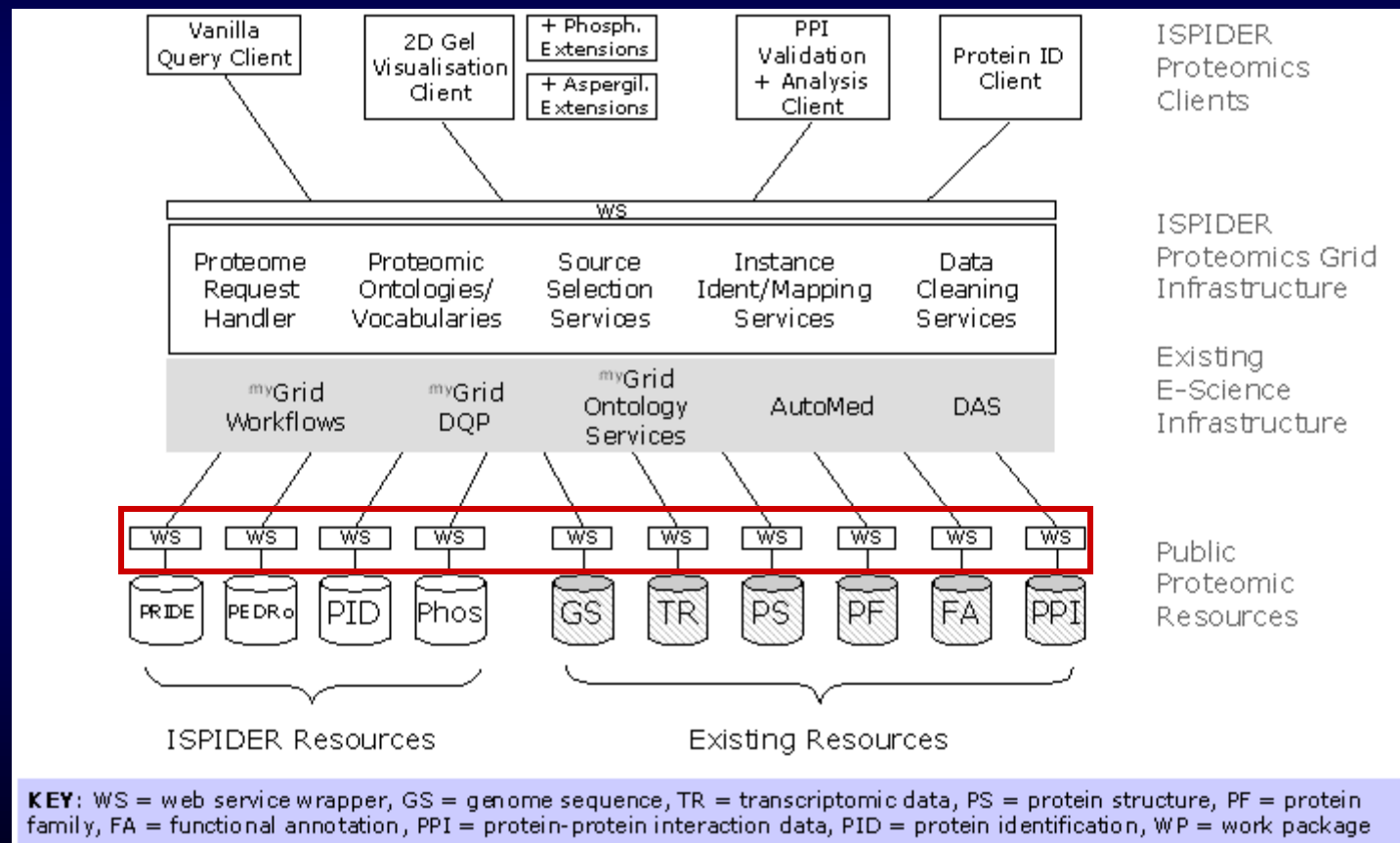
ISPIDER Objectives



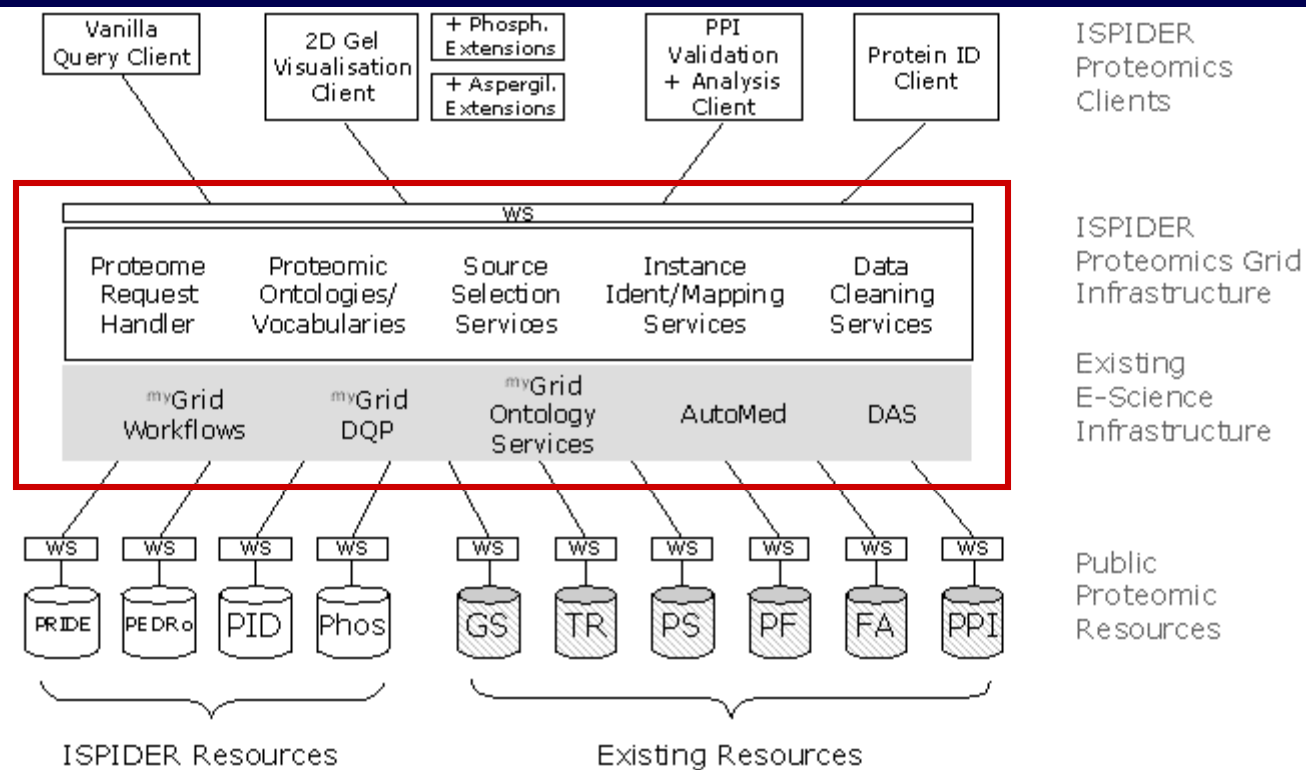
ISPIDER Objectives



ISPIDER Objectives

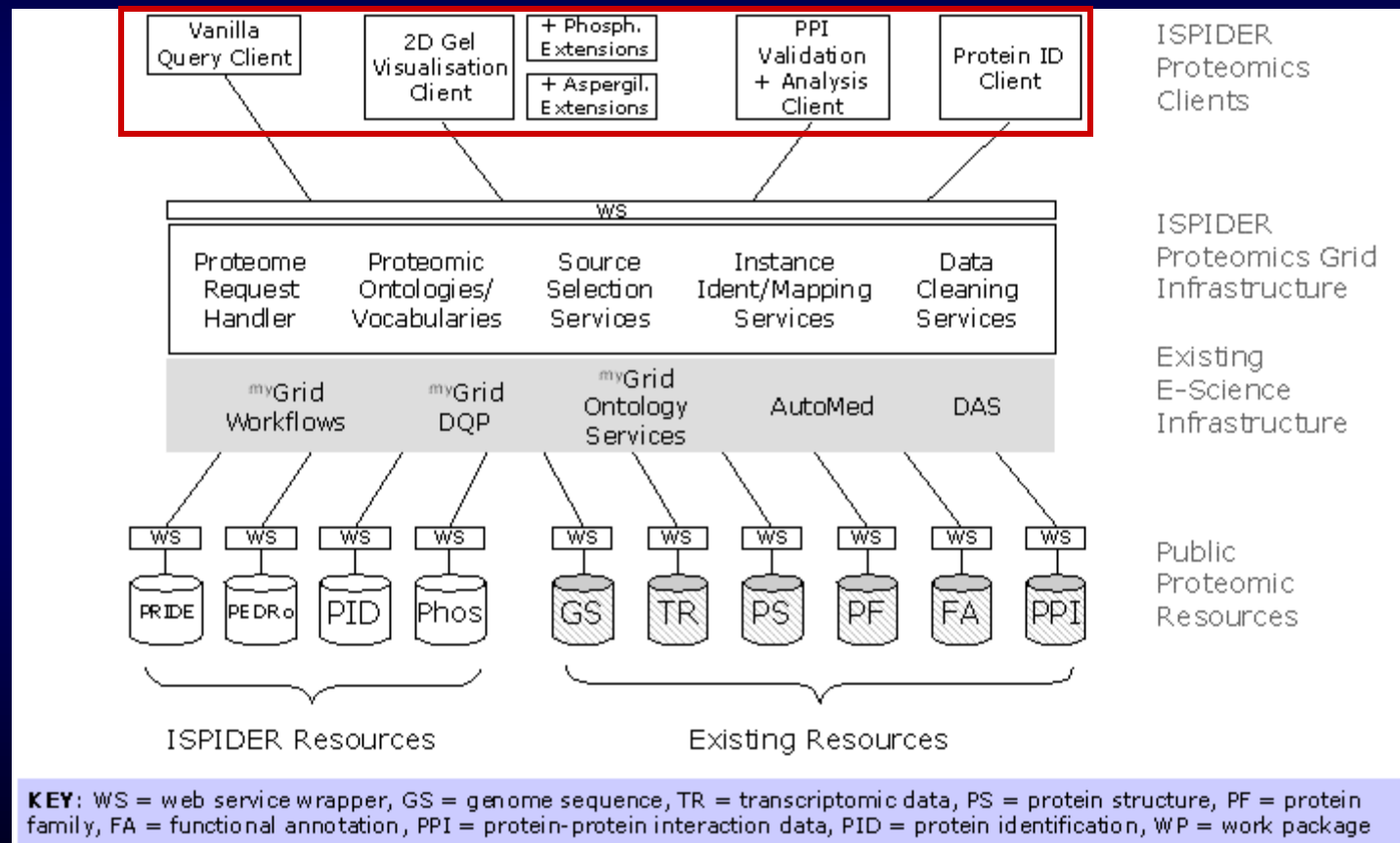


ISPIDER Objectives



KEY: WS = web service wrapper, GS = genome sequence, TR = transcriptomic data, PS = protein structure, PF = protein family, FA = functional annotation, PPI = protein-protein interaction data, PID = protein identification, WP = work package

ISPIDER Objectives

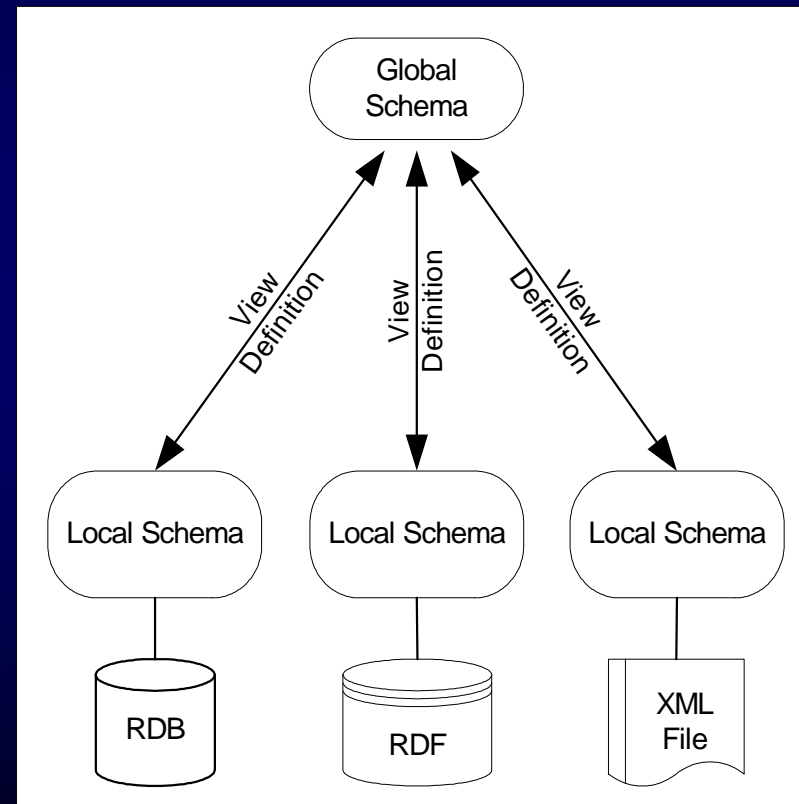


Overview

- Project overview
- Data Integration: the AutoMed project
- The OGSA-DAI project
- Project Implementation

Data Integration

- Global-As-View (GAV) approach: describe GS constructs with view definitions over LS_i constructs
- Local-As-View (LAV) approach: describe LS_i constructs with view definitions over GS constructs



GAV Example

S_g student(id, name, left#, degree)
 monitors(sno, id)
 staff(sno, sname, dept#)

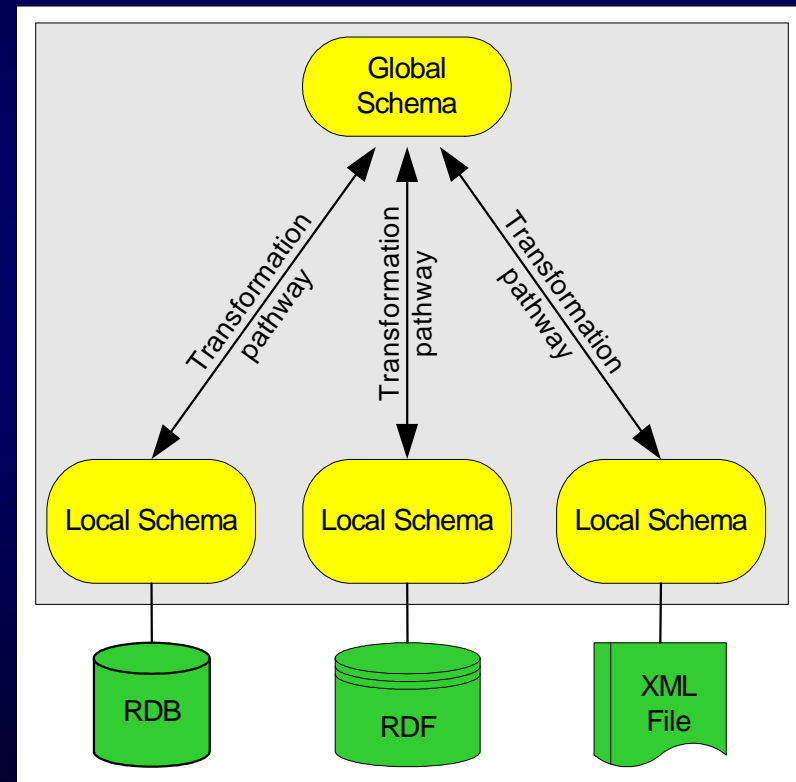
S_1 ug(id, name, left#, degree, sno)
 tutor(sno, sname)

S_2 phd(id, name, left#, title)
 supervises(sno, id)
 supervisor(sno, sname, dept)

- student(id, name, left, degree) =
 $[\{x, y, z, w\} \mid \langle x, y, z, w, _ \rangle \in \text{ug}$
 $\vee \langle x, y, z, _ \rangle \in \text{phd}$
 $\wedge w = \text{'phd'}]$
- monitors(sno, id) =
 $[\{x, y\} \mid (\langle y, _, _, _, x \rangle \in \text{ug}$
 $\wedge \langle x, _, _, _ \rangle \notin \text{phd})$
 $\vee \langle x, y \rangle \in \text{supervises}]$
- staff(sno, sname, dept) =
 $[\{x, y, z\} \mid \langle x, y, z \rangle \in \text{supervisor}$
 $\vee \langle x, y \rangle \in \text{tutor}$
 $\wedge \langle x, _, _ \rangle \notin \text{supervisor}]$

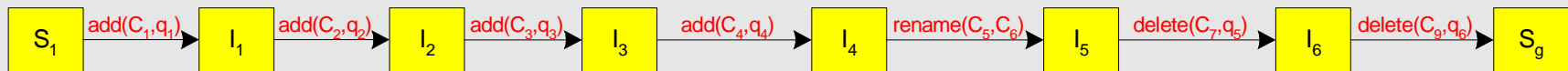
Both-As-View (BAV) Approach

- Schema transformation approach
- For each pair (LS_i, GS) : incrementally modify LS_i/GS to match GS/LS_i

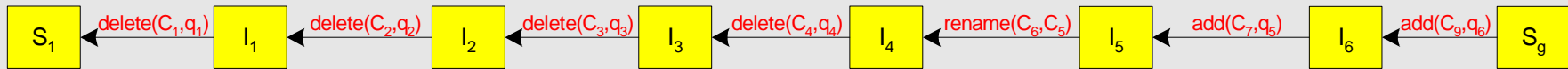


BAV Example

$S_1 \dot{\rightarrow} S_g$



$S_1 \dot{\leftarrow} S_g$



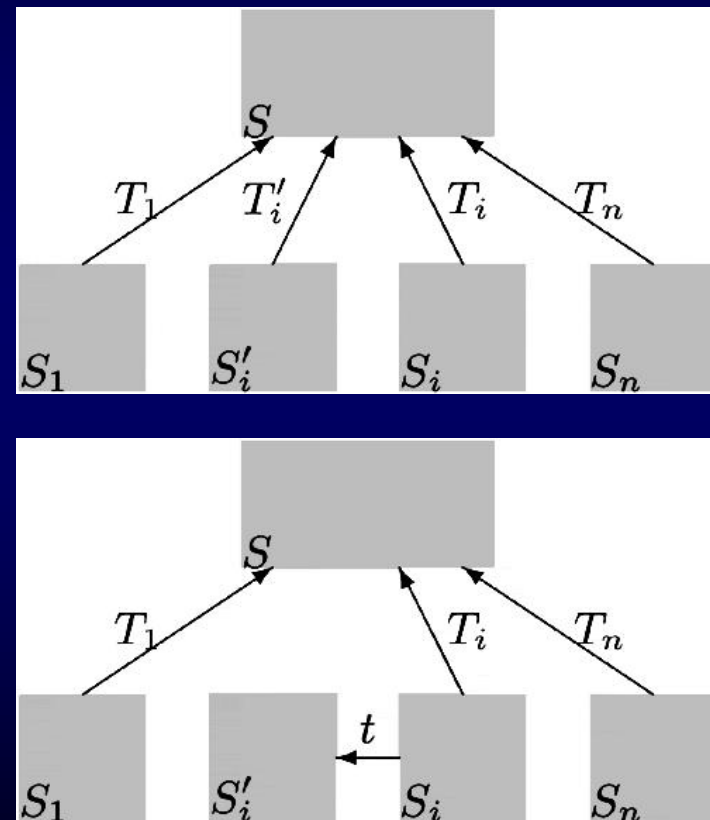
- Transformation pathway consists of primitive transformations
- Pathway contains both GAV & LAV definitions
- Transformations are automatically reversible
- Metadata in AutoMed Repository

AutoMed Toolkit

- Heterogeneous data integration system
- Birkbeck College & Imperial College
- AutoMed advantages
 - Subsumes traditional approaches
 - Handles heterogeneity – easily extensible
 - Virtual/materialised/hybrid integration
 - Schema evolution
 - Tools: data warehousing, schema matching, semi-automatic XML transformation/integration

Schema Evolution Example

- Define the evolution of the global or local schema as a schema transformation pathway from the old to the new schema



Overview

- Project overview
- Data Integration: the AutoMed project
- The OGSA-DAI project
- Project Implementation

Grids

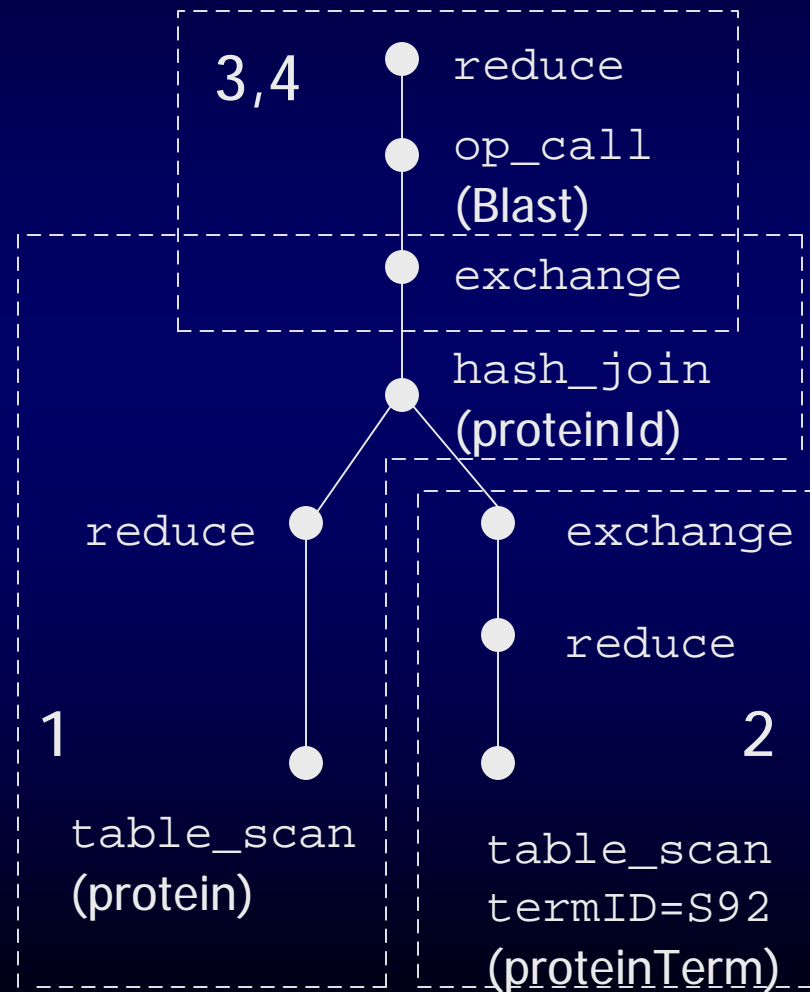
- What are Grids and why do we need them?
 - Collaborative research that is made possible by the sharing across the Internet of resources (data, instruments, computation, people's expertise...)
 - ISPIDER scope
- U.K. effort: OGSA-DAI
 - Open Grid Services Architecture – Data Access & Integration
 - Open Source
 - Service-Oriented Architecture (SOA)
 - Data Access
 - Data Integration

OGSA-DAI

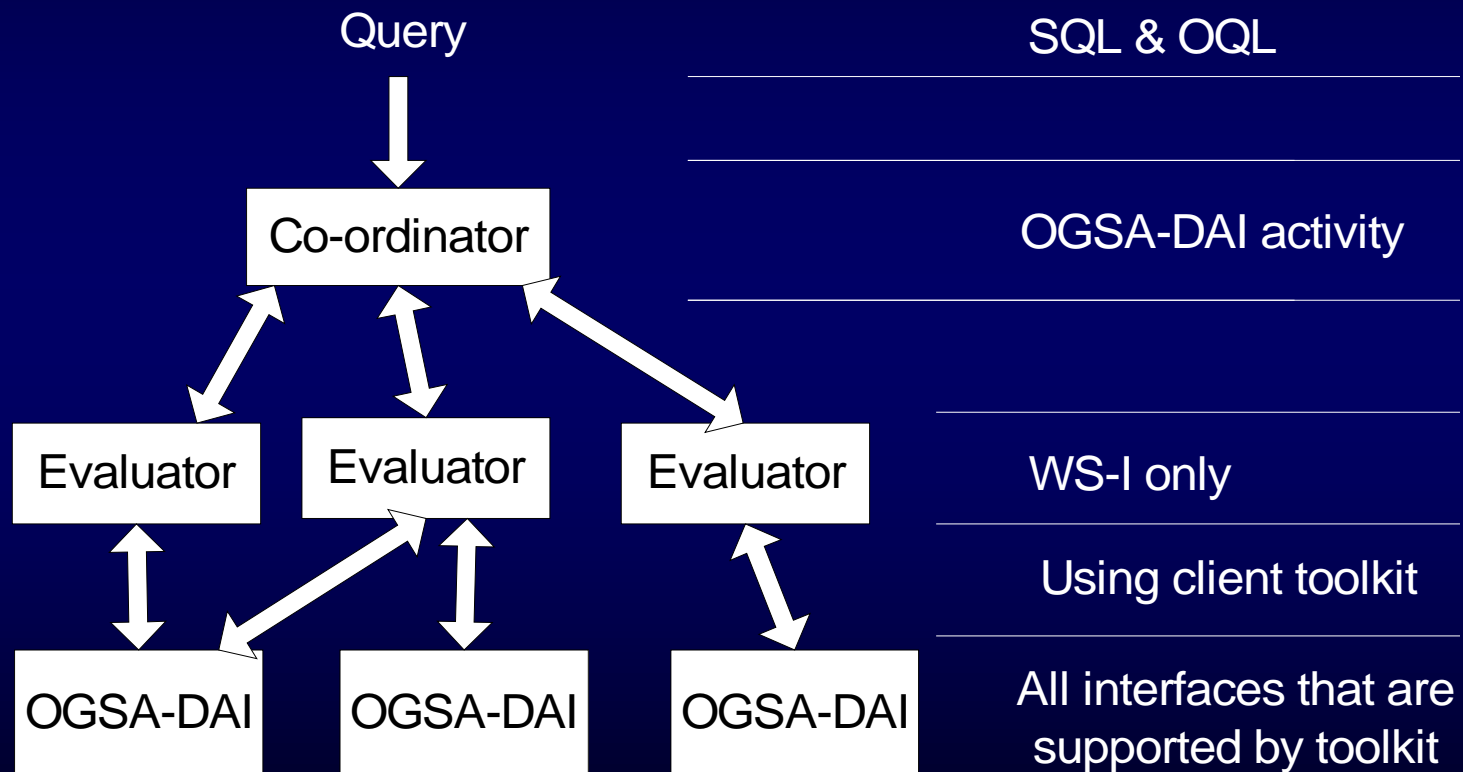
- A framework for building applications
- Supports data access, insert and update
 - Relational: MySQL, Oracle, DB2, SQL Server, PostgreSQL
 - XML: Xindice, eXist
 - Files – CSV, BinX, EMBL, OMIM, SWISSPROT,...
- Supports data delivery
 - SOAP over HTTP
 - FTP; GridFTP
 - E-mail
 - Inter-service
- Supports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

OGSA-DQP

- Distributed Query Processor
- Implicit parallelism
- Execution
 - Queries mapped to algebraic expressions for evaluation
 - Parallelism represented by partitioning queries



DQP architecture

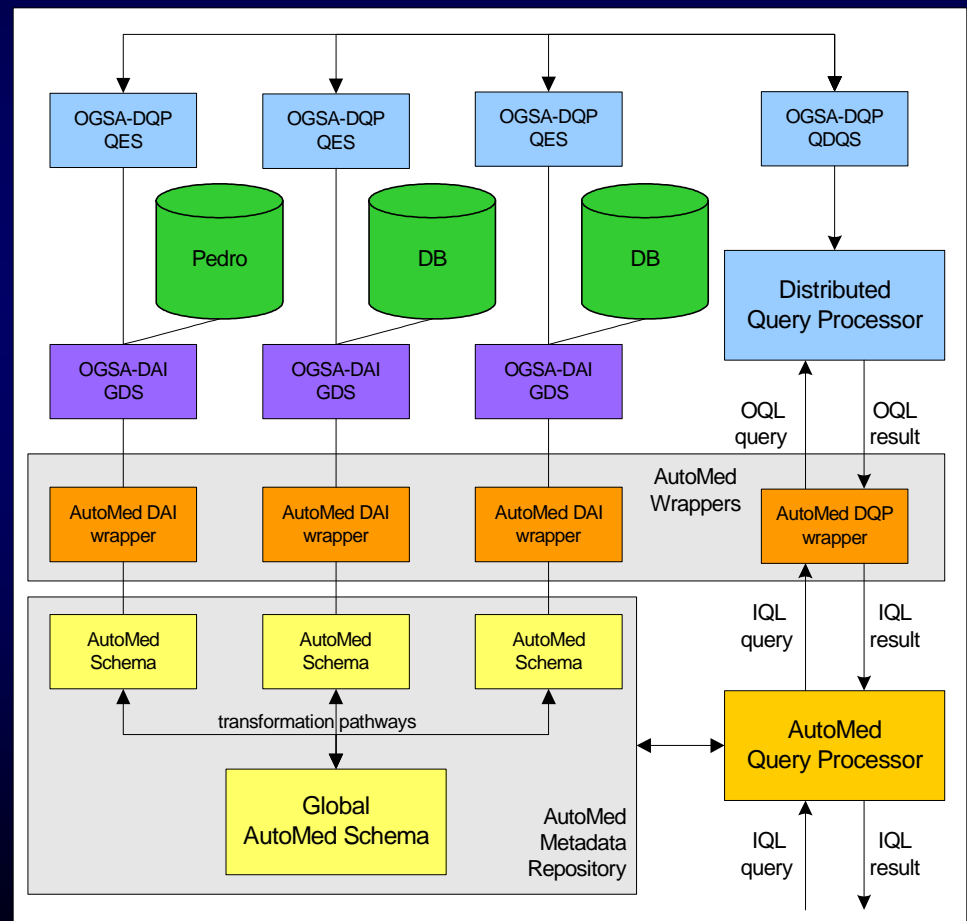


Overview

- Project overview
- Data Integration: the AutoMed project
- The OGSA-DAI project
- Project Implementation

Interoperability

- Sources wrapped with OGSA-DAI
- AutoMed wrappers extract source metadata
- Integration using AutoMed
- Queries submitted:
 - Reformulated using AutoMed metadata
 - Submitted to OGSA- DQP



Future Work

- AutoMed extensions:
 - Web/Grid Services for AutoMed
 - Data warehousing
 - Materialised/hybrid integration
 - Data provenance
 - Incremental view maintenance
 - Schema evolution

Summary

- ISPIDER aims to:
 - Build an integrated platform of proteomic resources
 - Use existing resources – produce new ones
 - Create clients for querying, visualisation, etc.
- ISPIDER is using:
 - myGrid – middleware for biological experiments
 - AutoMed – heterogeneous data integration system
 - OGSA-DAI – middleware for exposing resources on the Grid via web services
 - OGSA-DQP – distributed query processor

ISPIDER Project Members

- Birkbeck College
 - Nigel Martin
 - Alex Poulouvassilis
 - Lucas Zamboulis (R.A.)
 - Hao Fan (former R.A.)
- European Bioinformatics Institute
 - Rolf Apweiler
 - Henning Hermjakob
 - Weimin Zhu
 - Chris Taylor
 - Phil Jones
 - Nisha Vinod
- University of Manchester
 - Simon Hubbard
 - Steve Oliver
 - Suzanne Embury
 - Norman Paton
 - Carol Goble
 - Robert Stevens
 - Khalid Belhajjame (R.A.)
 - Jennifer Siepen (R.A.)
- U.C.L.
 - David Jones
 - Christine Orengo
 - Melissa Pentony (R.A.)