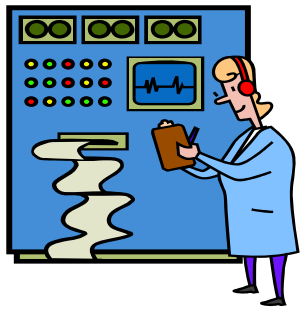
Grid Data Integration based on Schema-mapping

Carmela Comito

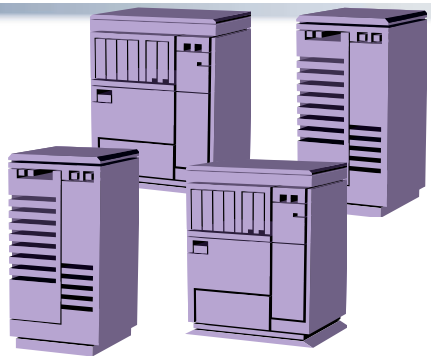
CoreGRID, University of Calabria – Italy

ccomito@deis.unical.it

Motivation(1)



Single Site
Clusters



Distributed
10's – 100's

Internet Scale
1000's – Millions

Database Community — Network Community

- The way data management is viewed is massively changed
  - Data is being increasingly processed within networks
- Challenges
  - How to run database style queries at Internet scale?
  - Can DB concepts influence the next Internet architecture?

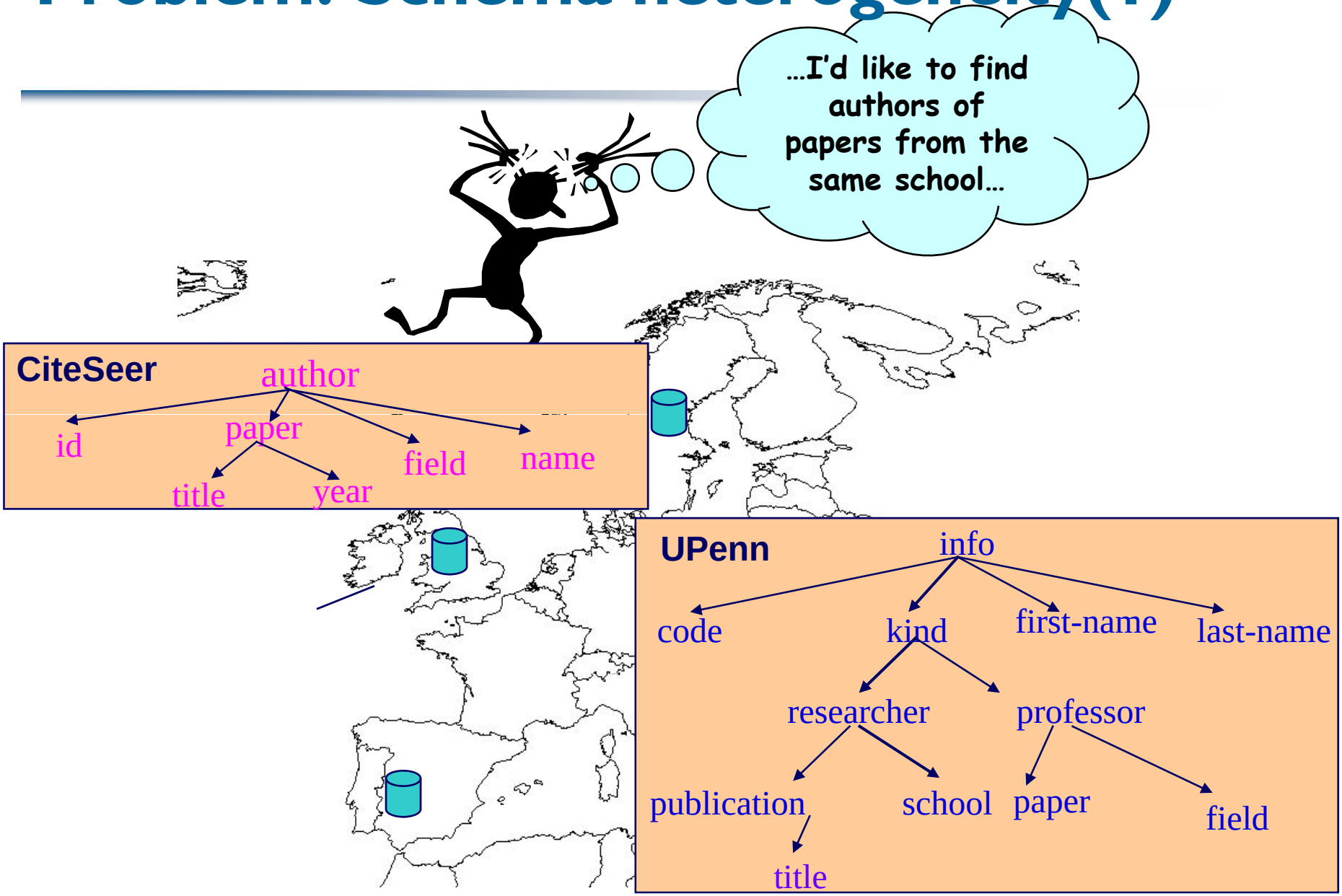
Motivation(2)

- Challenges
  - Data is naturally distributed (Distribution)
  - Centralized collection undesirable (Decentralization)
  - Heterogeneous schemas (Schema Heterogeneity)
  - Large-scale data sources (Scalability)
  - Dynamic environment (Dynamism)
  - Queries are posed using a node's schema. Answers come from anywhere in the system (Sharing and Cooperation

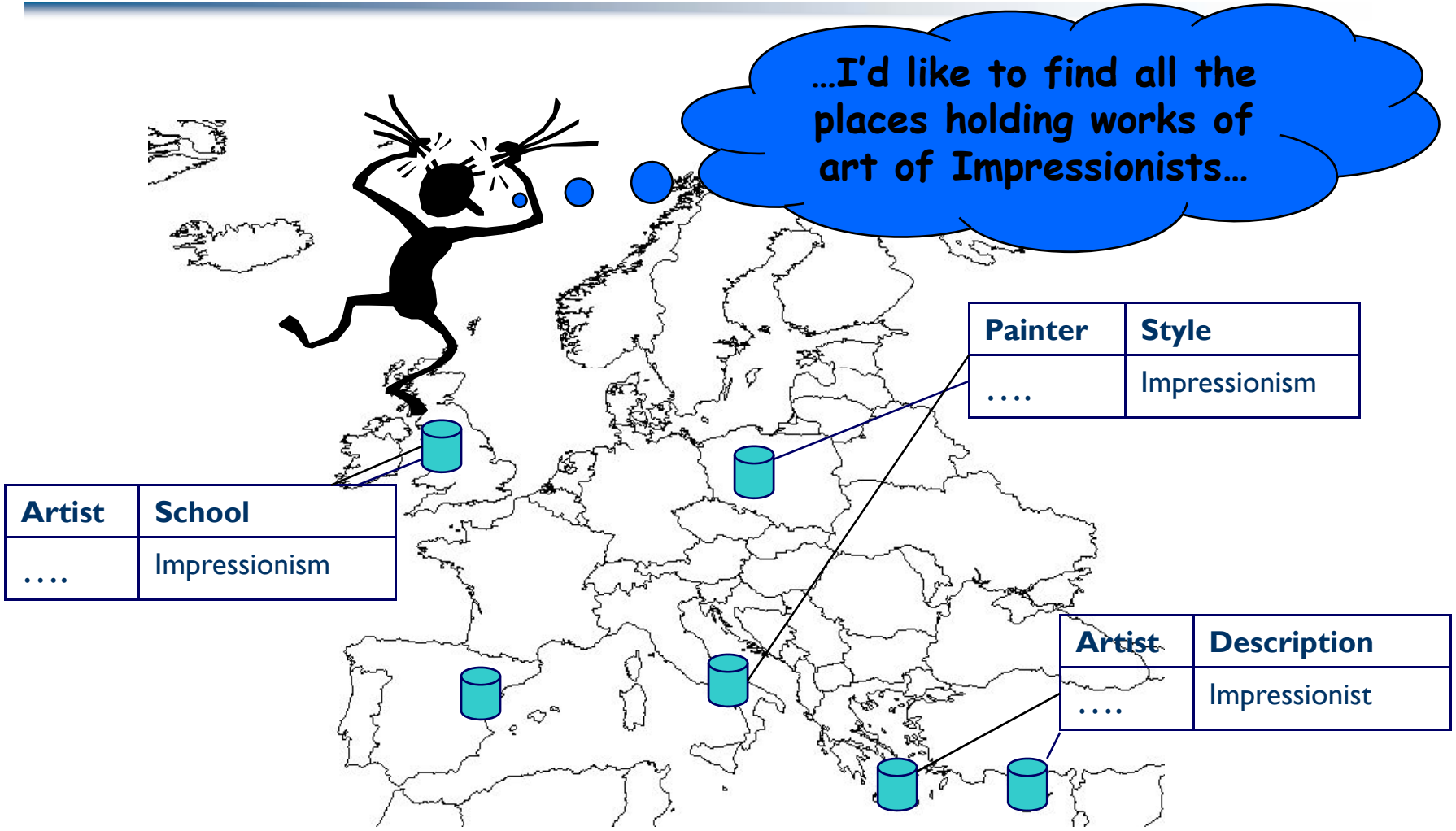
The focus on Semantics

- *How do we coordinate data sources that have different data model and/or source languages, and are on different machines?*
  - Challenges
    - Everything has a different semantics
    - Heterogeneity, distribution, large-scale, decentralization
  - Parts of the solution
    - Standard data formats: XML, XMLSchemas, Ontologies
  - Schema heterogeneity
    - "Schema mediation" and data translation
    - Ontologies

Problem: Schema heterogeneity(1)



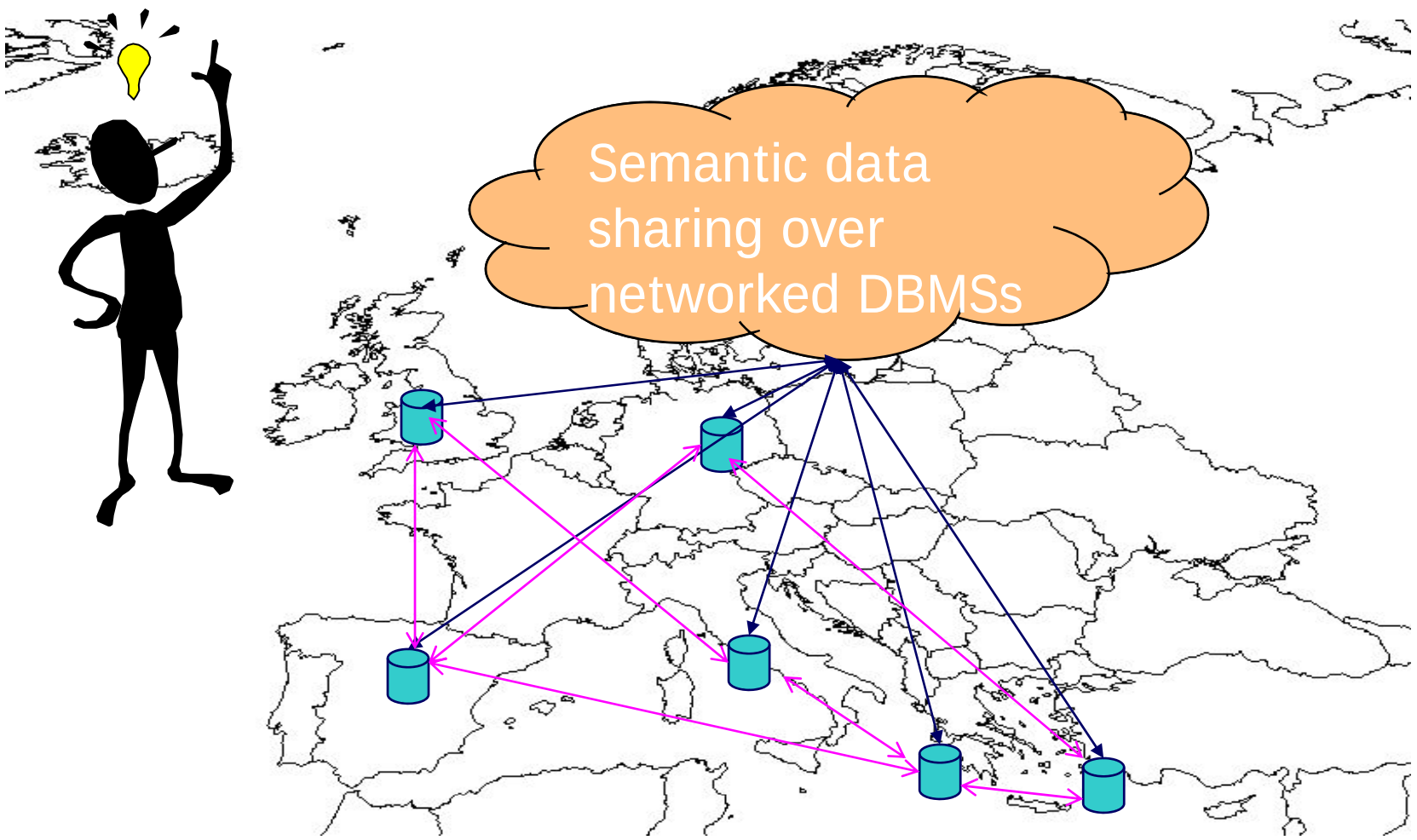
Problem: Schema heterogeneity(2)



What is the current practice?

- Simple… (and primitive… ☺)
- Just get access to each database and search, separately, using their own rules for querying!
- But:
  - Requires access rights everywhere (as well as familiarity with different schemas)
  - It is time-consuming
  - The output is not really data integration

Our Solution



Contribution (1)

- Peer-to-Peer and Grids
  - Commom goal: enabling resource sharing at large
  - Integrating computer into networks in which each can consume and offer services
  - Lack semantics
- ➢ *Filling this gap with an integrated approach*
- Large-scale data integration
  - Query reformulation based on decentralized P2P schema mappings
  - Robust and scalable peer-to-peer networks
  - Grid infrastructure
  - Comprehensive assembly of techniques from both domains

Contribution (2)

- **XMAP**: decentralized, scalable network of semantically-related XML data sources distributed over autonomous, dynamic nodes
- **XMAP-DQP** : service-based architecture for data integration enabled distributed query processing in Grids
- **PARIS**: P2P architecture where clusters of peers interoperate through XMAP

Outline

- Data Integration over Networked System
- The XMAP framework
- The XMAP-DQP system
- Conclusions

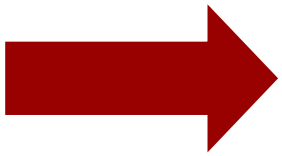
XMAP: XML Data Integration based on Schema Mapping

Data integration in Networked Systems

A data integration system provides a uniform query interface across autonomous, heterogeneous networked or local data sources

In the considered scenario we deal with:

- Multiple, autonomous, unpredictable sites
- Huge, highly dynamic, data volumes
- Heterogeneity and Distribution of data resources
- Sites both clients and servers



Traditional approaches to data integration are not suitable over networked data

Challenges of Networked Data Integration Architectures

- Networked data raises new challenges in data integration systems:
  - No need for a central mediated schema (*Decentralization*)
  - Ability to map data as is most convenient (*Flexibility, Dynamism*)
  - Wide-scale, ad-hoc nature *(Scalability*)
  - Queries are posed using the node's schema. Answers come from anywhere in the system (*Sharing and Cooperation*)

P2P data integration

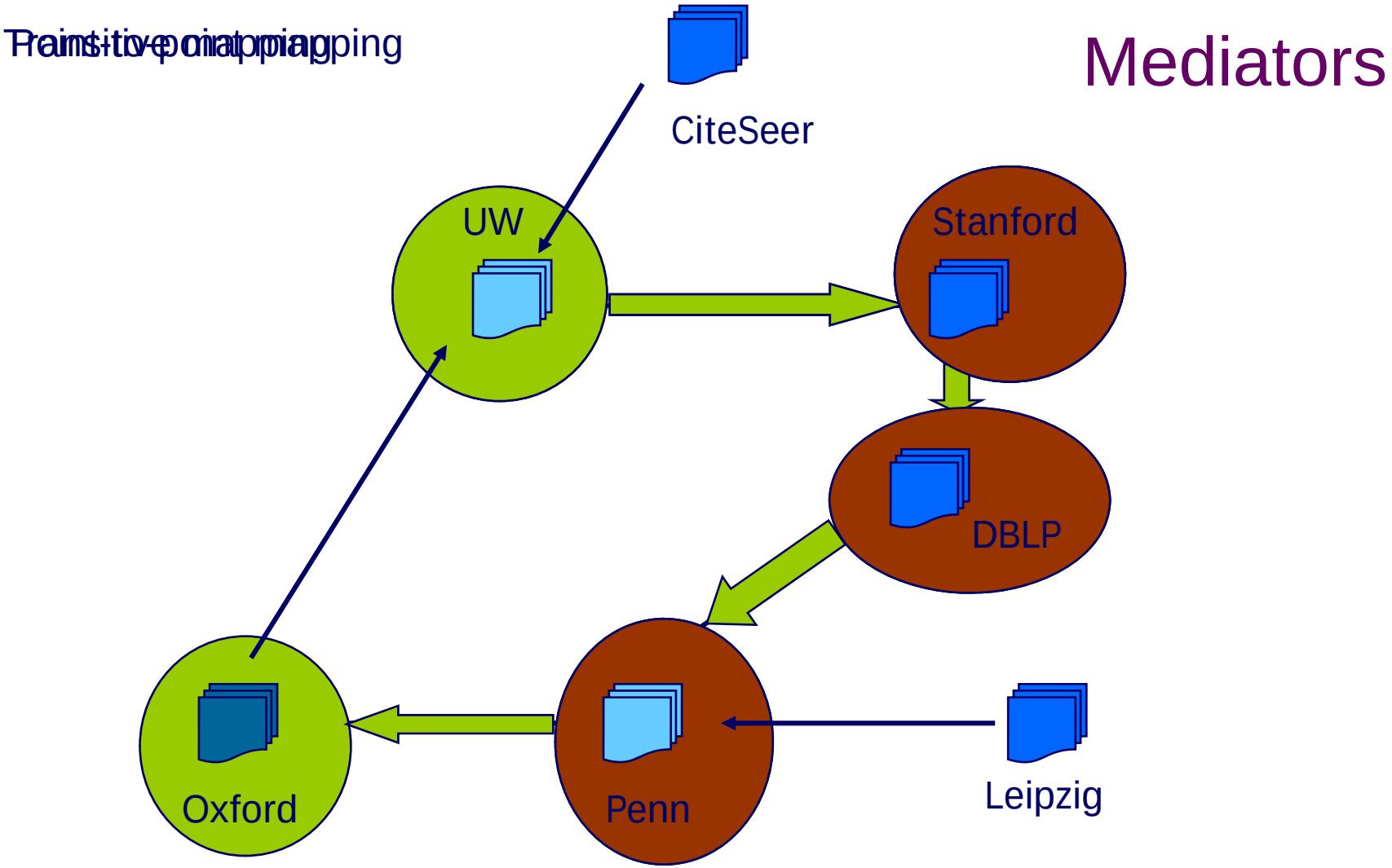
- Goal:
  - Large-scale autonomous semantic sharing of structured data
- A "logical" peer-to-peer model
  - Every autonomous peer can contribute:
    - Extensional data in its own schema
    - Mappings between schemas
    - Computation (query answering) and caching
  - Lacks of a global schema
  - NOT a P2P file sharing system

The XMAP framework

A decentralized network of semantically-related XML data sources

- A set of distributed, heterogeneous, autonomous XML data sources
  - Mapping is a key issue to any data sharing architecture
- XMAP integration model is based on schema mappings
- Mapping specification is flexible and scalable, not resorting to any hierarchical structure
  - Each source schema is directly connected to only a small number of other schemas (*point-to-point mapping*)
  - Each source schema is reachable from all other schemas belonging to its transitive closure (*transitive mapping*)

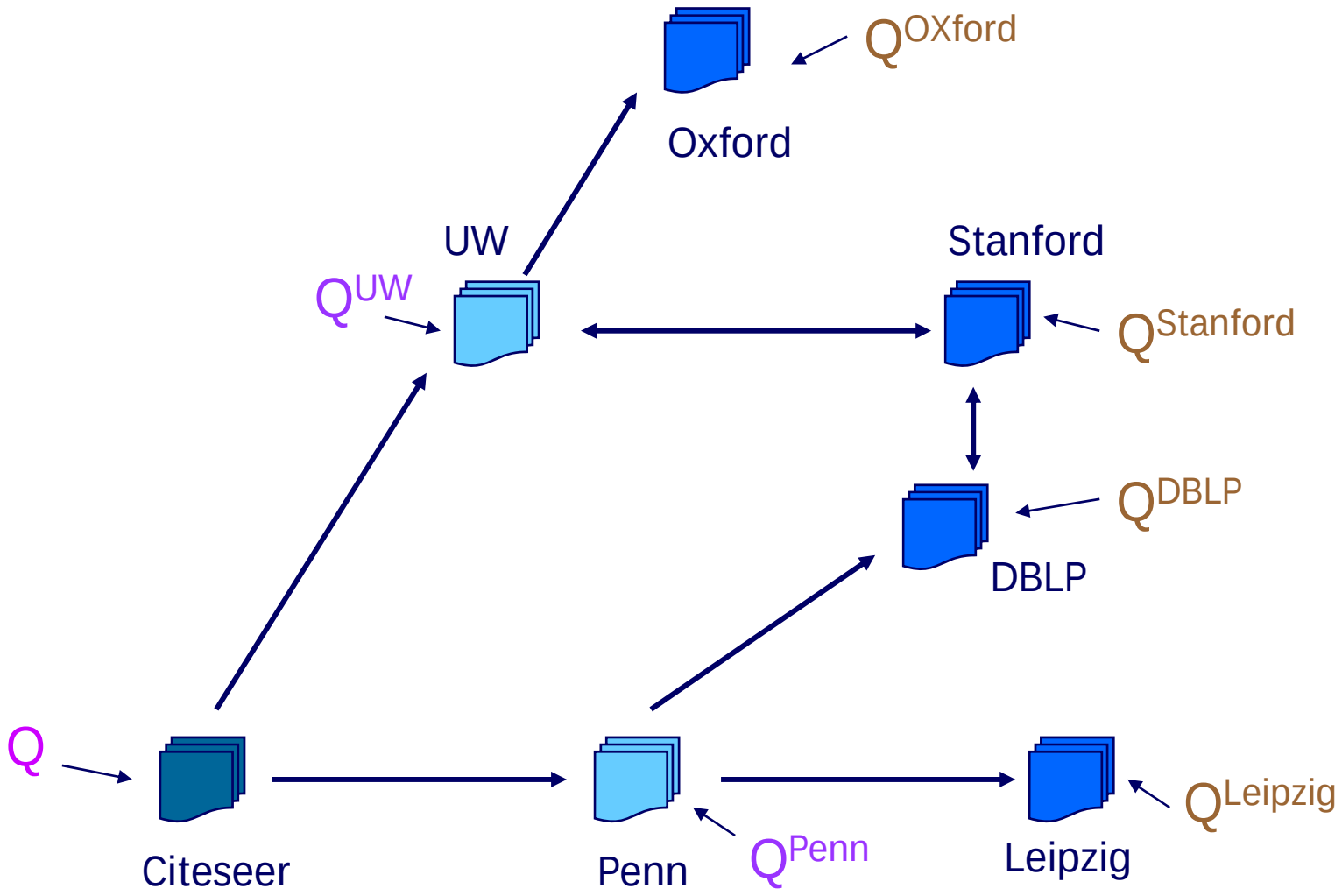
The XMAP Integration Approach



XMAP reformulation algorithm

- The XMAP reformulation algorithm reformulates an XPath query Q over all the schemas related to Q
- Query is answered by chaining of mapped sources using the mapping rules defined in XMAP documents
  - Direct reformulations of Q by using the mapping of its schema
  - Transitive reformulations are obtained by recursively invoking the algorithm over each reformulated query

XMAP reformulation algorithm - Example



XMAP-ODS: a Service-based Architecture for Data Integration in Grids

Done in cooperation with the Computer Science department of University of Manchester (UK)

Databases and the Grid

- Resource virtualization through a service-oriented architecture
  - Wrap databases as services
  - Metadata-driven access
- OGSA-DAI
  - The Grid Data Service (GDS) provides access to both relational and XML databases wrapped as Grid Services
- OGSA-DQP
  - A service-based distributed query processor exposed as an OGSA Grid Service
  - Extends OGSA-DAI with

    Grid Distributed Query Service (GDQS)

    Grid Query Evaluation Service (GQES)

Goal

- Develop a decentralized service-based architecture for data integration enabled distributed query processing
- Expose schema-mapping utilities as a Web Service allowing
  - Integrate different data models on distributed databases
  - integration features addressing autonomous, dynamic sources
- Exploit the middleware provided by XMAP, OGSA-DAI, OGSA-DQP
- Use the facilities of the OGSA to dynamically achieve resources discovery and allocation

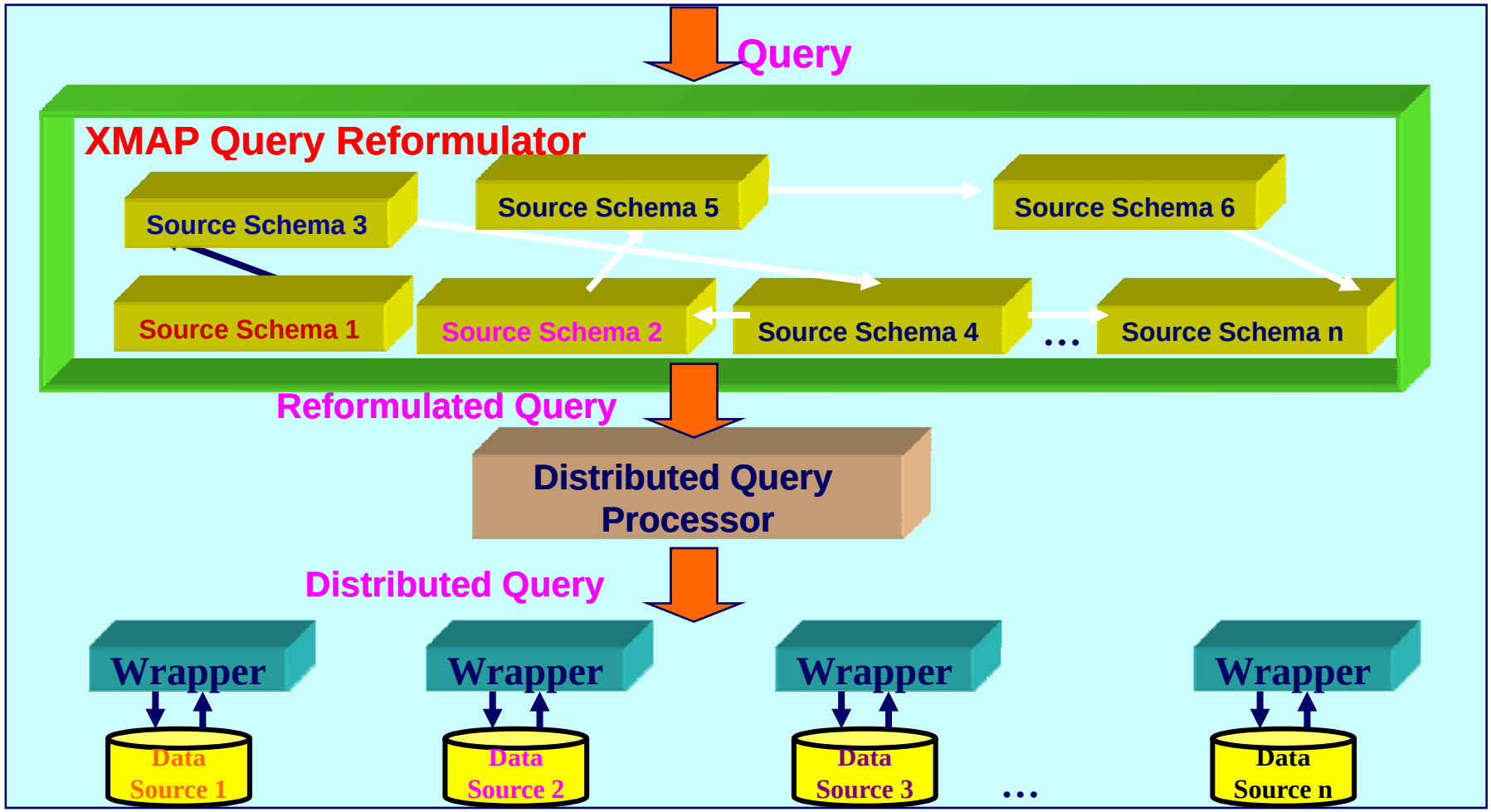
XMAP-DQP (1)

- OGSA-DQP limitation:
  - It relies on users for the semantic interpretation of the data
- XMAP:
  - Doesn't provide query processing facilities
- XMAP-DQP
  - Users need just to create a query that refers to a single database
  - The system automatically returns semantically equivalent queries that refer to all the related available databases and process them in parallel

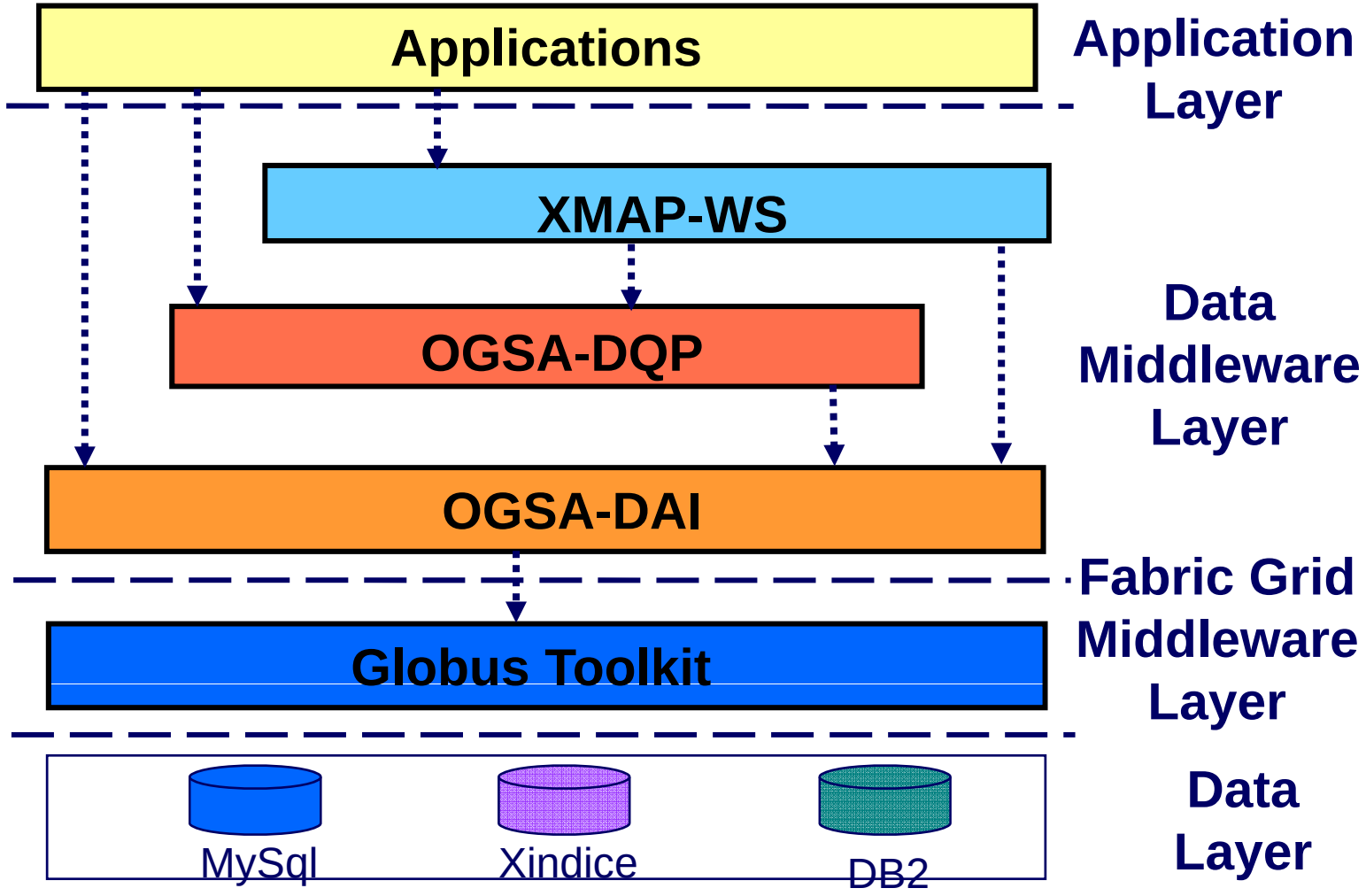
XMAP-DQP (2)

- *The XMAP-DQP is a service-based architecture integrating schema-mapping methodology with existing grid services allowing data-integration enabled distributed query processing*
  - Adopts the XMAP decentralized mediator approach to handle semantic heterogeneity over data sources
  - Adopts DQP as distributed query processor
  - Syntactic heterogeneity is hidden behind OGSA-DAI wrappers
- A prototype of a system that extends OGSA-DQP functionalities in order to support data integration based on XMAP schema-mapping

XMAP-DQP System Model



XMAP-DQP layered architecture



System Functionalities

- We modified the DQP coordinator to:

1. Locate the XMAP-WS forwarding it the XPath query

2. Wait until the XMPA-WS produces a set of semantically equivalent queries (*Query Reformulation*)

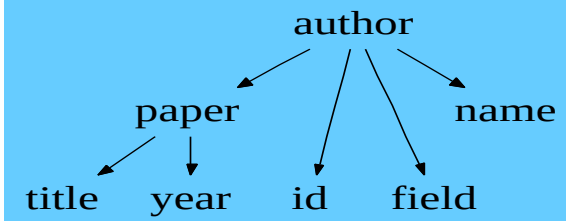
3. Call a module that given an XPath query it produces the equivalent OQL one (*Query Transformation*)

Query Reformulation: XMAP-WS

- The XMAP algorithm is wrapped as a web service (*XMAP-WS*)
  - The XMAP-WS sevice contains mappings information
  - By a single call to it is possible to retrieve all the semantically equivalent XPath queries

Database schemas

Site S1



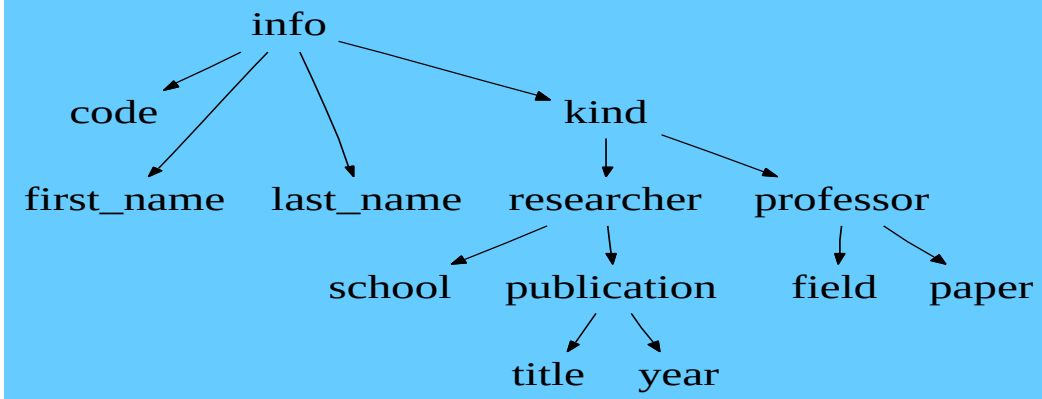
author

id	field	name

paper

paper_id	title	year

Site S2



info

code	first_name	last_name	kind	id

researcher

info_id	researcher_id	school

publication

researcher_id	title	year

professor

info_id	field	paper

Database Metadata

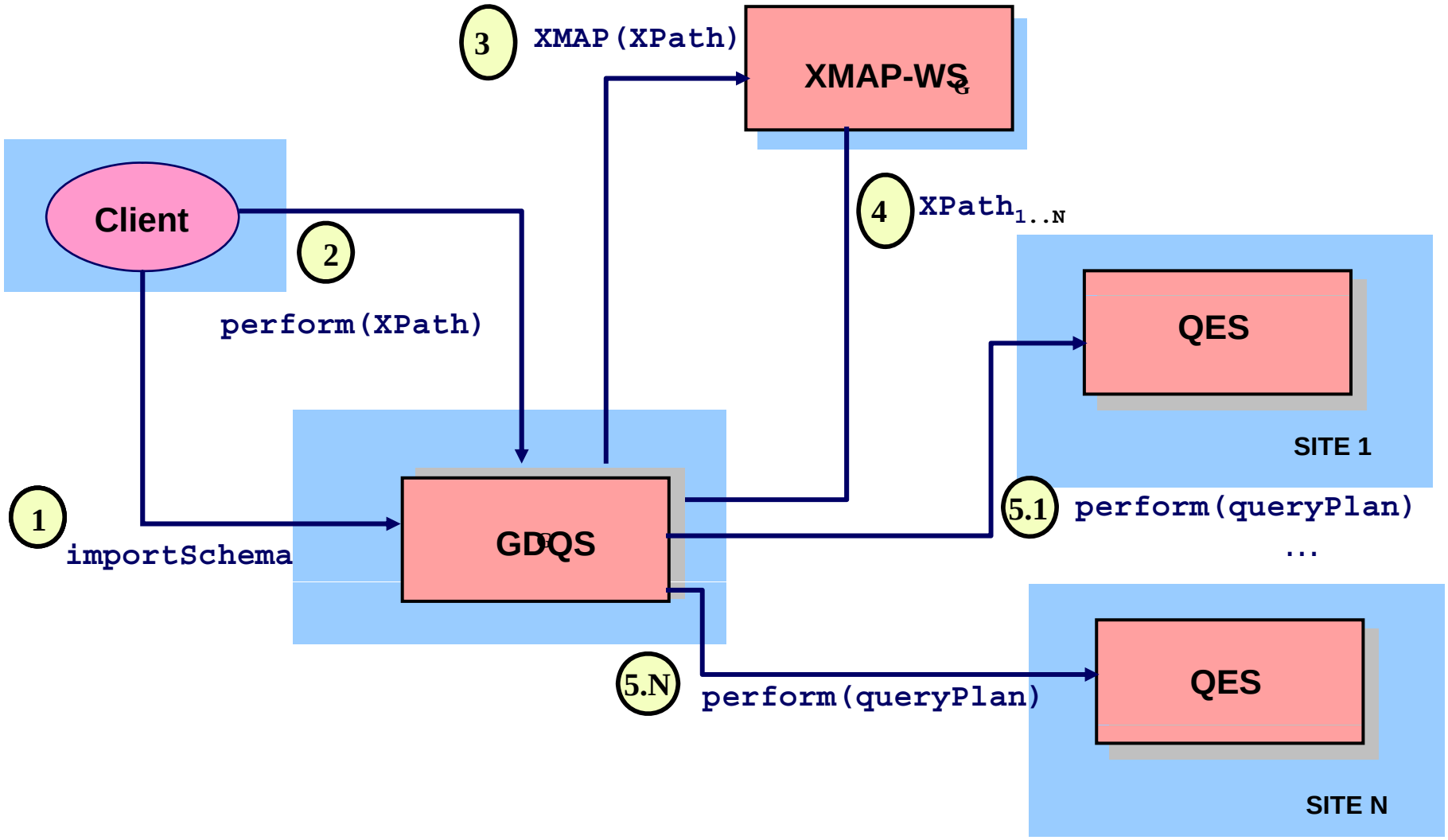
```
<databaseSchema dbname="S1">
    <table name="Author">
        <column name="id" />
        <column name="Field" />
        <column name="name" />
        <primaryKey>
            <columnName>id</columnName>
        </primaryKey>
    </table>
    <table name="Paper">
        <column name="title" />
        <column name="year" />
    </table>
</databaseSchema>

<databaseSchema dbname="S2">
    <table name="Info">
        <column name="id" />
        <column name="code" />
        <column name="first_name" />
        <column name="last_name" />
        <column name="kind" />
        <primaryKey>
            <columnName>id</columnName>
        </primaryKey>
    </table>
```

```
<table name="Researcher">
        <column name="researcher_id" />
        <column name="info_id" />
        <column name="school" />
        <primaryKey>

<columnName>researcher_id</columnName>
        </primaryKey>
    </table>
    <table name="Publication">
        <column name="publication_id" />
        <column name="title" />
        <column name="year" />
        <primaryKey>
            <columnName>title</columnName>
        </primaryKey>
    </table>
    <table name="Professor">
        <column name="info_id" />
        <column name="paper" />
        <column name="field" />
    </table>
</databaseSchema>
```

Service Interactions: Distributed Query Processing

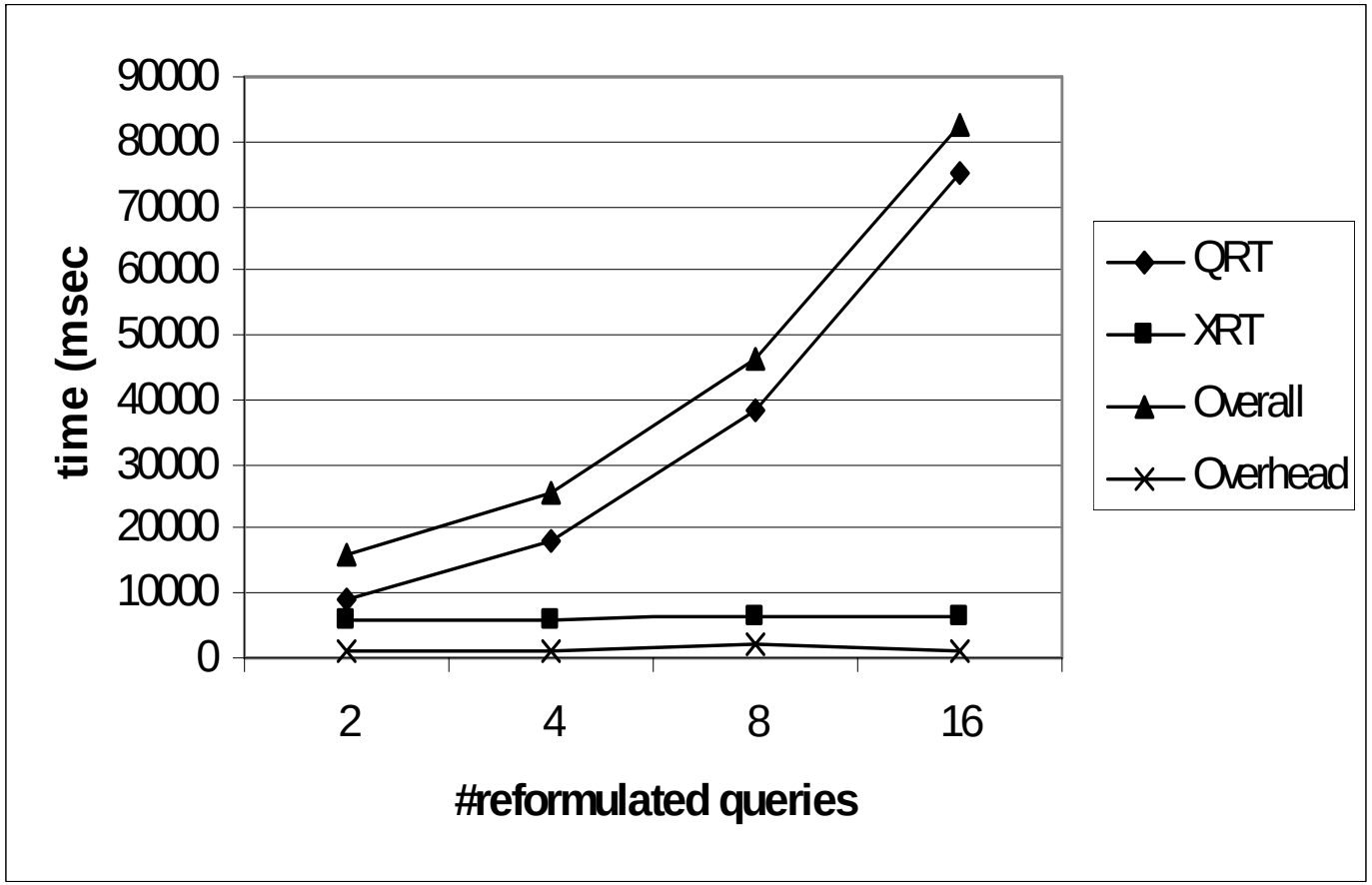


Experimental Results (1)

#T	QRT	XRT	Overall	OH (%)
2	1726(102.8)	2056(17.8)	3872(123)	90.2(2.33%)
4	2948(130.8)	1977(2.65)	5010(132.2)	85.4(1.7%)
8	5613(129)	2031 (7.2)	7730 (123.7)	86.6(1.12%)
16	11670(566)	2964(16.5)	13824(559)	89.2(0.65%)

#T	QRT	XRT	Overall	OH (%)
2	9155(191)	5956(55)	16069(146)	958(5.96%)
4	18308(617)	5947(103)	25344(755)	1089(4.3%)
8	38269(656)	6247(212)	46487(726)	1971(4.24%)
16	75266(1495)	6177(89.3)	82609(1476)	1166(1.41%)

Experimental Results (2)



Summary

- Data Integration is a key issue for exploiting the availability of large, heterogeneous, distributed data volumes on Grids
- Integration formalisms can benefit from OGSA-based Grids
  - Dynamic discovery, allocation, access and use of resources
- The XMAP integration framework provides the architecture, and query reformulation algorithm for data integration.
- OGSA-DAI/DQP provide the service infrastructure.
- XMAP-DQP: a service-based architecture for Grid data integration enabled distributed query processing