

A Uniform Approach to Data and Workflow Integration for the Life Sciences

L. Zamboulis^{1,2,3}, N. Martin^{1,3}, A. Poulouvasilis^{1,3}
{lucas,nigel,ap}@dcs.bbk.ac.uk

¹School of Computer Science & Inf. Systems, Birkbeck

²Dept. of Biochemistry and Molecular Biology, UCL

³London Knowledge Lab

Projects

- AutoMed (EPSRC: 2001-2003 – still running)
 - Birkbeck and Imperial College
 - Framework for the integration of heterogeneous data sources (mediator approach)
- ISPIDER (BBSRC: 2005-2007)
 - Birkbeck, UCL, Manchester, EBI
 - Integrated Grid platform of proteomic resources

Outline

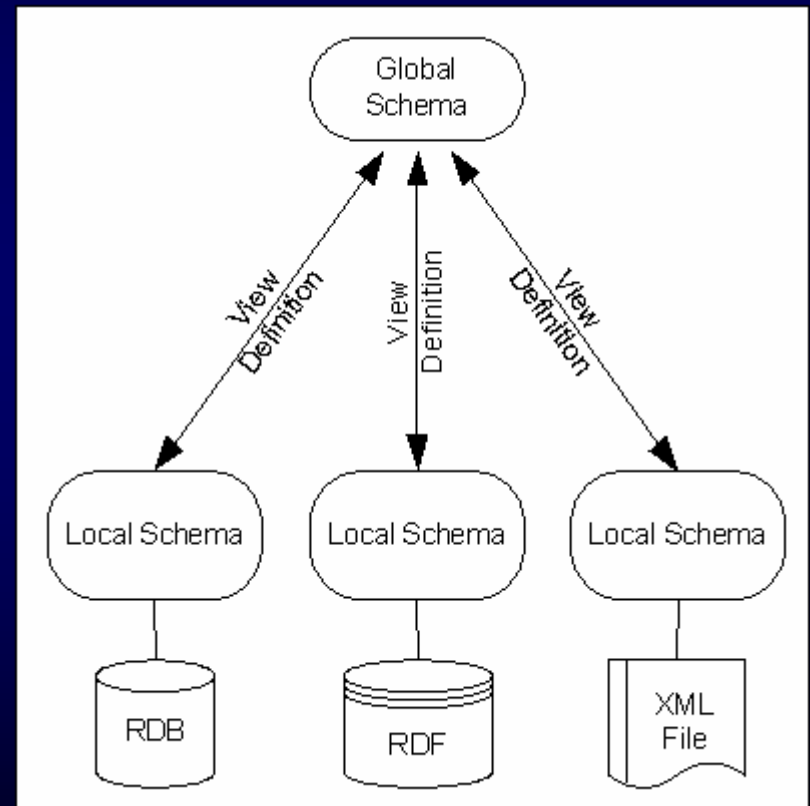
- AutoMed
 - Data integration approaches
 - The BAV approach
 - The AutoMed system
- ISPIDER
 - BioMap integration
 - ISPIDER integration
 - AutoMed – Taverna interoperation

Problem Definition

- Data federation: unification of data sources
 - Schema is a union of data source schemas
 - Users must know details of all data sources
 - Queries need to integrate data fetched from the different data sources
 - Exists in current DBMSs
- Data integration: integration of data sources
 - Schema is an integration of data source schemas
 - Users only need to know the integrated schema
 - Data are integrated internally through mappings
 - Does not exist in current DBMSs

Data Integration

- Global-As-View (GAV): describe global constructs with view definitions over local constructs
- Local-As-View (LAV): describe local constructs with view definitions over global constructs



GAV Example

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

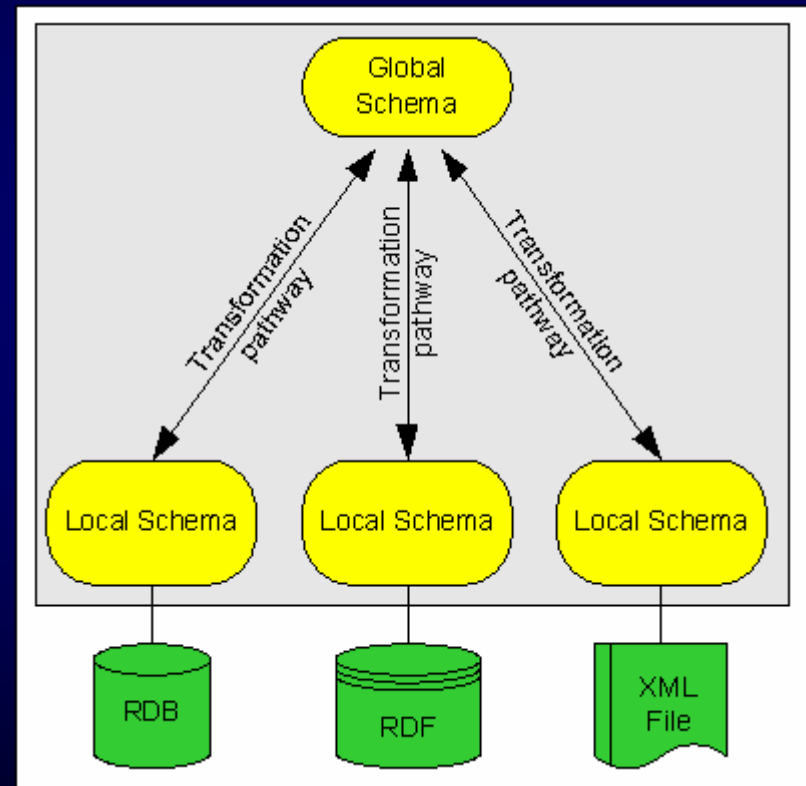
S_1 ug(id, name, left#, degree, *sno*)
 tutor(sno, sname)

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

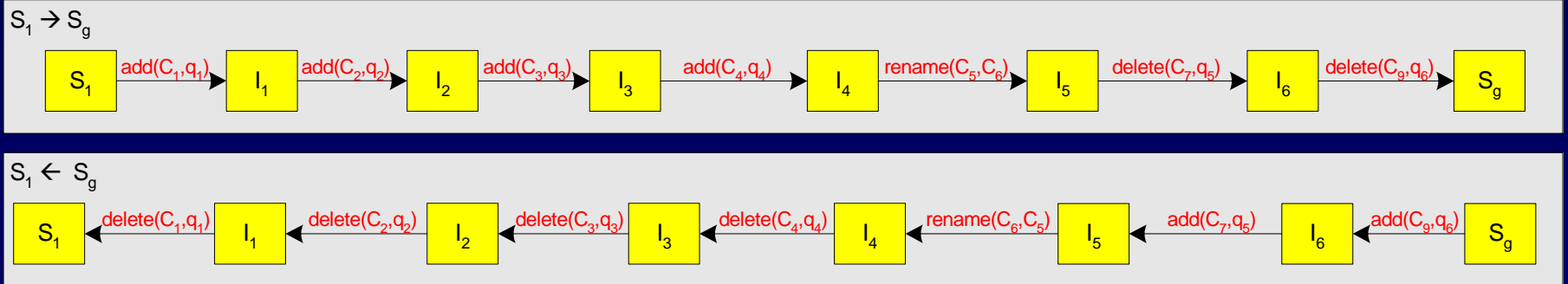
- student(id, name, left, degree) =
 $[\{x, y, z, w\} \mid \langle x, y, z, w, _ \rangle \in \text{ug}$
 $\vee \langle x, y, z, _ \rangle \in \text{phd}$
 $\wedge w = \text{'phd'}]$
- monitors(sno, id) =
 $[\{x, y\} \mid (\langle y, _, _, _, x \rangle \in \text{ug}$
 $\wedge \langle x, _, _, _ \rangle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



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

AutoMed

- Heterogeneous data integration system
- AutoMed advantages
 - Subsumes GAV, LAV and GLAV
 - Handles heterogeneity – easily extensible
 - Virtual/materialised/hybrid integration
 - Schema evolution
- Available components:
 - GUI
 - data warehousing (data provenance, incr. view maintenance)
 - schema matching (model independent)
 - schema evolution
 - semi-automatic XML transformation/integration
 - P2P infrastructure
 - Grid support: OGSA-DAI/DQP
 - Parallel & distributed query processing

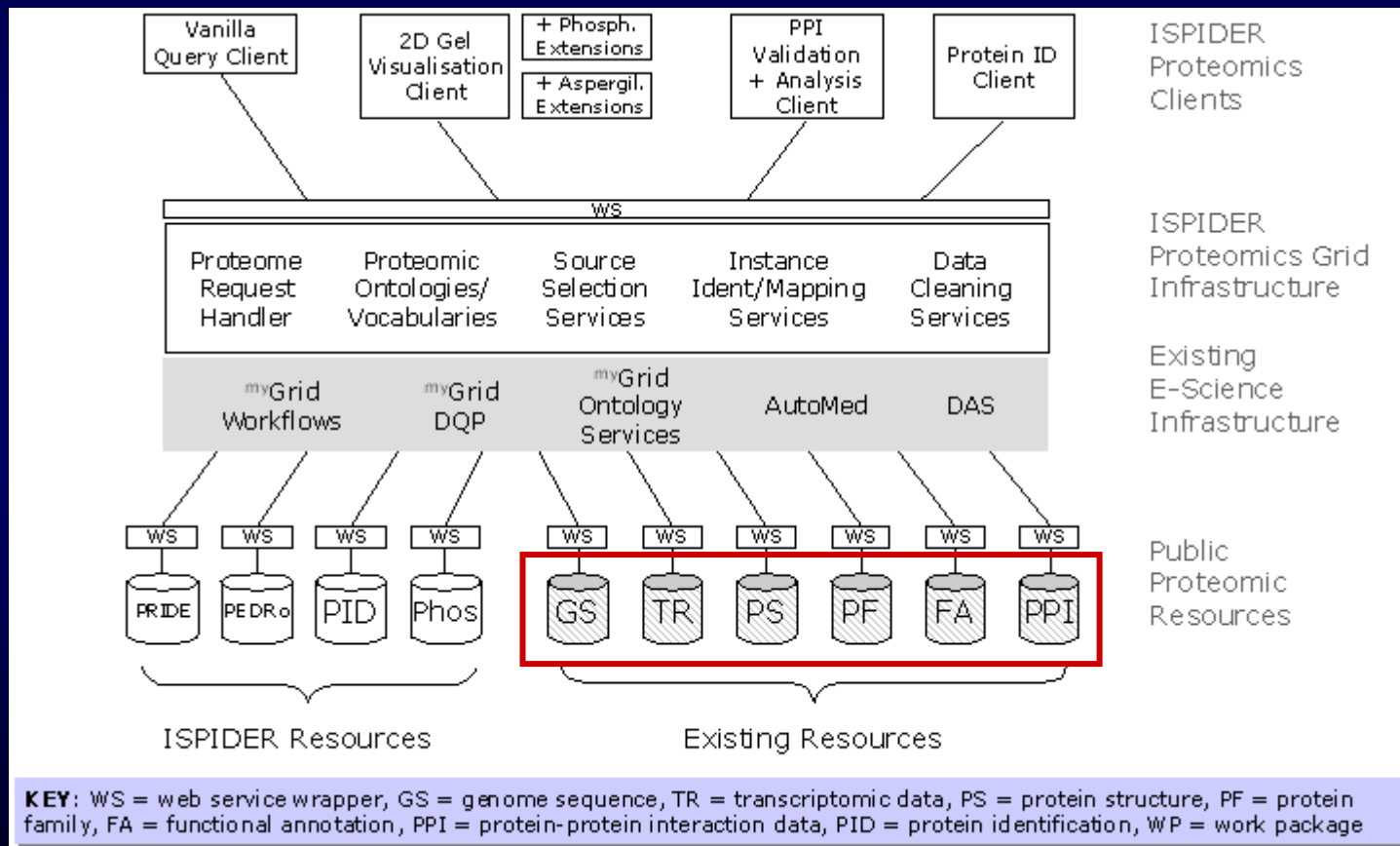
Outline

- AutoMed
 - Data integration approaches
 - The BAV approach
 - The AutoMed system
- ISPIDER
 - Review
 - BioMap integration
 - ISPIDER integration
 - AutoMed – Taverna interoperation

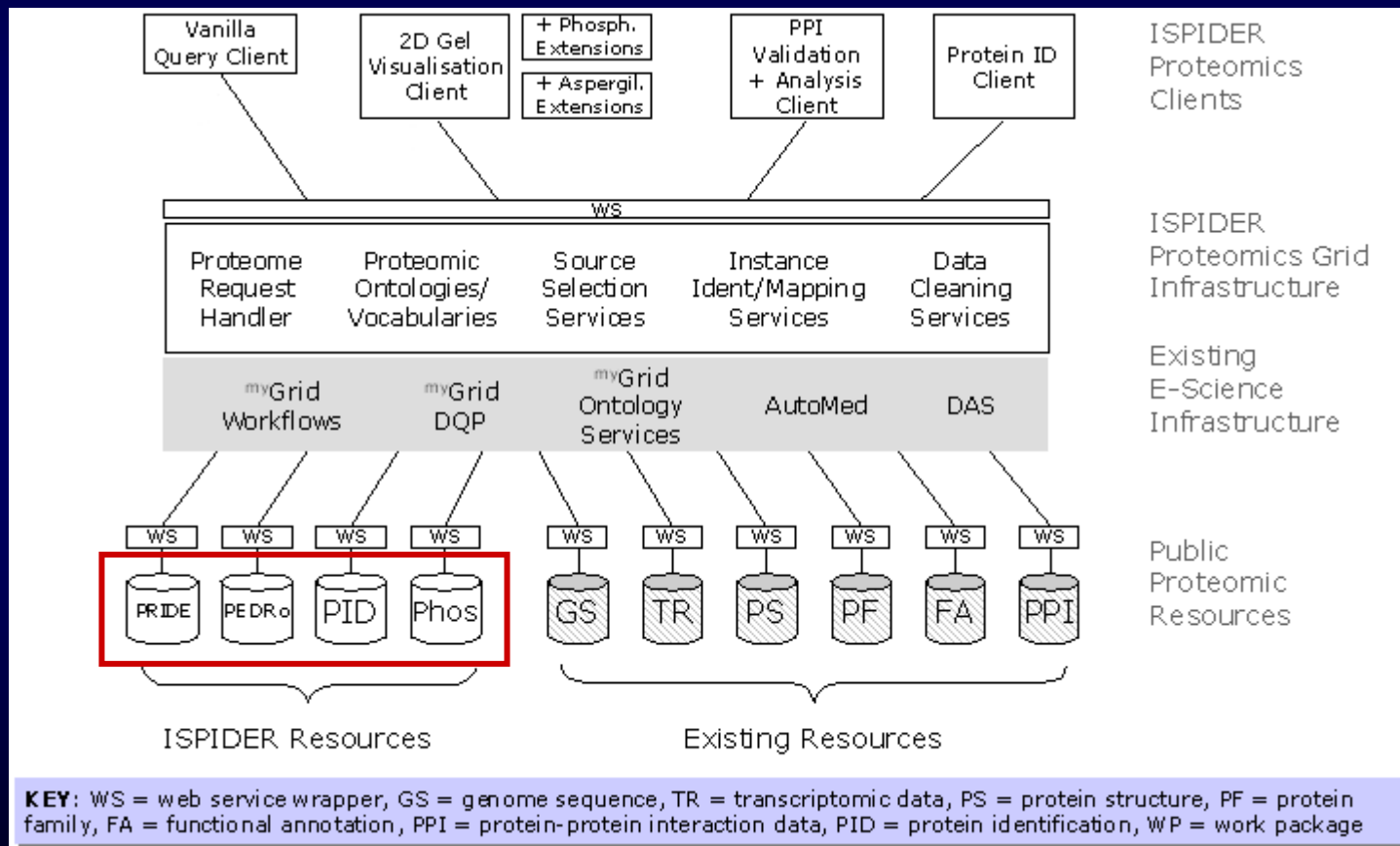
ISPIDER

- Produce an integrated platform for biologists
 - Laboratories across the world produce vast amounts of experimental data
 - Combining efforts will result in added value
- Challenges
 - Data resources: overlapping and heterogeneous
 - Data resources: rapidly updated and evolved
 - Physical distance between repositories
 - Need for efficient computation

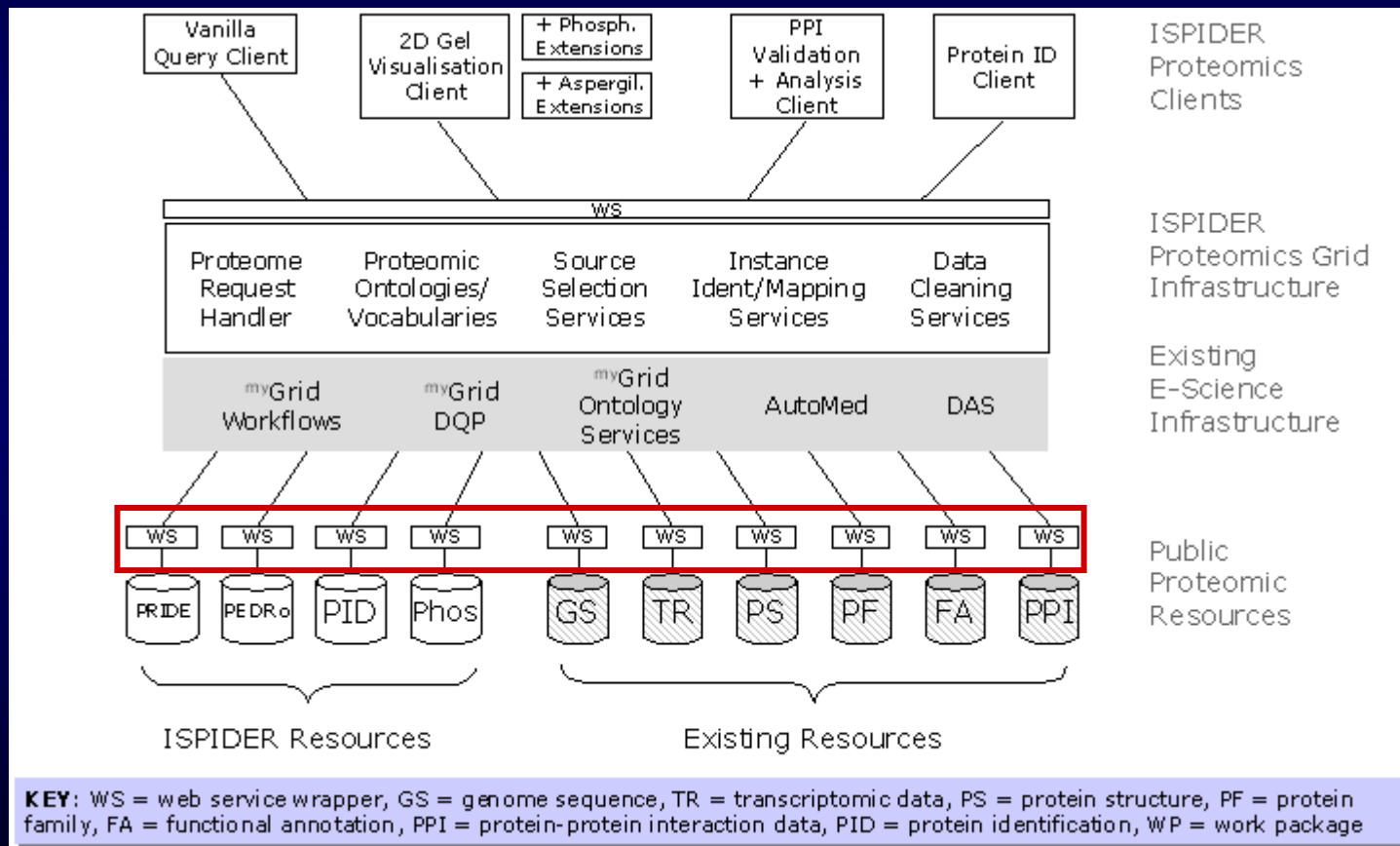
ISPIDER Objectives



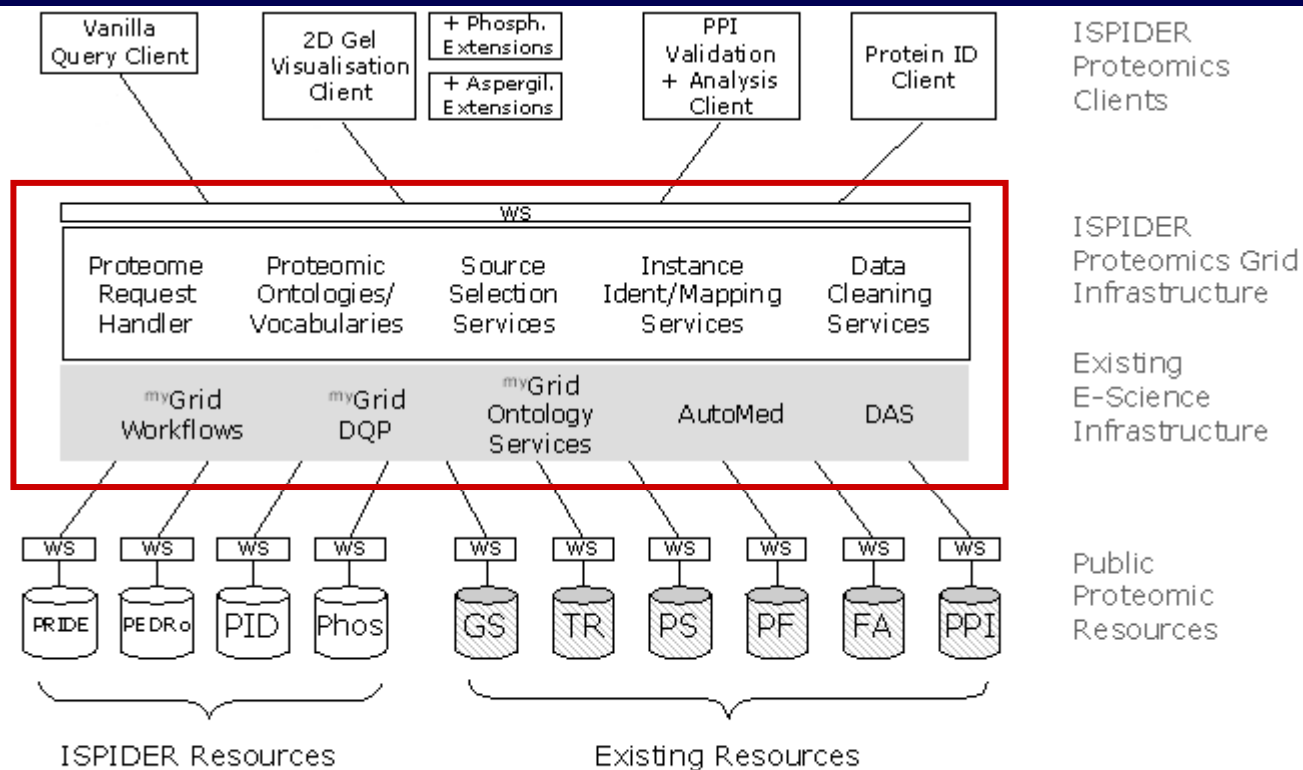
ISPIDER Objectives



ISPIDER Objectives

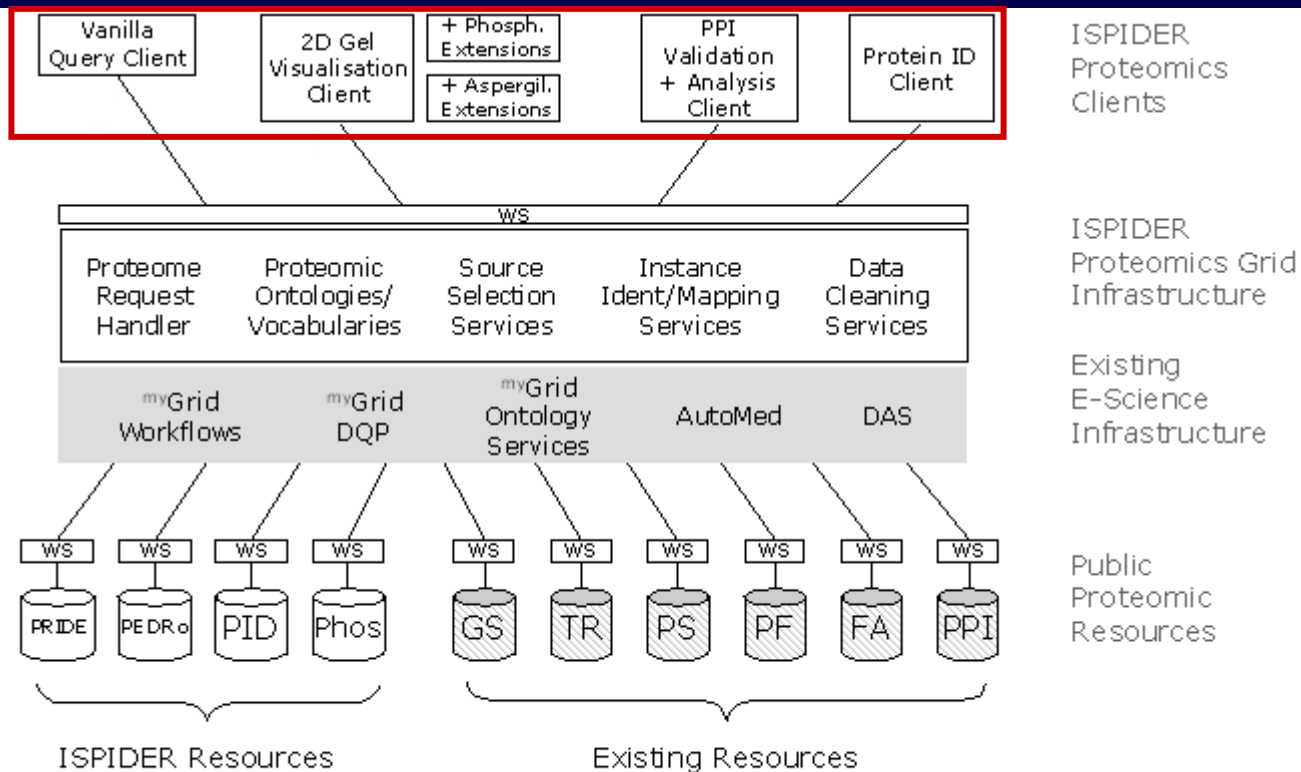


ISPIDER Objectives



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

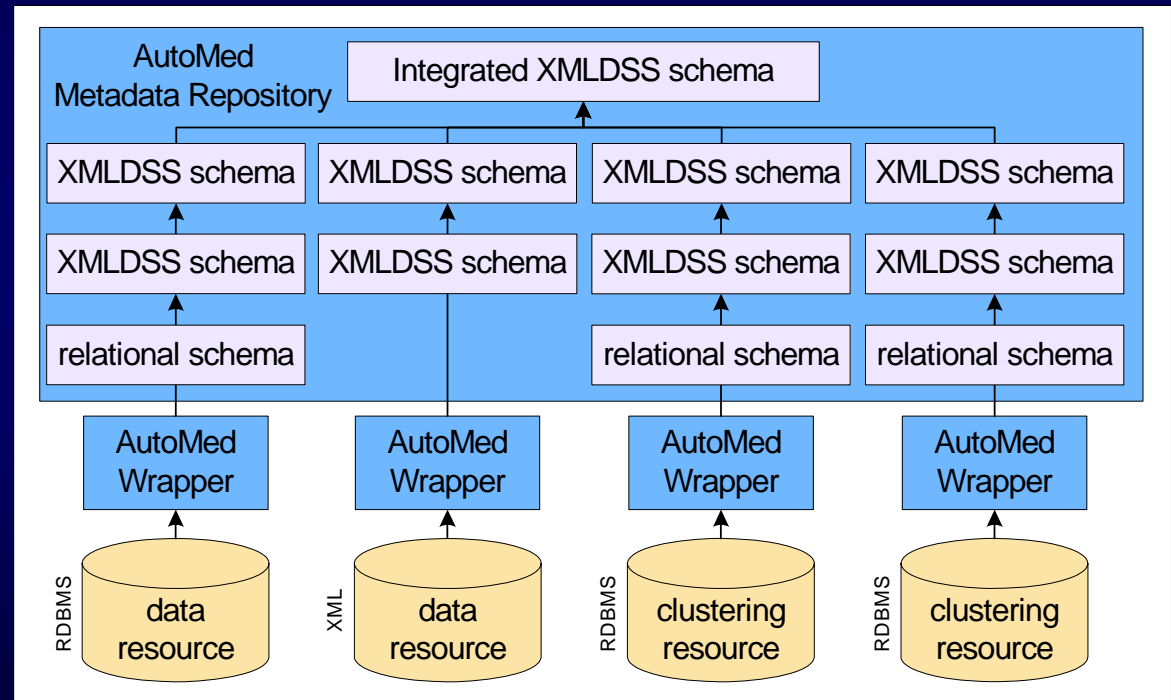
ISPIDER Objectives



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

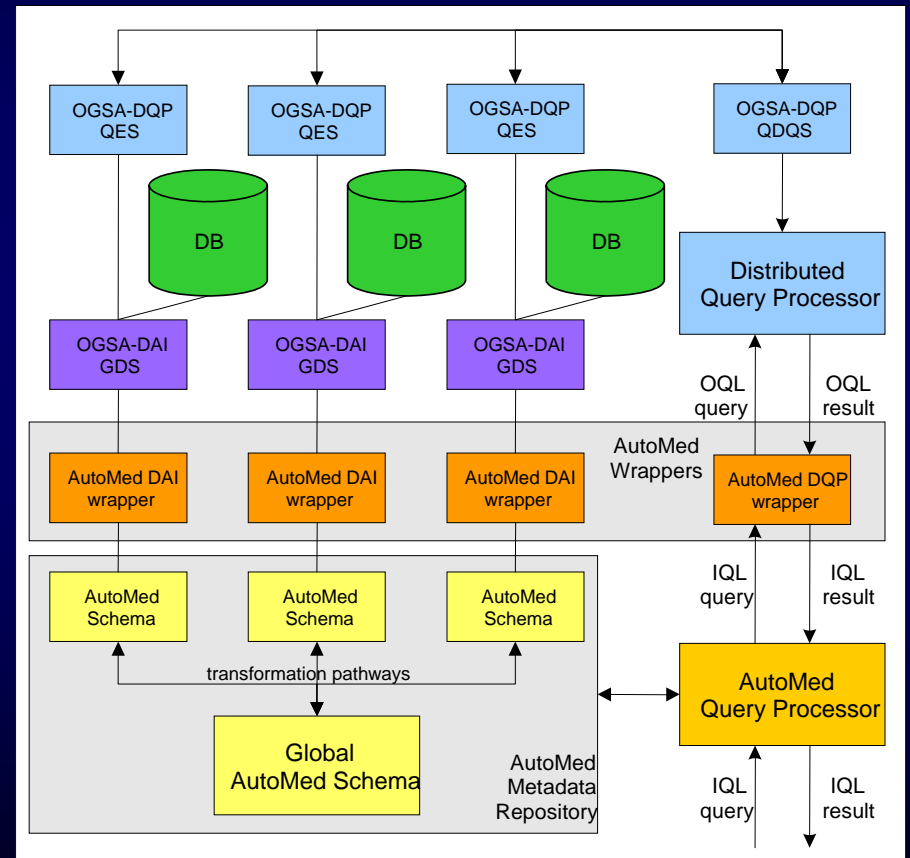
BioMap: Materialised Integration

- Relational → XML (automatic)
- Schema conformance & data cleansing (manual)
- XML integration (automatic)
- XML or SQL materialisation



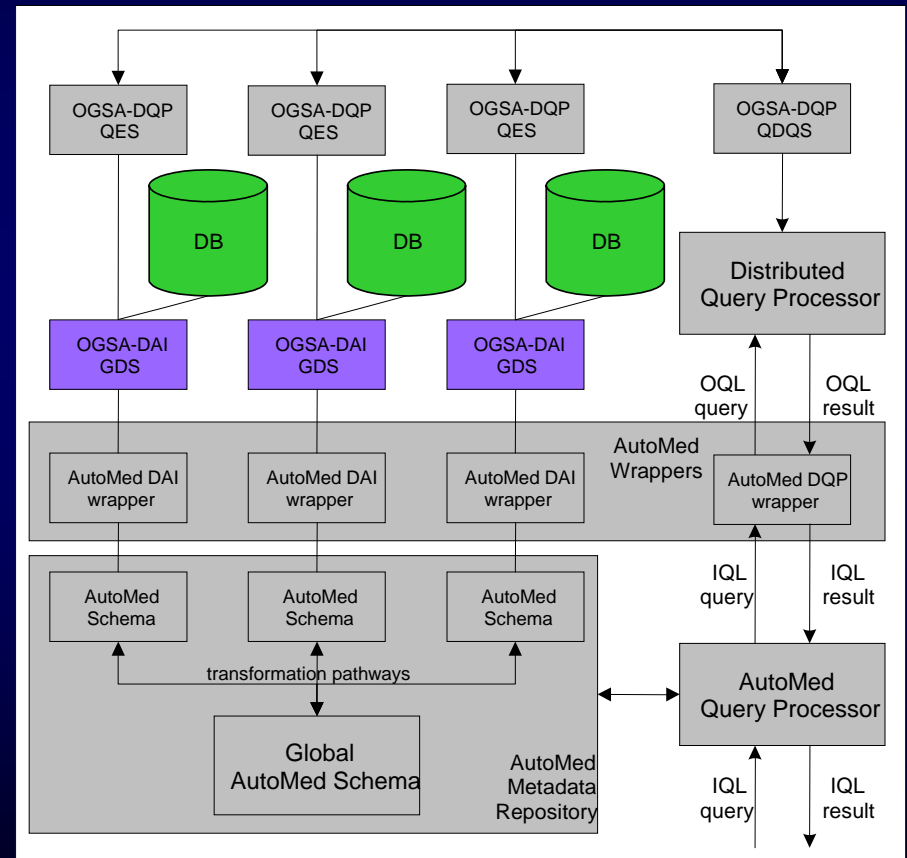
ISPIDER: Virtual Integration

- Sources wrapped with OGSA-DAI
- AutoMed toolkit wraps OGSA-DAI resources
- Integration of OGSA-DAI resources
- Queries submitted to AutoMed QP are evaluated with the help of OGSA-DQP



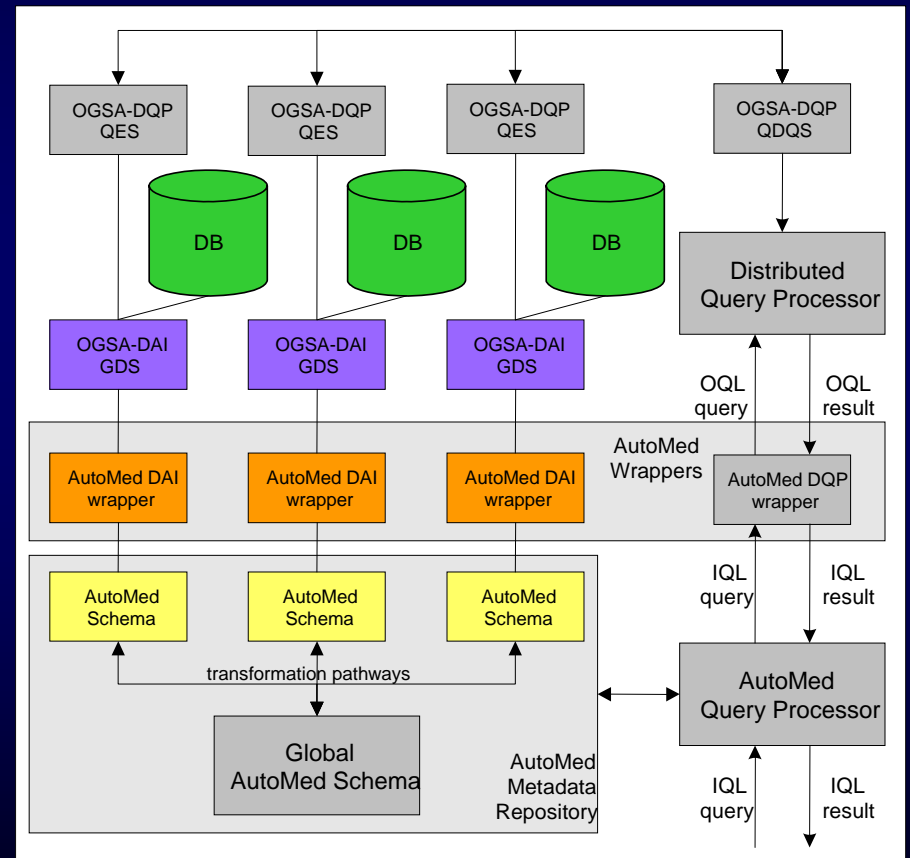
ISPIDER: Virtual Integration

- Sources wrapped with OGSA-DAI
- AutoMed toolkit wraps OGSA-DAI resources
- Integration of OGSA-DAI resources
- Queries submitted to AutoMed QP are evaluated with the help of OGSA-DQP



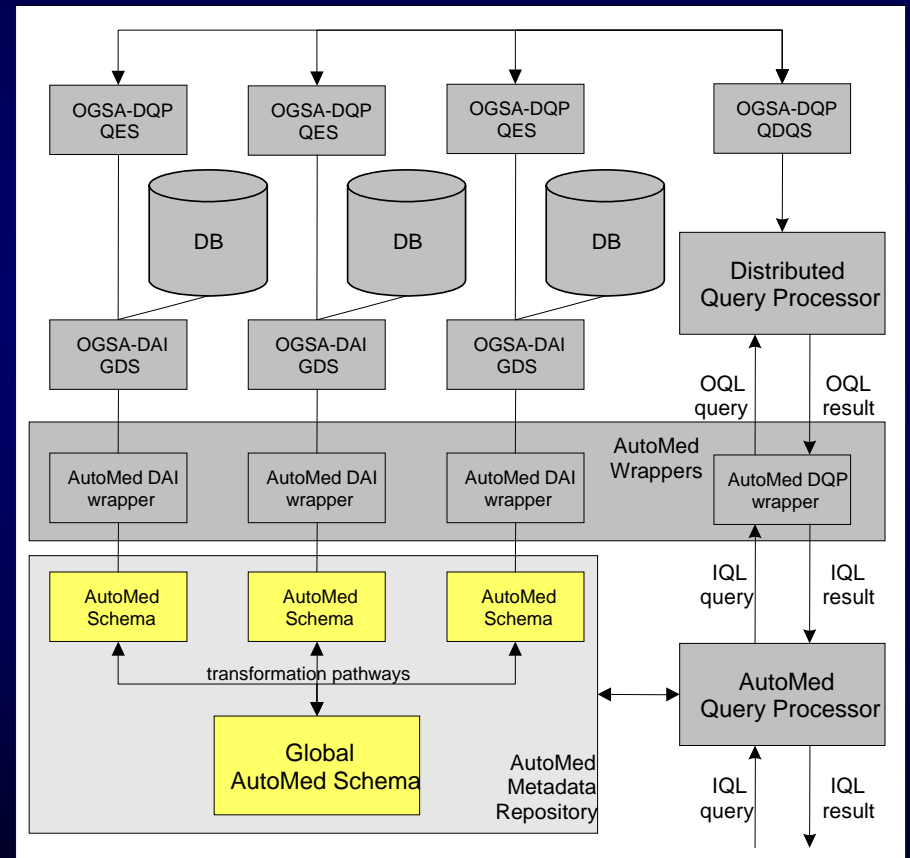
ISPIDER: Virtual Integration

- Sources wrapped with OGSA-DAI
- AutoMed toolkit wraps OGSA-DAI resources
- Integration of OGSA-DAI resources
- Queries submitted to AutoMed QP are evaluated with the help of OGSA-DQP



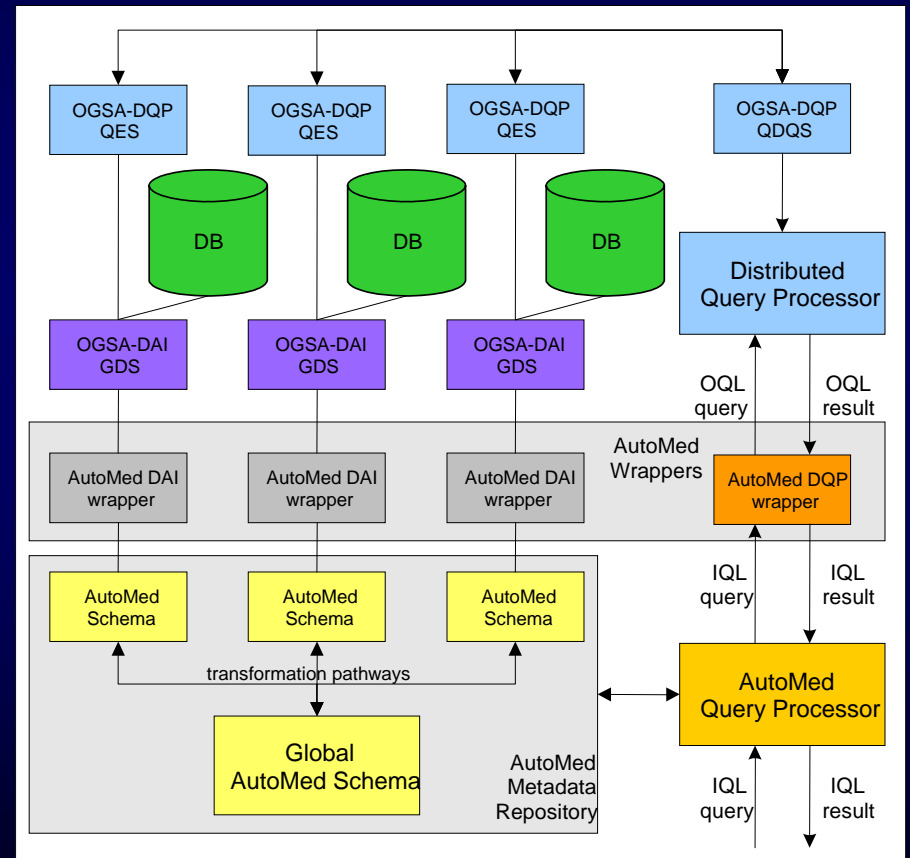
ISPIDER: Virtual Integration

- Sources wrapped with OGSA-DAI
- AutoMed toolkit wraps OGSA-DAI resources
- Integration of OGSA-DAI resources
- Queries submitted to AutoMed QP are evaluated with the help of OGSA-DQP



ISPIDER: Virtual Integration

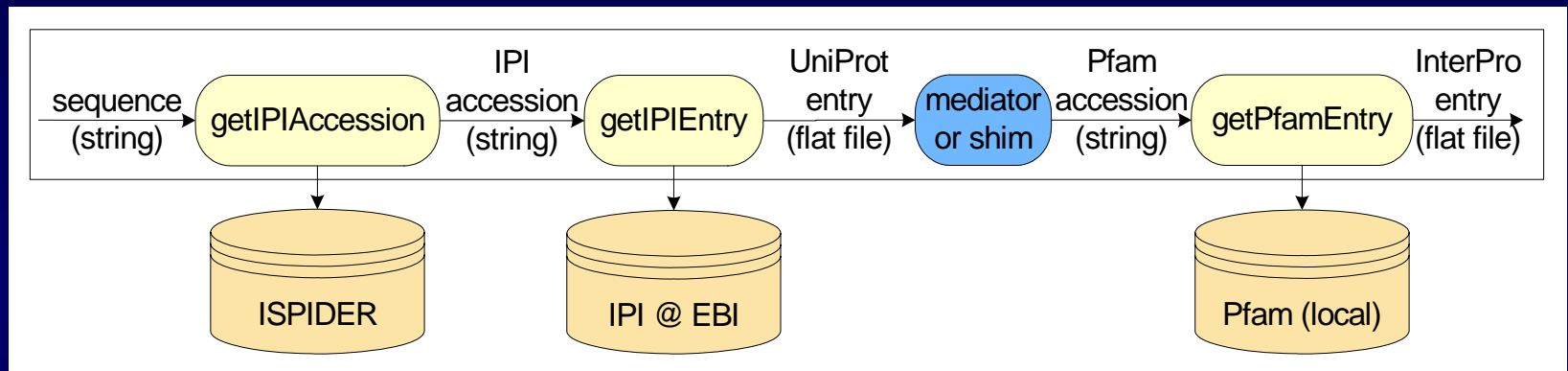
- Sources wrapped with OGSA-DAI
- AutoMed toolkit wraps OGSA-DAI resources
- Integration of OGSA-DAI resources
- Queries submitted to AutoMed QP are evaluated with the help of OGSA-DQP



ISPIDER: Scientific Workflows

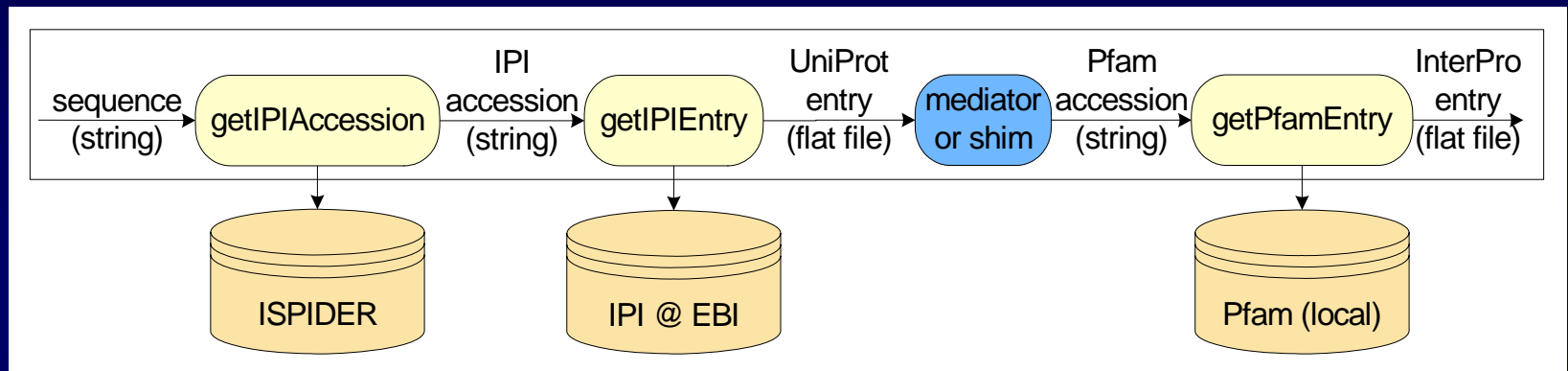
- Plethora of bioinformatics services impedes service composition
 - service discovery used to reduce the search space
- Semantically compatible services may not be able to interoperate:
 - service technologies
 - differences in data model, data structure, data types
 - need for **service reconciliation**.
- Reconciliation problem amplified due to:
 - simple strings used rather than complex types
 - service providers disinclined to supply annotations

ISPIDER: Scientific Workflows



- Current approach: a *shim* for each pair of services (shim: service that reconciles two other services)
- ...but ^{my}Grid project currently has more than 3000 services → does not scale!

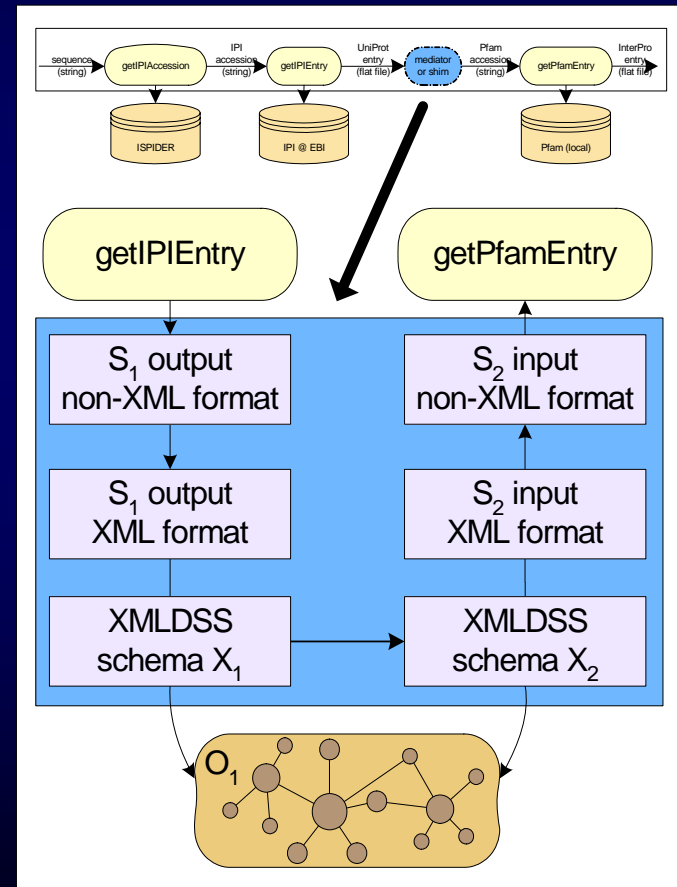
ISPIDER: Scientific Workflows



- Proposed approach: service reconciliation via mediation using the AutoMed system
- Requirements:
 - Wide coverage of interoperability issues
 - Scalability of approach, promote reusability
 - Static/dynamic mediation

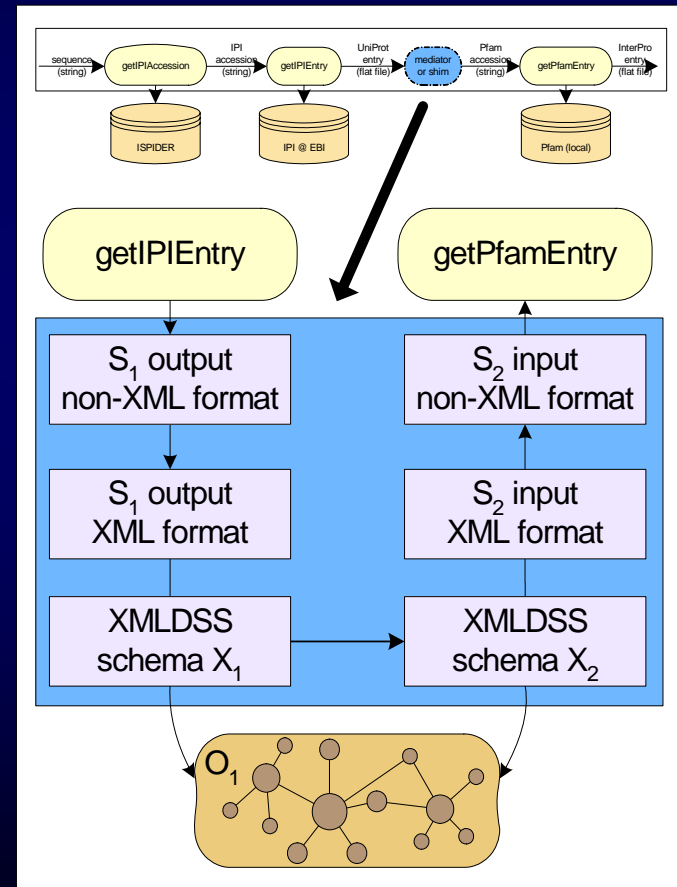
ISPIDER: Scientific Workflows

- Providers/users of services provide correspondences from service input/output to an ontology
- If service I/O not in XML → convert to XML
- Use correspondences to automatically reconcile services



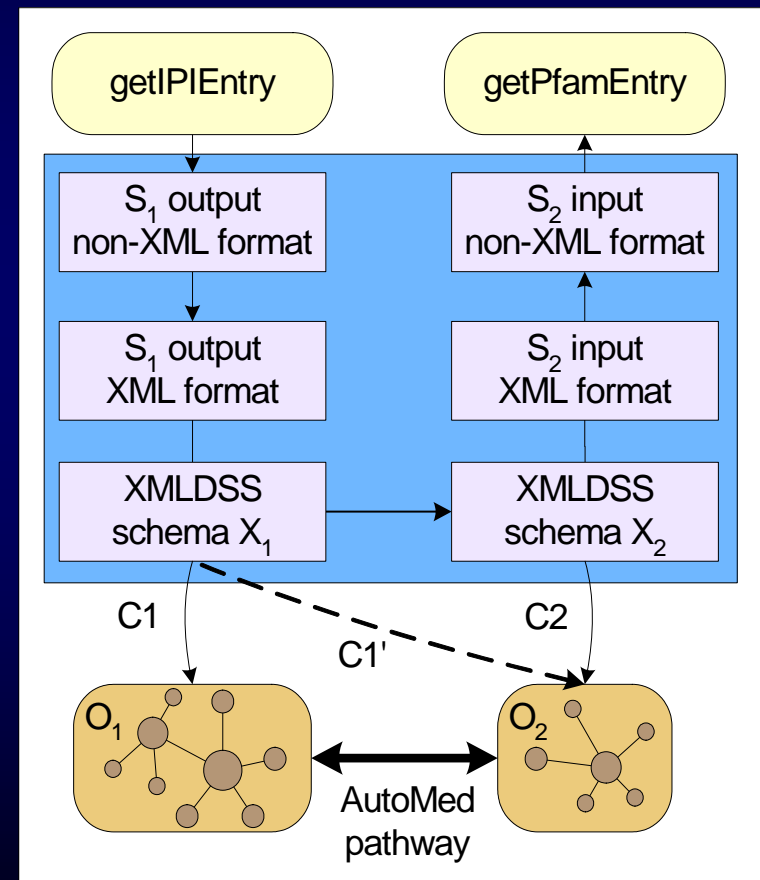
ISPIDER: Scientific Workflows

- Handles all heterogeneity types
 - data model
 - semantic
 - schematic
 - primitive data type/scaling
- Promotes reusability of annotations
- Assumption: workflow tool addresses service technology reconciliation (e.g. Taverna uses the Freefluo component)



ISPIDER: Scientific Workflows

- In a setting where:
 - X1 corresponds to O1 using C1
 - X2 corresponds to O2 using C2
 - there is a (direct or indirect) AutoMed pathway $O1 \leftrightarrow O2$
- Automatically produce new set of correspondences C1' for X1 and O2 (using query reformulation)
- Setting is now identical to single ontology setting.
- Proviso: C1' syntax must conform to our correspondences language.



References

- AutoMed – <http://www.doc.ic.ac.uk/automed>

P.J. McBrien and A. Poulovassilis, **Data Integration by Bi-Directional Schema Transformation Rules**, Proc. of International Conference on Data Engineering (ICDE), 2003

- ISPIDER – <http://www.ispider.manchester.ac.uk>

M. Maibaum, L. Zamboulis, G. Rimon, N. Martin, A. Poulovassilis, **Cluster based Integration of Heterogeneous Biological Databases using the AutoMed toolkit**, Proc. Data Integration in the Life Sciences (DILS), 2005

L. Zamboulis, H. Fan, K. Belhajjame, J. Siepen, A. Jones, N. Martin, A. Poulovassilis, S. Hubbard, S. M. Embury, N. W. Paton, **Data Access and Integration in the ISPIDER Proteomics Grid**, Proc. Data Integration in the Life Sciences (DILS), 2006

L. Zamboulis, N. Martin, A. Poulovassilis, **Bioinformatics Service Reconciliation By Heterogeneous Schema Transformation**, Proc. Data Integration in the Life Sciences (DILS), 2007