ANALYSIS AND COMPARISON OF SYSTEMS FOR HETEROGENEOUS INFORMATION RESOURCES INTEGRATION

Tenth All-Russian Science Conference Digital Libraries: Advanced Methods and Technologies, Digital Collections
Dubna, Russia

Session 12: Informational model mapping and resource integration
October 9, 2008

Leonid Kalinichenko, Alexey Vovchenko
Institute of Informatics Problems of RAS
TALK OUTLINE

- Information Integration Problem
- Heterogeneous Information Resources Integration
- Analyzed Integration Systems
- Important Integration Principles and Comparison Criteria
- Results
The current period of IT development is characterized by an explosive process of information models creation.

- **Distributed infrastructures**: OMG, semanticWeb, SOA, digital library, information grid, ...
- **Information models**: data models, workflow models, process service composition models, semantic models
- **Accumulation** of based on such models information resources, the **number** of which **grows exponentially**

- [Dr. Patrick Ziegler](http://www.ifi.uzh.ch/~pziegler/IntegrationProjects.html)
- [http://www.ifi.uzh.ch/~pziegler/IntegrationProjects.html](http://www.ifi.uzh.ch/~pziegler/IntegrationProjects.html)
- 183 Integration Projects
TYPES OF INFORMATION INTEGRATION SYSTEMS

- Data warehousing
- Virtual Data Integration
- Message Mapping
- Object Relational Mapping
- Document Management
- Portal Management
**Data Warehousing**

- Data warehouse – database that consolidates data from multiple sources
- Each resource may have a **DB schema** that **differs** from the **warehouse schema**. So **data has to be reshaped** into common warehouse schema
- Extract-Transform-Load (ETL) tools
  - cleansing operations
  - reshaping operations
VIRTUAL DATA INTEGRATION

- Gives the illusion that data sources have been integrated **without materializing** data
- Offers a **mediated schema** against which users can pose queries
- The **implementation**, often called a query mediator system, translates the user’s query into **queries over the data sources** and integrates the result of those queries so that it appears to have come from a single integrated database
- Resources are **heterogeneous** in that they may use **different database** systems and structure the data using **different schemas**
MESSAGE MAPPING

- **Message-oriented middleware** helps integrate independently developed applications by moving **messages** between them.

- If a broker is avoided through all applications’ use of the same protocol, then the product is called an **enterprise service bus**.

- If the focus is on defining and controlling the order in which each application is invoked, then the product is called a **workflow system**.
OBJECT RELATIONAL MAPPING

- Application programs today are typically written in an object-oriented language, but the data they access is usually stored in a relational database.
- Mapping applications to databases requires integration of the relational and application schemas.
- Differences in schema constructs can make the mapping rather complicated.
- Object-to-relational mapper offers a high-level language in which to define mappings.
- Resulting mappings are then compiled into programs that translate queries and updates over the object-oriented interface into queries and updates on the relational database.
**DOCUMENT MANAGEMENT**

- Much of the **information** is contained in **documents**.
- To promote **collaboration** and avoid **duplicated work** in a large organization, this information needs to be integrated and published.
- **Integration** may simply involve **making the documents available** or integration may mean **combining** information from these documents **into a new document**.
- In some applications, it is useful to extract structured information from documents. The ability to **extract structured information** of this kind may also allow businesses to **integrate unstructured documents**.
PORTAL MANAGEMENT

- One way to integrate related information is simply to present it all, side-by-side, on the same screen.
- A portal is an type of integration in mind.
- Portal design requires a mixture of content management (to deal with documents and databases) and user interaction technology (to present the information in useful and attractive ways).
TALK OUTLINE

- Information Integration Problem
- **Heterogeneous Information Resources Integration**
  - Analyzed Integration Systems
  - Important Integration Principles and Comparison Criteria
- Results
HETEROGENEOUS INFORMATION RESOURCES INTEGRATION

- Information Resource driven approach
  - moving from sources to problems (an integrated schema of multiple sources is created independently of a definition of specific application)
- is not scalable with respect to the number of sources
- does not make semantic integration of sources in a context of specific application possible
- does not lead to justifiable identification of sources relevant to specific problem,
- does not provide the required information system stability w.r.t. evolution of the observation sources (e.g., appearance of a new information source relevant to the problem lead to reconsideration of the integrated schema)
HETEROGENEOUS INFORMATION RESOURCES INTEGRATION (2)

- **Problem driven** approach
  - moving from a problem to the sources (a description of an application subject domain (in terms of concepts, data structures, functions, processes) is created, into which sources relevant to the application are mapped)
- assumes creation of subject mediator that supports an interaction between an application and sources on the basis of the application subject domain definition
- removes the disadvantages mentioned for the approach driven by information sources
INTEGRATION USING VIEWS

- **Global As View (GAV)**
  - According to GAV a global schema is defined in terms of the pre-selected sources

- **Local As View (LAV)**
  - Sources are defined as views over the mediator schema

- **Both As View (BAV)**
  - Based on the use of reversible schema transformation sequences. LAV and GAV view definitions can be fully driven from BAV

- **GLAV**
  - Later a variation of LAV allowing the head of the LAV view definition rules to contain any source schemas query and hence is able to express the case where a source schemas are used to define the global schema constructs (GAV)
TALK OUTLINE

- Information Integration Problem
- Heterogeneous Information Resources Integration
- **Analyzed Information Integration Systems**
- Important Integration Principles and Comparison Criteria
- Results
INFORMATION INTEGRATION SYSTEMS

- Agora
- AutoMed
- Infomaster
- PICSEL
- SIRUP
- Information Manifold
- MedMaker
- SYNTHESIS
AGORA

- **Approach:** LAV
- **Canonical model:** XML
- **Query language:** Xquery
- **Resources:** XML, Relational

Implemented in LaSelect
**AUTOmed**

- **Approach:** BAV
- **Canonical model:** HDM
- **Query language:** AIQL
- **Resources:** Relational, XML, flat files
INFOMASTER

- **Approach:** LAV
- **Canonical model:** KIF
- **Query language:** KQML
- **Resources:** Relational, Z39.50, custom pages
SIRUP

- **Approach:** LAV
- **Canonical model:** ICONCEPT
- **Query language:** SQL-like
- **Resources:** Relational, XML, ontology
**MedMaker**

- **Approach:** GAV
- **Canonical model:** OEM
- **Query language:** MSL
- **Resources:** Relational, Semi-Structured
INFORMATION MANIFOLD

- **Approach**: LAV
- **Canonical model**: CARIN-Classic
- **Query language**: Datalog-like
- **Resources**: XML, Relational, semi-structured, ...
**PICSEL2**

- **Approach:** LAV
- **Canonical model:** CARIN KB
- **Query language:** CARIN (Datalog like)
- **Resources:** Services
SYNTHESIS

- **Approach:** LAV
- **Canonical model:** SYNTHESIS
- **Query language:** Syfs

- **Resources:** XML, services, Relational, Objec-Relational, e.t.c.
TALK OUTLINE

- Information Integration Problem
- Heterogeneous Information Resources Integration
- Analyzed Information Integration Systems
- **Important Integration Principles and Comparison Criteria**
- Results
IMPORTANT INTEGRATION PRINCIPLES

- ASME Criteria
  - Abstraction
  - Selection
  - Modeling
  - Explicit Semantic

- Principles
  - Integration **Approach**
  - **Extensible** Canonical Informational Model
  - Semantic Schema Matching
  - Problem solving specification
ASME CRITERIA

- **Abstraction** refers to shielding users from low-level heterogeneities and underlying data sources.
- **Selection** means the possibility of user-specific selection of data and data sources for individual integration.
- **Modeling** corresponds to the availability of means to incorporate user-specific ways to perceive a domain of interest for which integrated data is desired in the process of data integration.
- **Explicit semantics** refers to means for explicitly representing the real-world semantics of data.
INTEGRATION PRINCIPLES

Integration Approach
- LAV removes the disadvantages of GAV
- Abstraction + Modeling = Approach (LAV, GAV, ...)
- Criteria – Approach (“A”)

Extensible Canonical Informational Model
- Resources are heterogeneous, so the unification of resources models in the frame of some unifying information model called canonical is required
- Unification requires a technique of matching the specifications of various resources
- Refinement relation: It is said that specification A refines specification D, if it is possible to use A instead of D so that the user of D does not notice this substitution
- Criteria – Unification (“U”)
- Criteria – Selection (“S”)
INTEGRATION PRINCIPLES (2)

- **Semantic Schema Matching**
  - Resource Registration require metadata (ontology)
  - Criteria – Explicit Semantic (“E”)

- **Problem solving specification**
  - Application domain specification includes: concepts, data structures, functions, processes
  - Criteria – Functionality (“F”)
  - Architecture Extensibility
  - Criteria – Hybrid (“H”)
  - User Friendly Integration Tools Availability
  - Criteria – Tools (“T”)


COMPARISON CRITERIA

• AUSEFHT
  • Approach
  • Unification
  • Selection
  • Explicit Semantic
  • Functionality
  • Hybrid
  • Tools
## RESULTS

<table>
<thead>
<tr>
<th>System</th>
<th>A</th>
<th>U</th>
<th>S</th>
<th>E</th>
<th>F</th>
<th>H</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agora</td>
<td>LAV</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>AutoMed</td>
<td>BAV</td>
<td>Yes</td>
<td>No</td>
<td>Partially</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Infomaster</td>
<td>LAV</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>SYNTHESIS</td>
<td>GLAV</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>PICSEL</td>
<td>LAV</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>SIRUP</td>
<td>LAV</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Information Manifold</td>
<td>LAV</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>MedMaker</td>
<td>GAV</td>
<td>No</td>
<td>Yes</td>
<td>Partially</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>
CONCLUSION

- **SYNTHESIS** – ex facte Excellent Project
- **MedMaker** – is interesting, cause automatic mediator generation
- **AutoMed** – is interesting, cause BAV views, and their transformation into LAV or GAV views. HDM, model mappings (Relational, XML, ER, UML, ORM), inter model transformation.
- **SIRUP** – ontology oriented approach. AIQL query.
- **PICSEL** – service integration oriented approach.

- Criteria must be wider
- More projects must be analyzed