



Grid-Based Integration of Biological
Data Using AutoMed

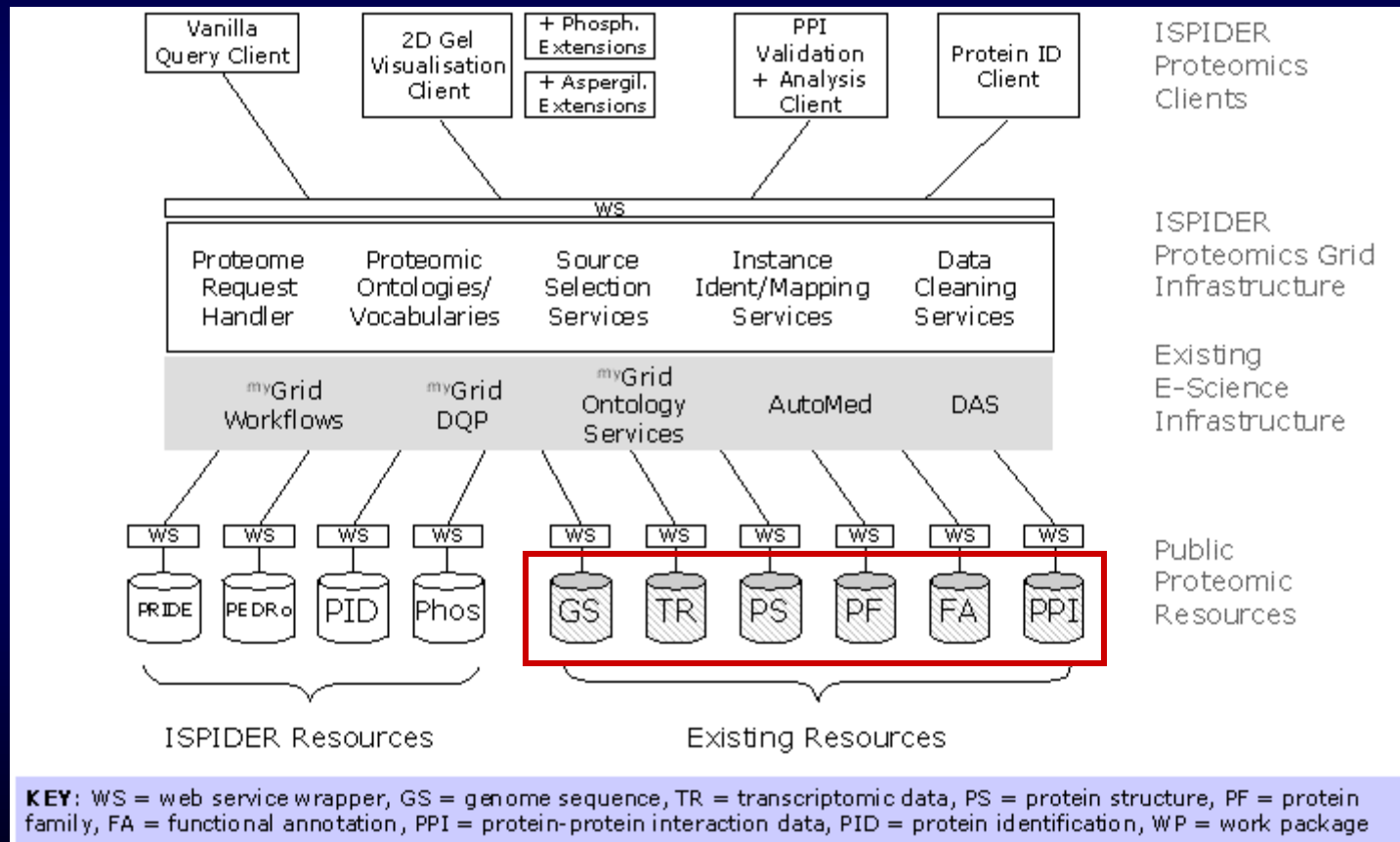
Project Details

- Members
 - Birkbeck College
 - European Bioinformatics Institute
 - University of Manchester
 - University College London

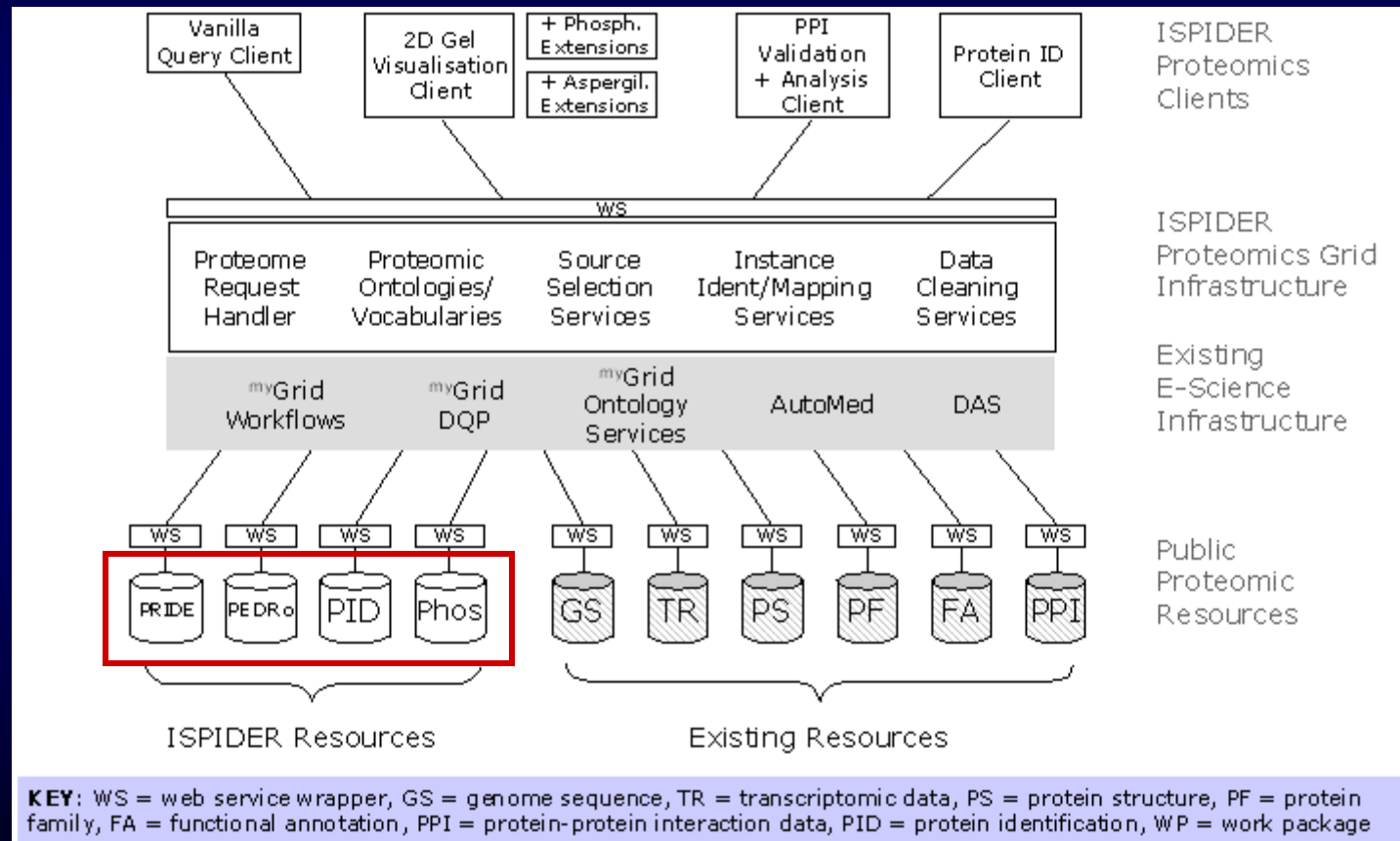
Problem Definition

- Vast biological data
- Need for interoperability
- Need for processing power

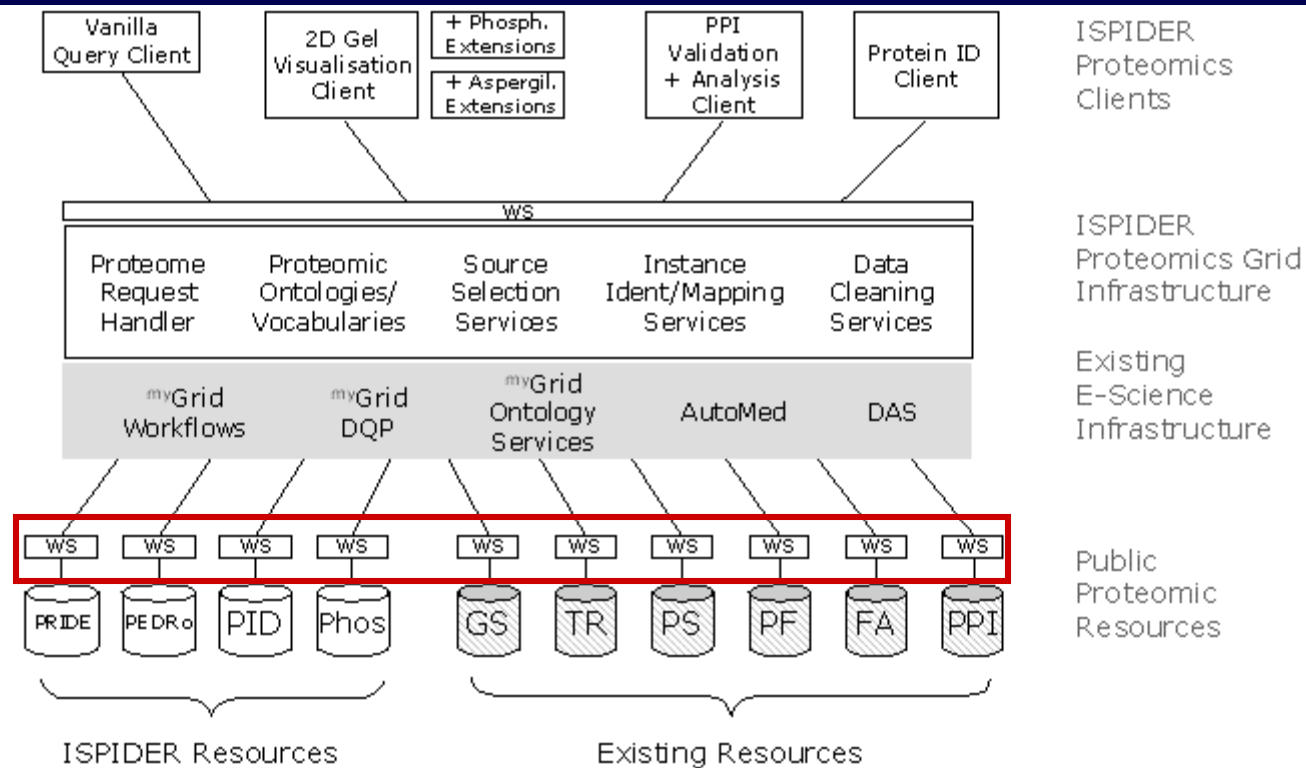
Project Aims



Project Aims

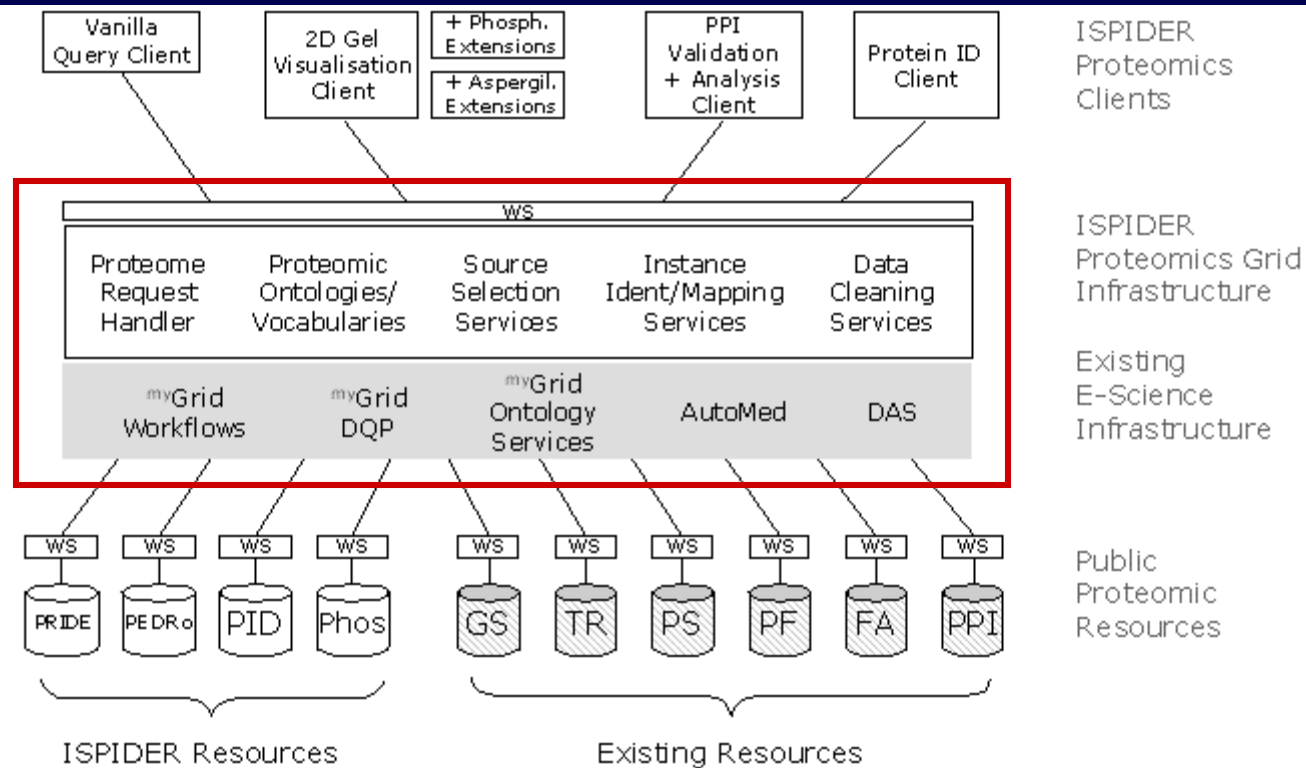


Project Aims



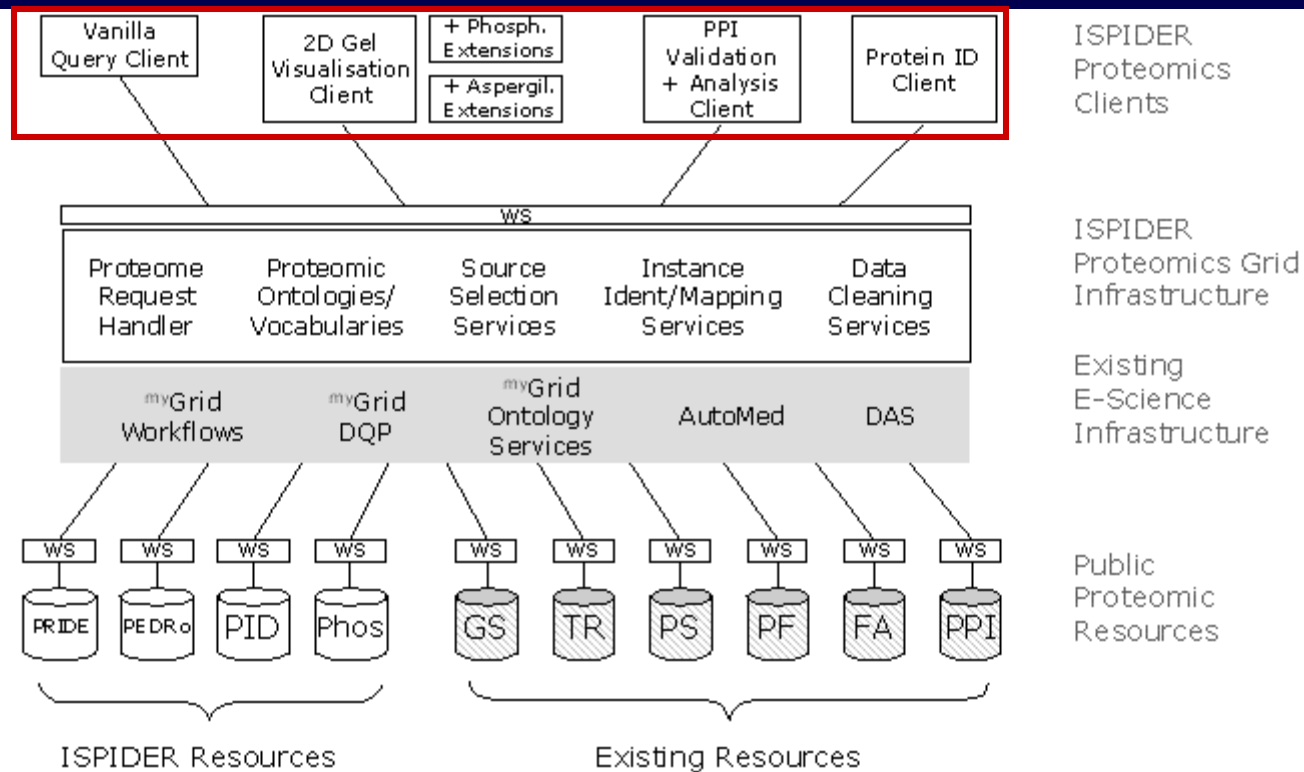
KEY: WS = web service wrapper, GS = genome sequence, TR = transcriptomic data, PS = protein structure, PF = protein family, FA = functional annotation, PPI = protein-protein interaction data, PID = protein identification, WP = work package

Project Aims



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

Project Aims



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

myGrid – DQP

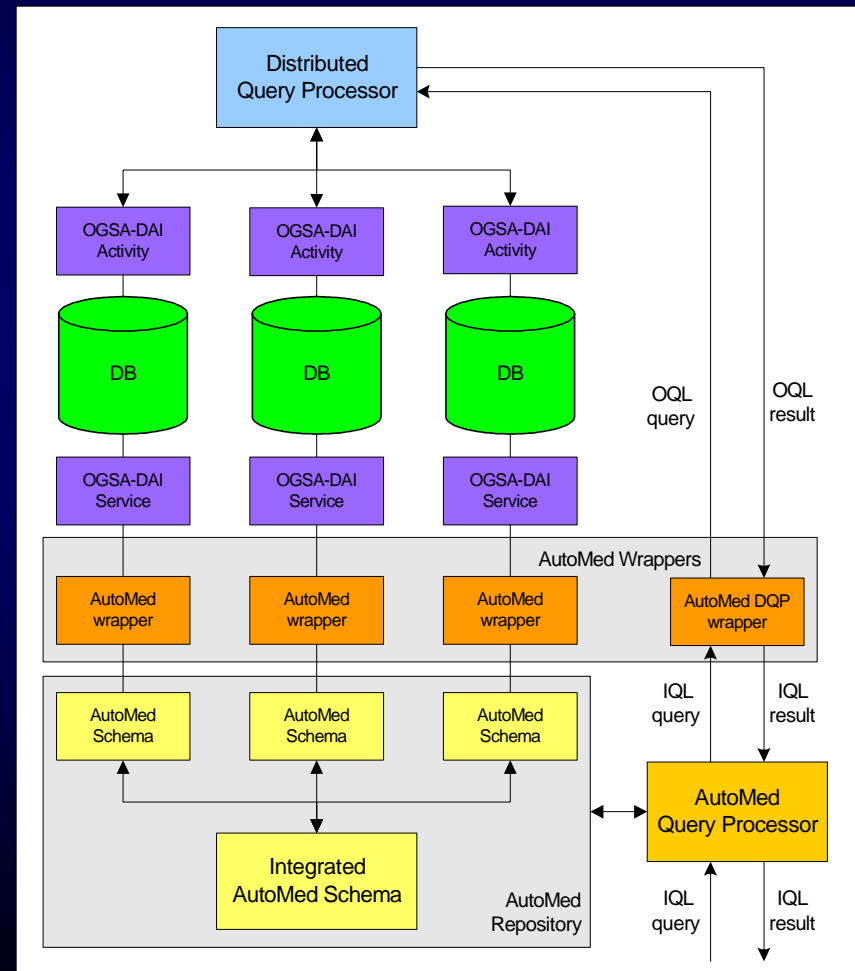
- myGrid: collection of services/components allowing high-level integration of data/applications
- DQP
 - OGSA-DAI (Open Grid Services Architecture Data Access and Integration)
 - Why DQP?
 - AutoMed – DQP cooperation

AutoMed Toolkit

- Heterogeneous data integration system - developed by Birkbeck College/Imperial College
- Why AutoMed?
 - Powerful modelling capabilities
 - Handles various data models – easily extensible
 - Virtual/materialised/ hybrid integration
 - Schema evolution

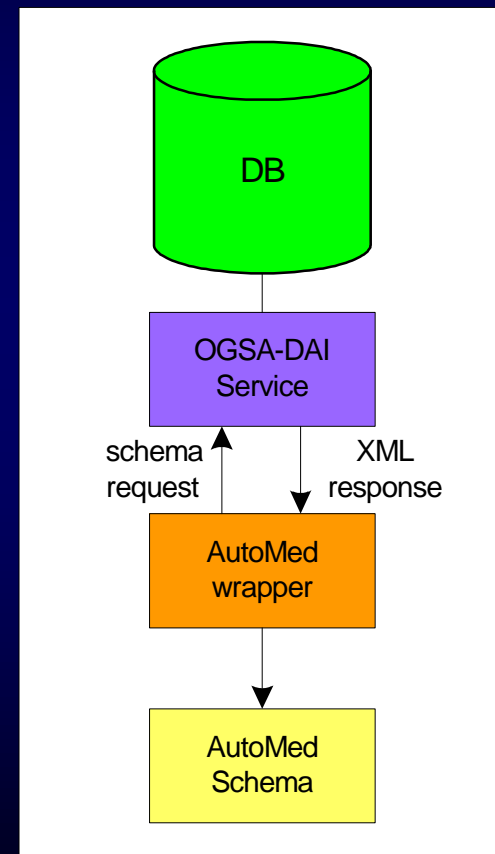
Interoperability

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



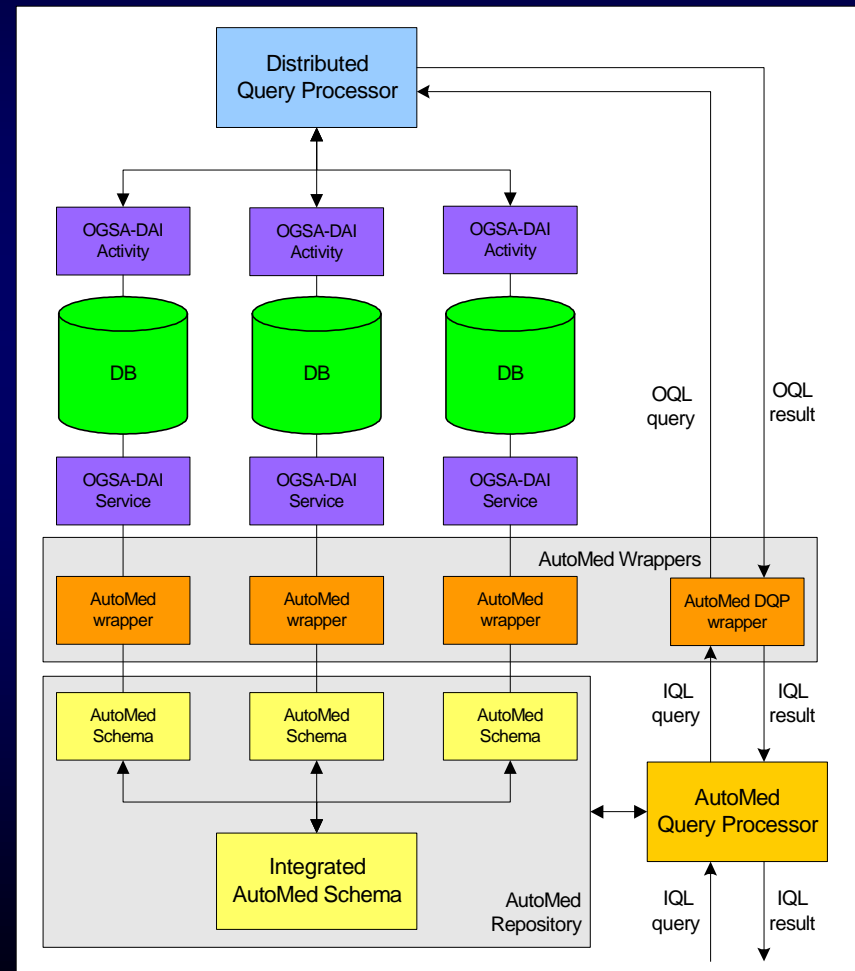
Schema extraction

- AutoMed wrapper requests the schema of the data source using an OGSA-DAI service
- The service replies with the source schema encoded in XML
- The AutoMed wrapper creates the corresponding schema in the AutoMed repository



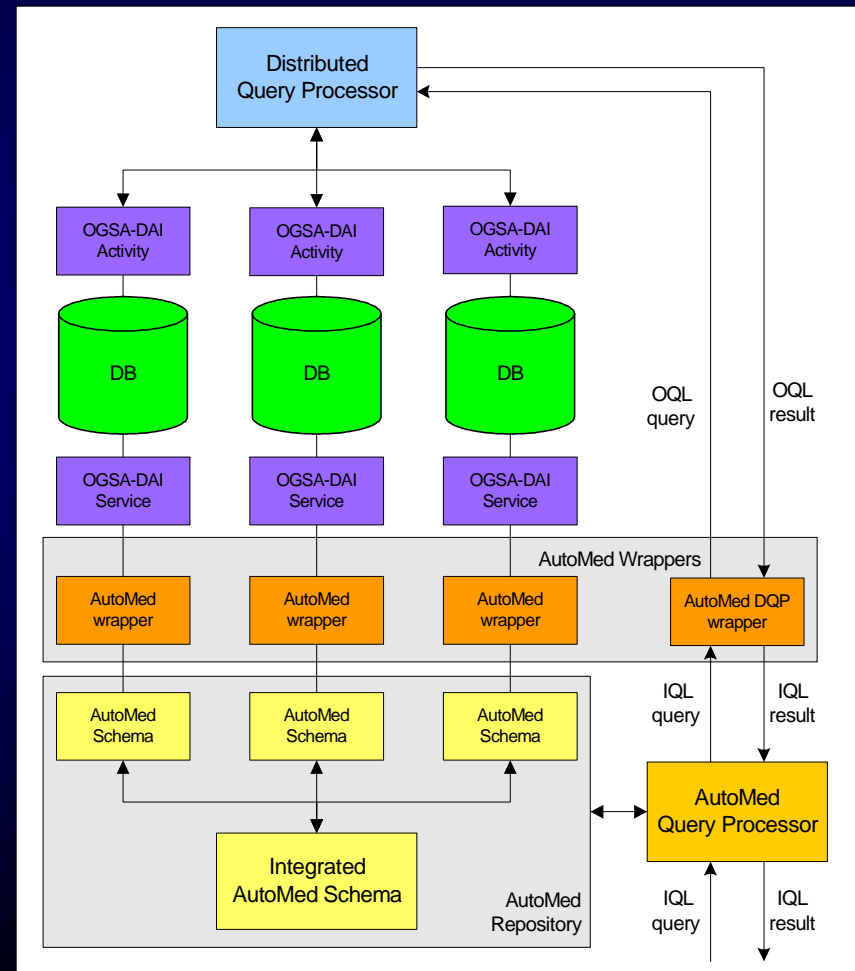
Query Processing

- Query is:
 - Submitted to AutoMed's GQP
 - Reformulated
 - Optimised
 - Translated from IQL into OQL
 - Submitted to DQP



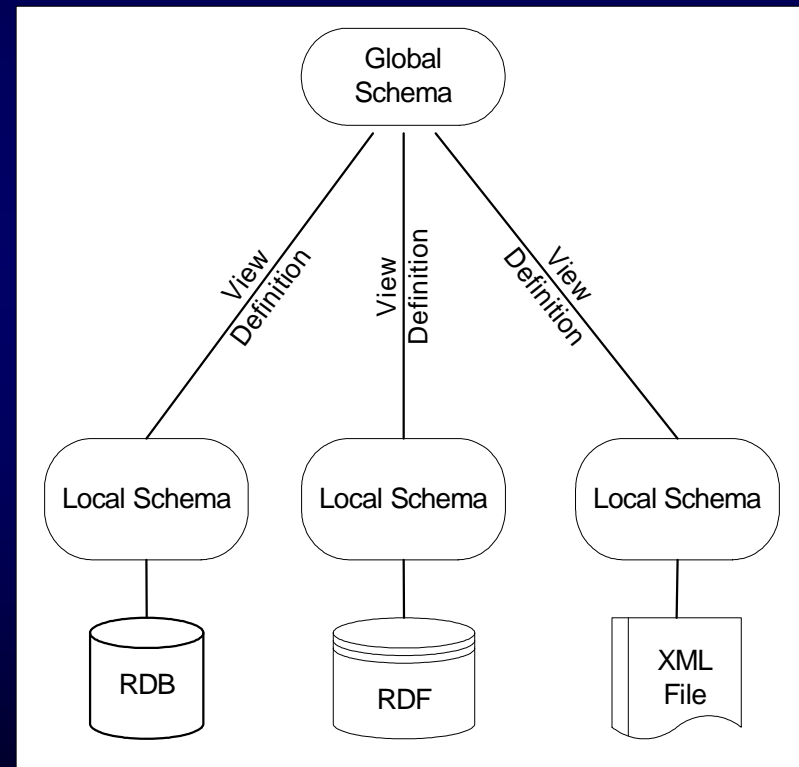
Query Processing

- DQP
 - Evaluates query using OGSA-DAI activities
 - Sends the result to AutoMed's GQP



GAV & LAV Approaches

- 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#)

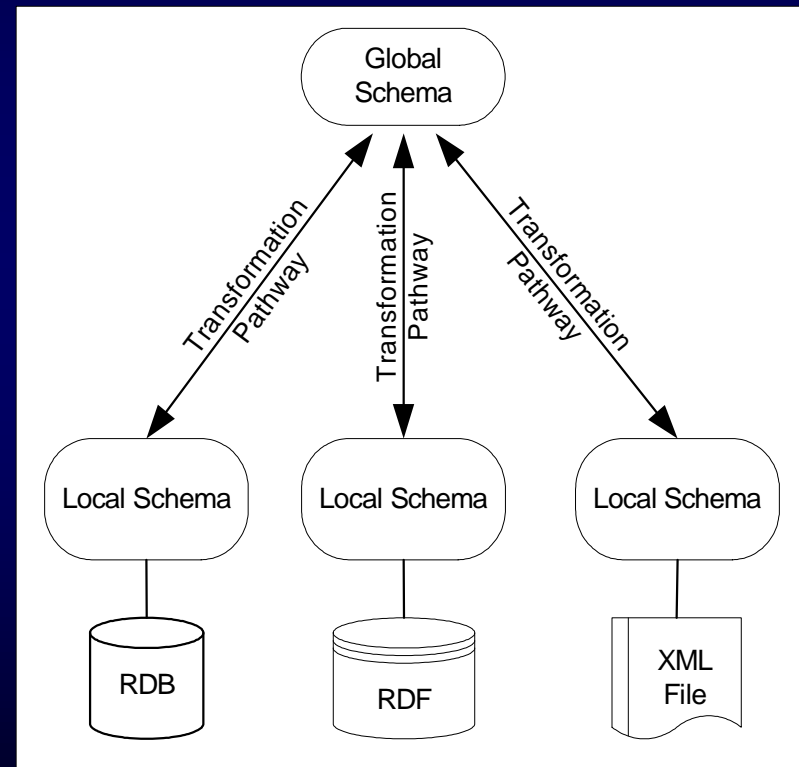
S_1 ug(id, name, left#, degree, sno)
 tutor(sno, sname)

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

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

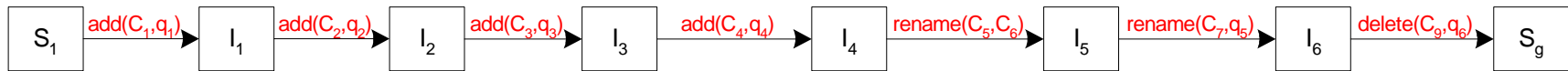
Both-As-View (BAV)

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

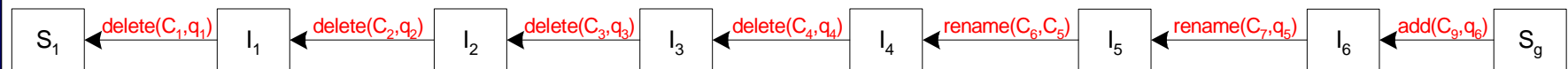


BAV Example

$S_1 \dot{\rightarrow} S_g$



$S_1 \dot{\leftarrow} S_g$



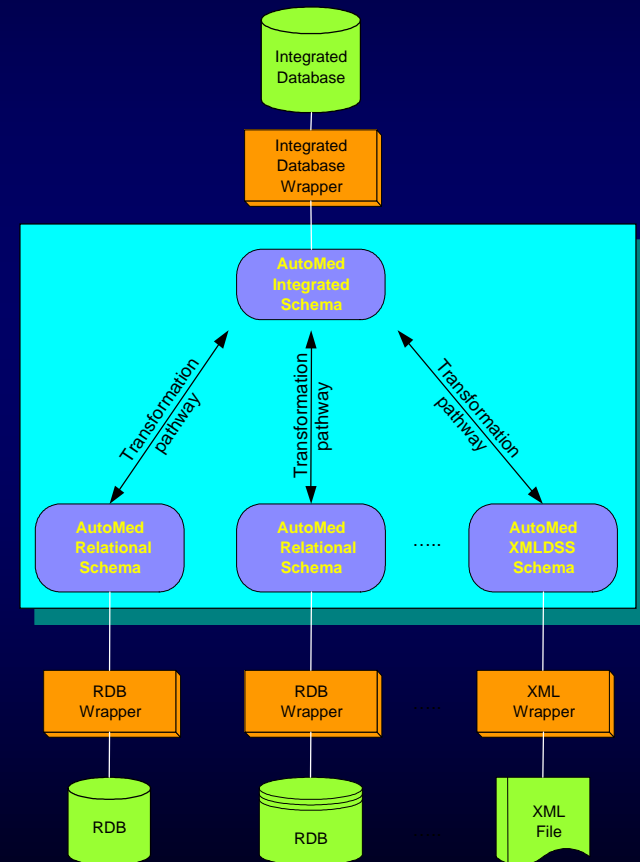
- Transformation pathway consists of primitive transformations
- Pathway contains both GAV & LAV definitions
- Transformations are automatically reversible

BioMap Integration

- Relational/XML sources - relational global schema
 - Wrapping of sources
 - Translation of source and global schemas into the XML schema type used within AutoMed
 - Domain expert provides mappings between sources & global schema
 - Automatic schema transformation/integration algorithm

- DILS'05

www.doc.ic.ac.uk/automed



Summary

- ISPIDER aims to:
 - Create an integrated platform of proteomic resources
 - Use existing resources – produce new ones
 - Create clients for querying, visualisation, etc.
- ISPIDER is using:
 - myGrid – middleware for in silico experiments in biology
 - OGSA-DQP – service-based distributed query processor
 - AutoMed – heterogeneous data integration system

Project Members

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