Data Integration
The problem of providing uniform source access to users, and eventually updates to multiple (even 2 is a problem) autonomous (not affect the behavior of sources) heterogeneous (different data models, schemas) structured (at least semistructured) data sources (not only databases)

What is Data Integration?
Motivation

Produced by countries in Europe.

- **Culture**: uniform access to all the cultural databases
- **Environment**: Puget Sound Regional Synthesis Model
- **Astrophysics**: monitoring events in the sky
- **Medical Genetics**: integrating genomic data
- **Science & Culture**

**XML Integration**

- Portals integrating data from multiple sources
- Comparison shopping (Netbot, Junglee)

**World-wide Web**

- Enterprise data integration, web-site construction.
Principle Dimensions of Data Integration

Virtual vs. materialized architecture

Access: query only or query/update?

Mediated schema: yes or no?

Mediated schema requires schema integration and then query reformulation.

Without mediated schema, we lose some of the advantages of data integration.

Need distributed transactional services.

Problem similar to updating through views.
Figure 1: Materialization architecture
Figure 2: General data integration architecture
Query Reformulation

Query Execution Engine

Wrapper

Query in the global schema

Distributed

union of exported local schemas

Query in the exported local schema

Query Optimization

Global Data Model

Local Data Model

Query in the local schema

Query in the exported local schema

Query in the exported local schema

Query in the global schema
6. Closed world assumption vs. open world assumption.

- speed
- handling inconsistencies
- machine readable vs. human readable
- completeness
- accuracy

5. Requirements from responses:

4. How structured are the data in the source?

3. How much knowledge do we have about source?

2. How autonomous are the sources?

1. How many sources are we accessing?

Additional Dimensions of Data Integration
Data mining: Discovering properties and patterns in data.

Information retrieval: Keyword search, no semantics.

Similarities at the optimization and execution level:

- Sources are not autonomous.
- Data is distributed a priori.
- Sources are homogeneous.

Distributed databases:

Related Technologies / Problems
(some) Research Prototypes

- WHIRL (AT&T)
- DISCO (INRIA)
- HERMES (U. of Maryland)
- Cargic (IBM)
- SIMS, ARCADIAE (USC/ISI)
- SIMS (Stanford), XMAS (UCSD)
- The Internet Software Ocean / Razor / Tukwila (UW)
- IRIS-DB (Versailles)
- InfoMaster (Stanford)
- InfoMaster (AT&T)
Outline

- Introduction and motivation
- Wrappers (briefly)
  - Semantic integration and source descriptions
  - Schema mappings and reformulation
- Data Integration
  - Modelingsource capabilities
  - Model incomplete
- Optimization (briefly)
  - Execution (mostly in the Tukwila paper)
How intelligent is the wrapper?

Built w.r.t. a mediator?

Where is the wrapper?
Catalogs contain descriptions of:

- mirror sources
- source reliability
- statistics about the data
- physical properties of the source and network
- source completeness
- source capabilities
- logical source contents

Data Integration
User queries refer to the mediated schema.

Sources store data in their local schemas.

Content descriptions provide the mappings between the mediated and local schemas.

Content descriptions describe the mediated schema.

Data Integration
Elements of source descriptions:

Capabilities: Source must receive movie title or director as input, source can perform selections.

Completeness: Source contains all American movies.

Constraints: All movies produced after 1965.

Contents: Source contains movies, their directors, cast.
Desiderada from source descriptions

Distinguish between sources with closely related data.

Because sources are dynamically being added and removed, we can prune access to irrelevant sources.

Be able to find sources relevant to a query: reformulate queries such that we obtain guarantees on which sources we access.

Enable easy addition of new information sources: because sources are dynamically being added and removed.
Problem: Reformulate the user query referring to the mediated schema onto the local schemas.

Given a query $Q$ in terms of the mediated-schema relations, such that it uses only the data source relations, such that

$Q \circ \rho = |Q|$ (i.e., answers are correct) and

provides all the possible answers to using the sources, $Q$ provides all the possible answers to using the sources.

Find a query $Q^*$ that uses only the data source relations, such that

descriptions of the data sources,

a query $Q^*$ that uses only the data source relations,

Given schema onto the local schemas.
Data Integration

Source relations described as concepts in a description logic: SIMS (ISI/USC), Catarci and Lenzerini (Rome).

Mediated schemas viewed over the source relations: TSIIMS (Stanford), HERMES (Maryland), Information Manifold (AT&T), Occam, Razor, Tukwila.

Source relations defined as views over mediated-schema relations:

Approaches to Specification of Source Descriptions
The mediated-schema relations are described in terms of the source relations.

The Global as View Approach

Movies and their years can be obtained from either DB1 or DB2.

Movie reviews can be obtained by joining DB1 and DB3.

Data Integration
Query Reformulation in GAV

Query reformulation is done by rule unfolding.

A containment check shows that the second rule is redundant.

\[ DB_2^{\text{title}}, \text{director}, \text{year} \cong DB_3^{\text{title}}, \text{review} \]
\[ DB_9^{\text{title}}, \text{review} \cong DB_9^{\text{title}}, \text{review} \]
\[ DB_9^{\text{title}}, \text{review} \cong DB_9^{\text{title}}, \text{review} \]

Reformulated query on the sources:

\[ DB_2^{\text{title}}, \text{director}, \text{year} \cong DB_3^{\text{title}}, \text{review} \]
\[ DB_9^{\text{title}}, \text{review} \cong DB_9^{\text{title}}, \text{review} \]

Reformulated query on the sources:

\[ Movie_{\text{title}}, \text{review} \]
\[ \text{year} \]
\[ Movie_{\text{title}}, \text{year} \]

Query: Find reviews for 1997 movies:

Query reformulation is done by rule unfolding.

Query Reformulation in GAV
Every data source is described as a query expression over mediated-schema relations:

\[ \text{Review}(\text{title}, \text{review}) \subset \text{Movie}(\text{title}, \text{year}, \text{director}, \text{genre}) \]

\[ \text{year} \geq 1990 \wedge \text{Director}(\text{American}(\text{director})) \]

The Local AS View Approach
query reformulation

\[ q'(\text{title, review}) : = \neg \forall (\text{title, year, director}) \land \forall (\text{title, revenue}) \]

The reformulated query on the sources:

\[ \exists \text{year} \geq 1990 \land \forall \text{Review}(\text{title, review}) \]
\[ \forall (\text{Movie}(\text{title, year, director, genre}) \land \text{American(director, genre)} \leq 1960 \land \text{genre} = \text{Comedy}) \]

Find reviews for comedies produced after 1950:

query reformulation
Comparision of the Approaches

See [Ullman, ICDT-97] for a detailed comparison.
How do we translate the query on the mediated schema to a
query on the source?

Reformulation Algorithm
Given the local completeness of the sources?

Answer completeness problem: Is the answer to a query complete to sources.

By exploiting source completeness we can avoid useless access to sources.

is complete for \texttt{American(publisher)}.

\texttt{Book(title, publisher, author)}

is complete for \texttt{Year > 1960}.

\texttt{Movie(title, director, year)}

Often, sources are complete, or locally complete.

Incomplete sources require that we look at all relevant sources.

Local Completeness Information
Additions to **Movie** may change the answer to $Q_1$:

```
WHERE $\textit{m.\text{TITLE}} = \textit{s.\text{TITLE}}$ AND $\textit{city} = \textit{Seattle}$
```

Query: Find movies (and their directors) playing in Seattle:

```
SELECT $\textit{m.TITLE}$, $\textit{m.DIRECTOR}$
FROM Movie $\textit{m}$, Show $\textit{s}$
WHERE $\textit{m.TITLE} = \textit{s.TITLE}$ AND $\textit{city} = \textit{Seattle}$
```

Example #1: Incomplete Query Answer

**Relations:**

- Movie(n28title/,director/,year/n29)
- Show(n28title/,theater/,city/,hour/n29)
- Movie(n28year, ,director, ,year)
- Movie(n28title, ,theater, ,city, ,hour)
- City(n29city)
- Year(n29year)
Example # 2: Complete Query Answer

No, completeness is not simply disjointness of the query with the incomplete parts of the database.

```
\text{SELECT} \text{m.DIRECTOR} \\
\text{FROM Movie} m, Oscar o \\
\text{WHERE} m.TITLE = o.TITLE \\
\text{AND} m.YEAR = o.YEAR \\
\text{AND} o.YEAR \geq 1965. \\
\text{WHERE m.TITLE = o.TITLE} \\
\text{AND} o.YEAR 
```

Query: directors whose movies won Oscars after 1965.

Realted:

Oscar(title, year)

Movie(title, director, year)