

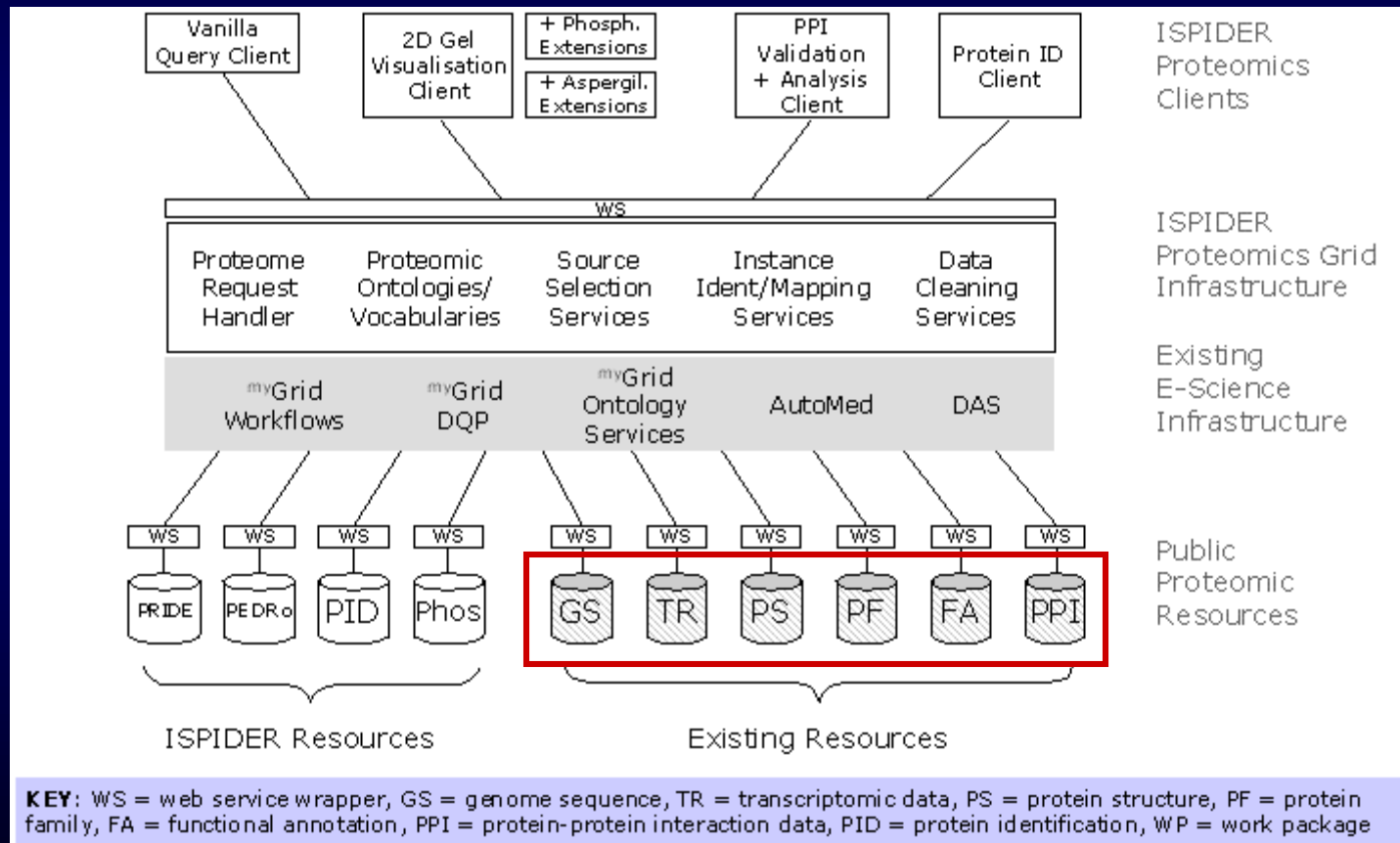


Data Access & Integration in the ISPIDER Proteomics Grid

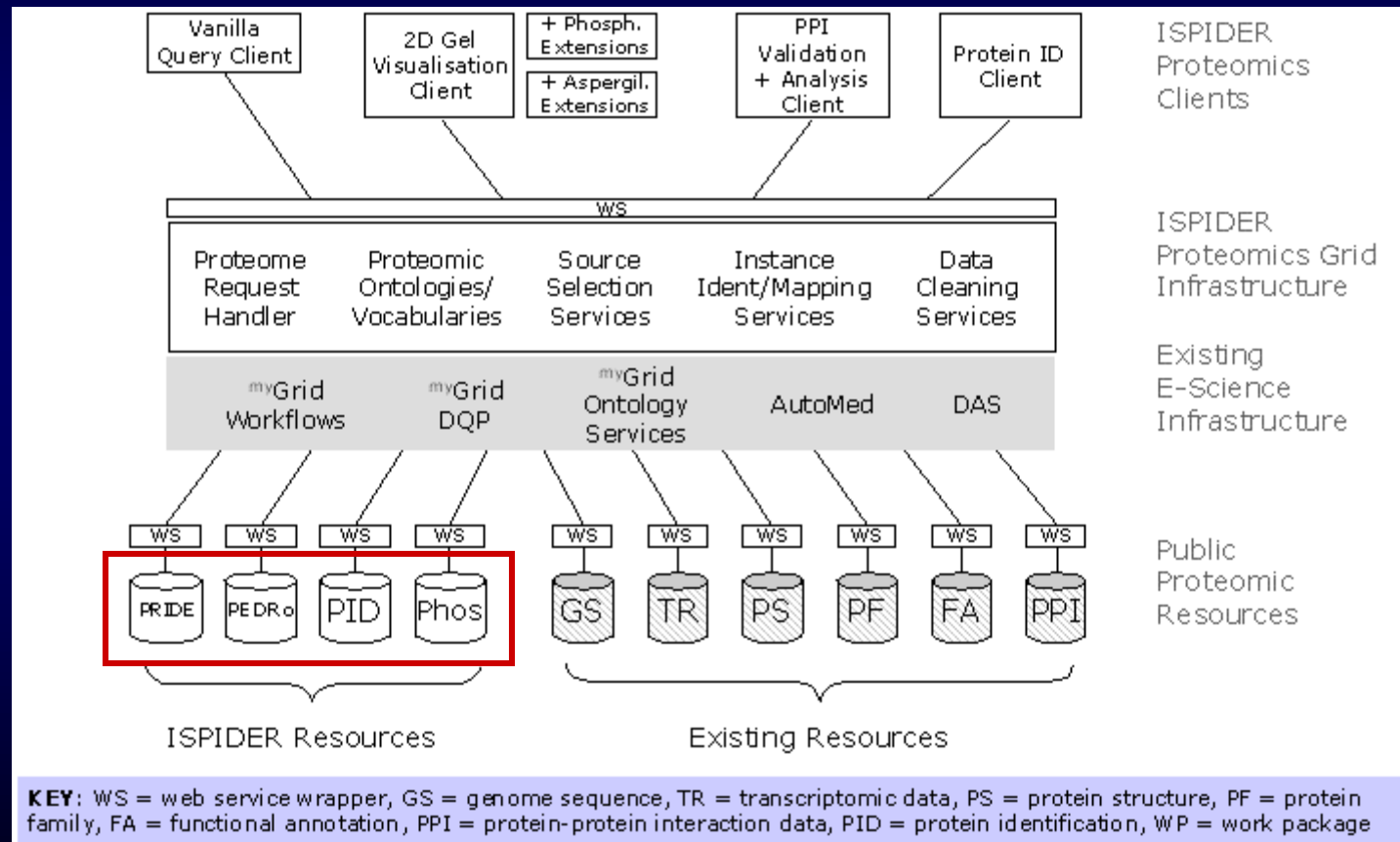
to appear in Data Integration in the Life Sciences 2006

L. Zamboulis, H. Fan, K. Belhajjame, J. Siepen, A. Jones, N.
Martin, A. Pouloussilis, S. Hubbard, S. M. Embury, N. W. Paton

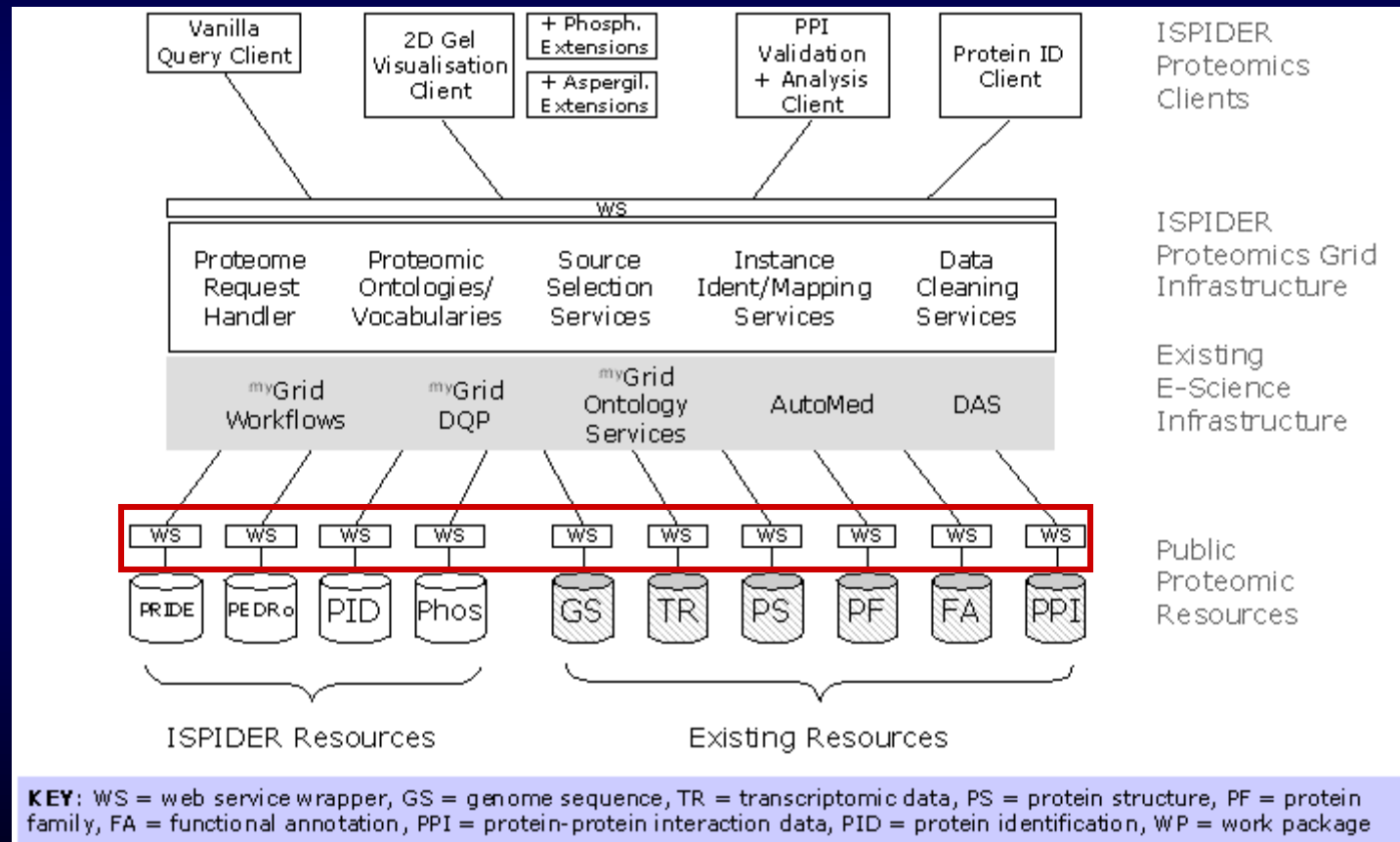
ISPIDER Architecture



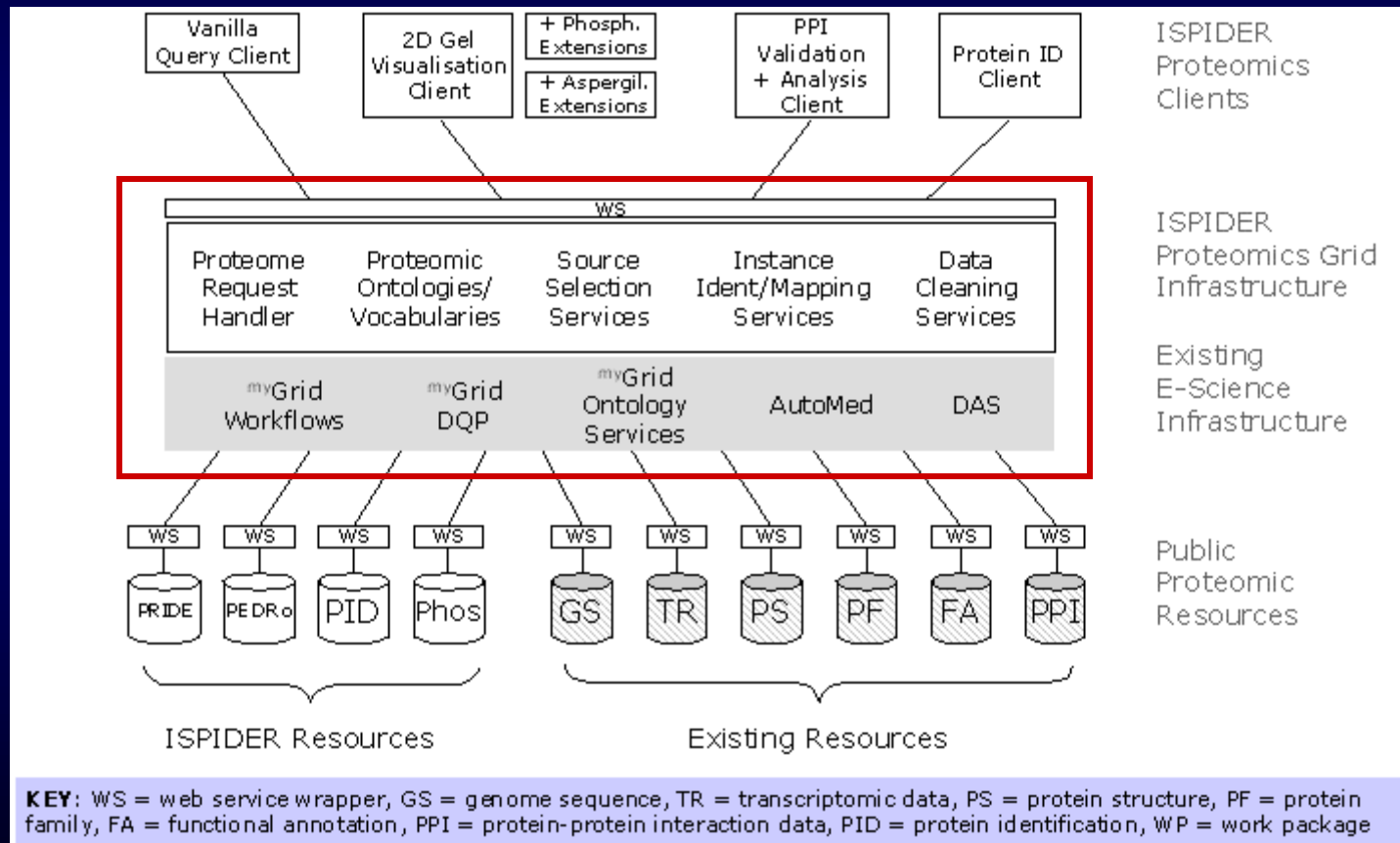
ISPIDER Architecture



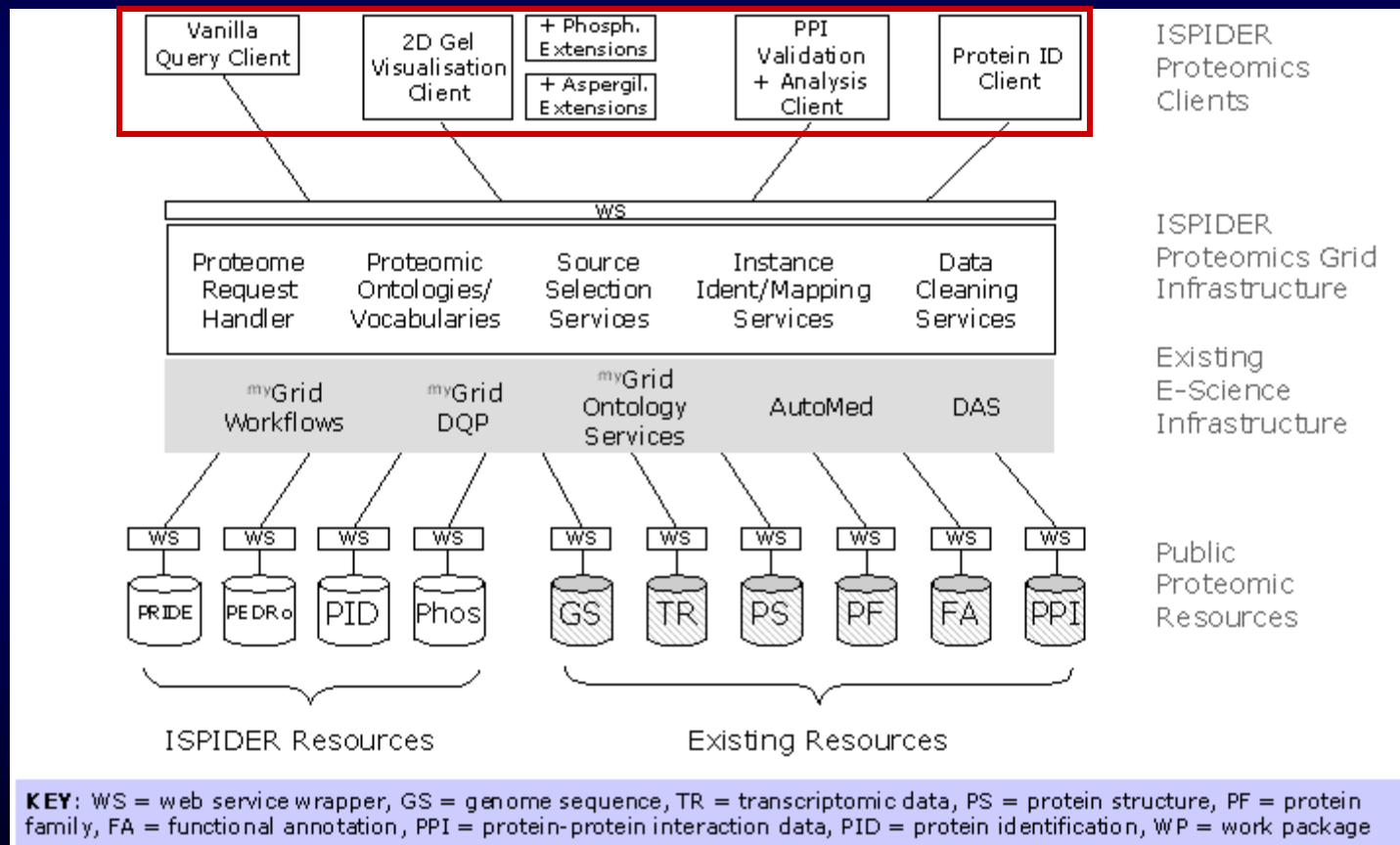
ISPIDER Architecture



ISPIDER Architecture



ISPIDER Architecture



Resource Integration

- Objective: develop an integrated platform of proteome-related resources, using existing standards
- Benefits:
 - Access to increased breadth of information
 - More reliable analyses
 - Integration, not union → added value

Resource Integration

- Multiple large proteomics repositories in disparate locations → *need for a distributed solution*
 - Common access to all resources
 - Efficient query processing
- *Multiple large* proteomics repositories in disparate locations → *need for integration*
 - Overlapping data
 - Same information with different representations

Middleware (1/2)

- OGSA-DAI: middleware product exposing data resources on Grids via web services
 - Open-source and extendable
 - Supports variety of data sources
 - Uniform access to data sources
 - Supports a variety of operations, such as querying/ updating, data transformation and data delivery
- OGSA-DQP: service-based distributed query processor
 - Efficient querying of OGSA-DAI resources through parallelism

Middleware (2/2)

- AutoMed: heterogeneous data transformation and integration system
 - Subsumes traditional data integration approaches
 - Handles various data models – easily extensible
 - Virtual/materialised/ hybrid integration
 - Data warehousing tools
 - Schema evolution

Proteomics Resources

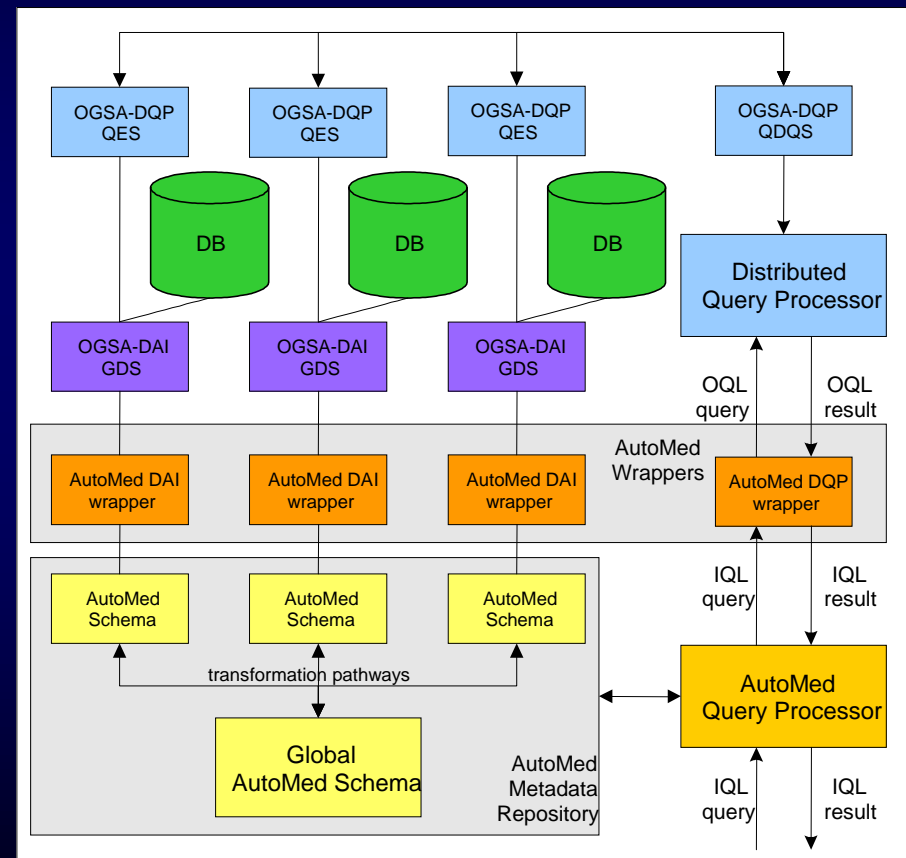
- PEDRo
 - collection of descriptions of experimental data sets in proteomics
 - has been used as a format for exchanging proteomics data
- gpmDB
 - contains a large number of proteins and peptide identifications
 - initially designed to assist in the validation of peptide MS/MS spectra and protein coverage patterns
- PepSeeker
 - developed as part of the ISPIDER project
 - comprehensive resource of peptide/protein identifications
- PRIDE
 - centralised, standards compliant, public proteomics repository
 - contains protein/peptide identifications + evidence supporting them

Global Schema

- Incremental build
 - based on PEDRo's peptide/protein identification section
 - expanded with information unique in other resources
- Scope trade-off
 - able to answer very specific user queries
 - a full integration
- Entities identified by LSIDs
 - *URN:LSID:ispider.man.ac.uk:pedro.protein:99*

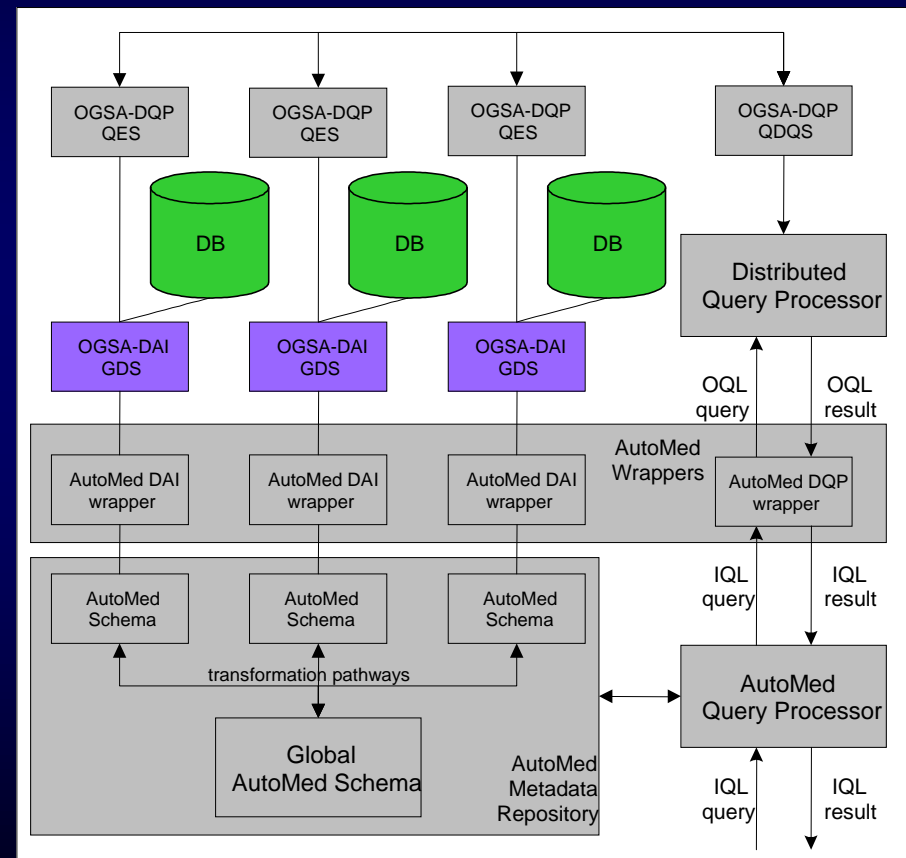
System Architecture

- 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



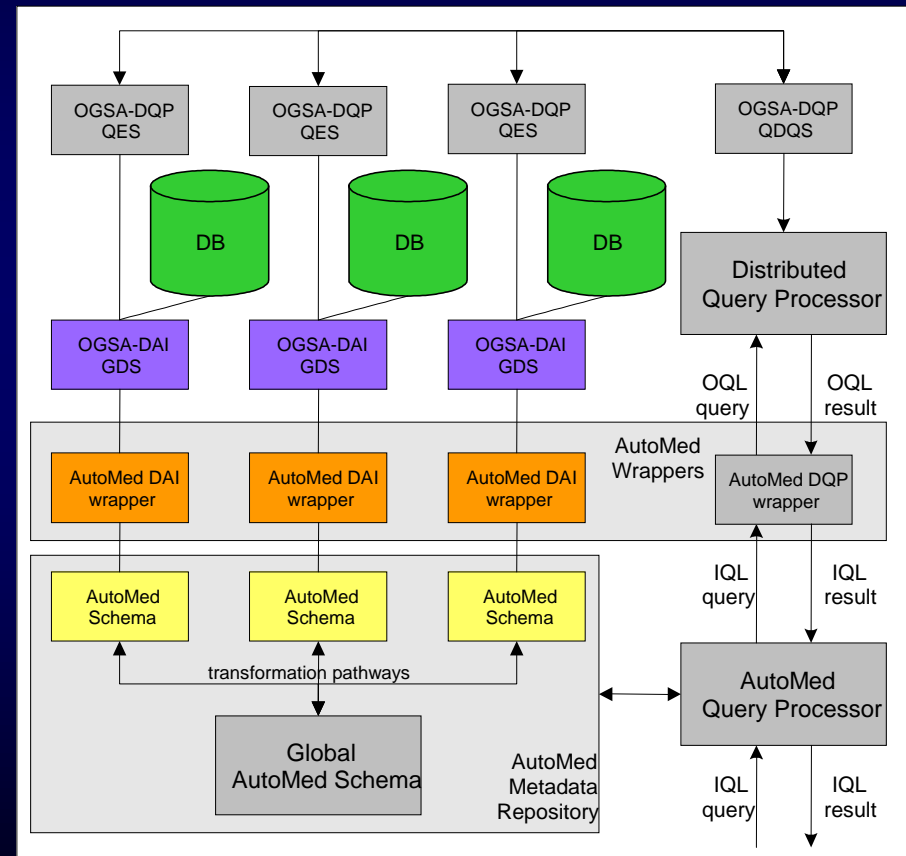
System Architecture

- 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



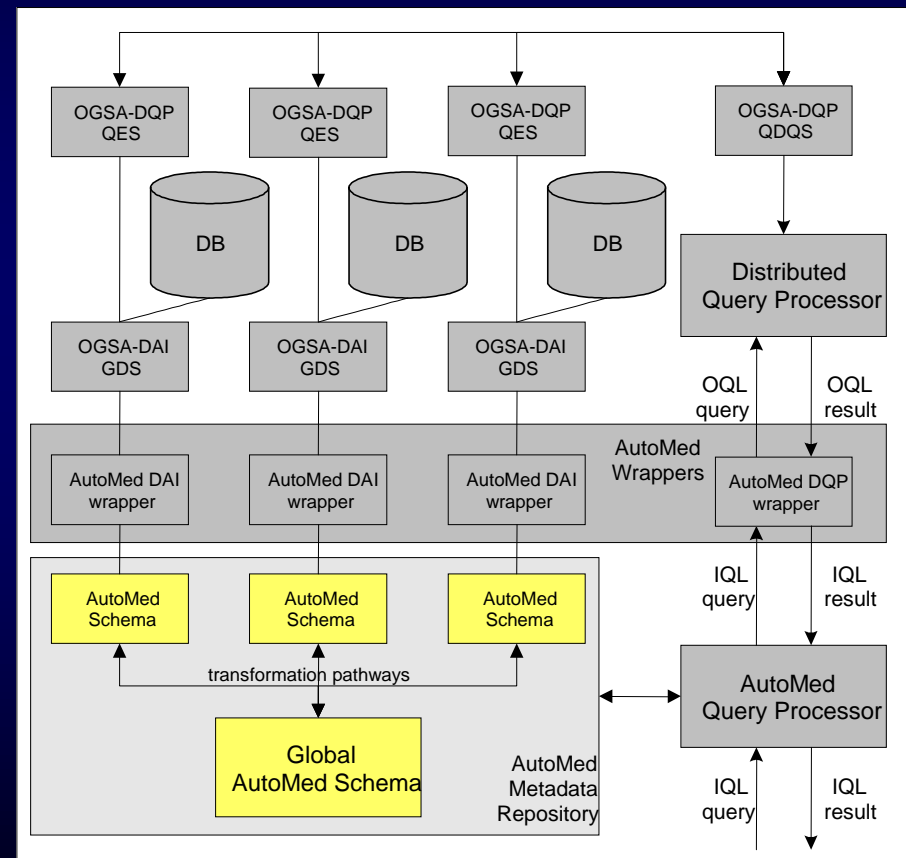
System Architecture

- 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



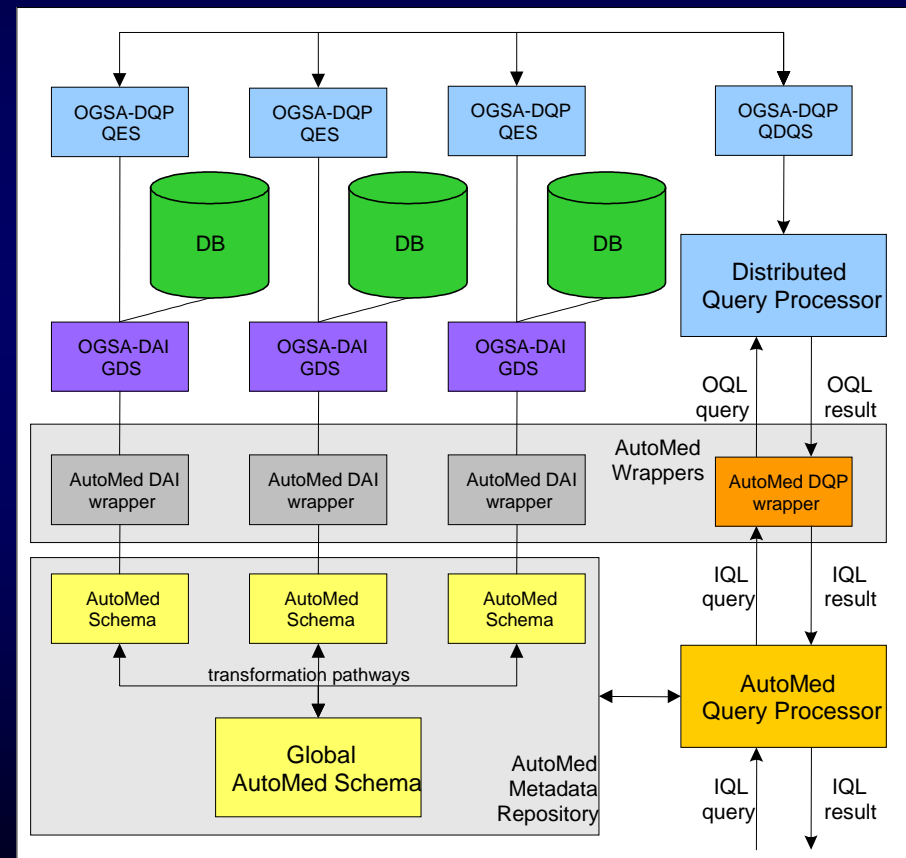
System Architecture

- 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



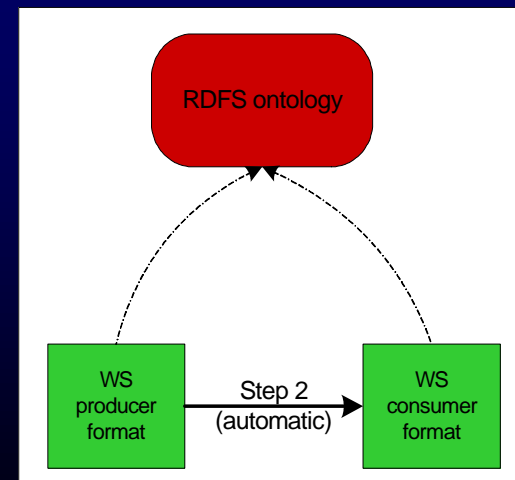
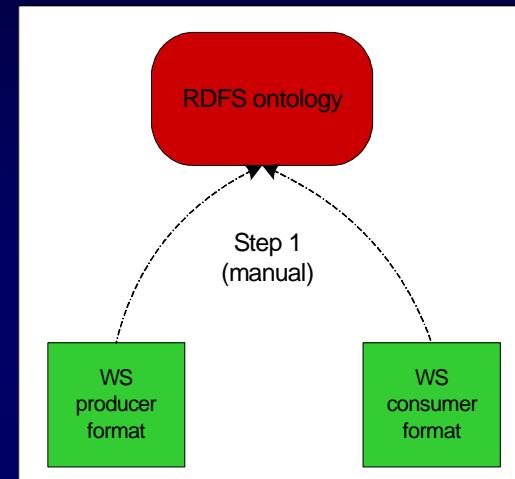
System Architecture

- 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



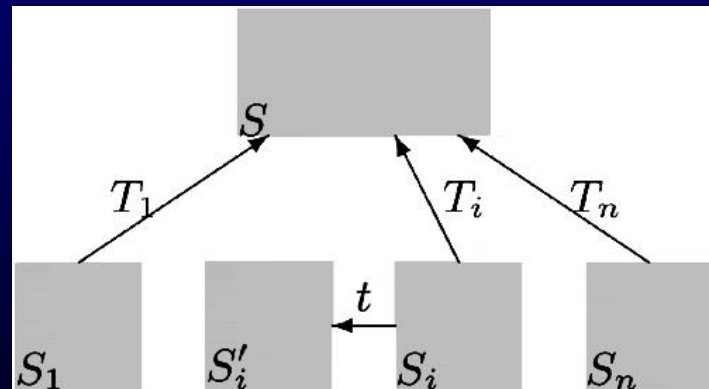
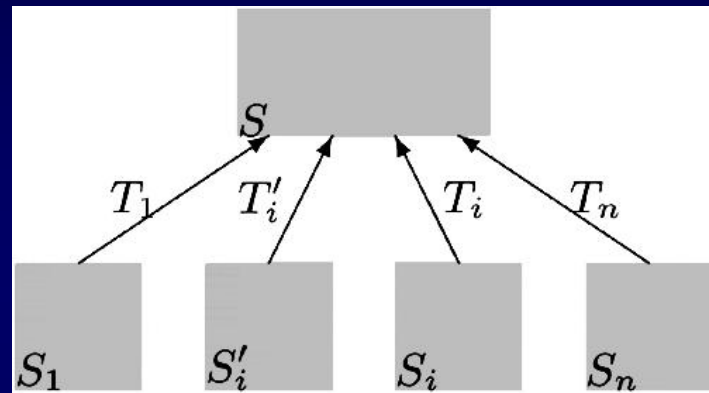
Future Work

- AutoMed and Taverna for WS integration
- Schema evolution



Future Work

- AutoMed and Taverna for WS integration
- 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.)