

## Chapter 1

# Meta-Model Based Information Mediation

Lutin Zhao  
Philips Research, USA

Keng Siau  
The University of Nebraska at Lincoln, USA

### ABSTRACT

*Information mediation is one of the major approaches to solve interoperability problems related to heterogeneous information integration. This paper first discusses the concept of information mediation and typical mediation architecture. Two major mediation research projects, TSIMMIS and MIX, and their limitations, are discussed. Meta-model, a way for exchanging meta-data, is then introduced for the purpose of improving information mediation. Finally, a meta-model based mediation approach is proposed.*

### INTRODUCTION

Information mediation is a research area that deals with integrating information from different, usually heterogeneous, information sources, including regular databases, XML source, record files, email systems, etc. The software that handles or masks data heterogeneity from end users is called a *mediator*.

In the information mediation research community, there are several research projects that have been completed (Molina, 1997; Baru, 1999a). Two of the most important ones are TSIMMIS (by Stanford University) and MIX (by University of California at San Diego). Although both of them use typical mediation architecture, different data models and query languages are used to solve mediation problems.

Exchanging meta-data based on meta-model (meta-meta-data) is always considered a major interoperability solution. Currently this approach is strongly backed up by the emergence of eXtensible Markup Language (XML), which is considered a breakthrough solution for interoperability. For example, Common Warehouse Meta-model (CWM) (OMG, 2001; Poole, 2002) is the first meta-model standard established by the Object Management Group (OMG) to enable the exchange of meta-data, mainly in data warehouse domain using XML. Although still being improved, this standard becomes our major motivation for creating new meta-model driven information mediation architecture.

## INFORMATION INTEROPERABILITY AND MEDIATION

A common problem that the distributed information system faces is the need to integrate heterogeneous information sources, including regular databases, file systems, web database, email system, etc. Business decision makers need to access multiple information sources to gather enough information; but this is usually hindered by information heterogeneity, which includes data model difference, data format inconsistency, data semantic difference, naming inconsistency, etc. Unlike Intranet, data models of distributed information sources are usually unknown (Saelee, 2001). This brings even more interoperability difficulty. The solution to solve heterogeneous data access problems is called information interoperability, including middleware-based interoperability and mediation-based interoperability.

Middleware-based interoperability is similar to other business service interoperability. The basic idea is to encapsulate data access functions into methods and publish them using implementation independent Interface Definition Language (IDL). This type of interoperability is at the *service* level because users only invoke data access methods rather than query data itself. Many commercial products (CORBA products, Microsoft DCOM, etc.) could be used to support this approach.

Mediation-based interoperability provides users with (probably converted) data view and query language for querying heterogeneous information sources. This type of interoperability is considered at the *data* level, in contrast to the *service* level

interoperability (CORBA, DCOM). More specifically, mediation solutions provide users with a way to send on-demand queries to heterogeneous information sources. In other words, in users' eyes, there is a homogeneous (common) view despite the heterogeneous information sources. User queries issued on this view are intercepted by the mediation system and converted to query formats that can be accepted by heterogeneous information sources. Currently there is no good commercial product support for mediation.

## A TYPICAL MEDIATION MODEL

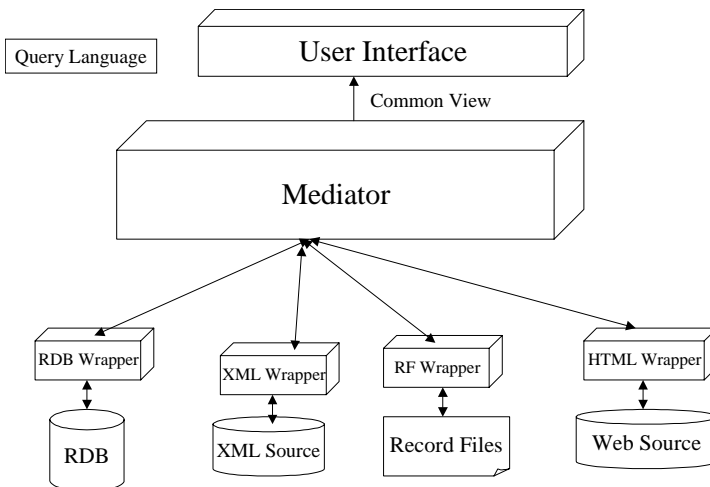
Figure 1 shows a typical information mediation model. To make mediation work, the following requirements must be met:

- **Common data model (view) definition**

The reason for defining a common data model is because heterogeneous information sources have different data models or schemas (e.g., relational database has a relational schema, XML database has a tree-like schema, etc.) Common data model provides users with a common view so that queries can be issued on this view. Molina (1997) stated that the common data model must be able to handle:

- (1) *A rich collection of structures, including nested structures that are found in typical modern programming languages.*

Figure 1: A typical information mediation model



- (2) *Missing information or related information of widely differing structures.*
- (3) *Meta-information, that is, information about the structures themselves and about the meanings of the terms used in the data.*

- **Common query language definition**

Common query language provides a single query language for querying different information sources. According to Molina (1997), the common query language must allow:

- (1) *New mediators to join old ones for augmented functionality.*
- (2) *New sources to provide input to an existing mediator.*

- **Wrapper design**

Wrappers sit on top of heterogeneous sources to export data in a uniform format to the mediator (Baru, 1999a). Wrappers provide access to heterogeneous information sources by converting application queries into source-specific queries or commands. Wrappers also accept users' queries, decide whether they are allowed, translate them into queries that underlying information sources can recognize, and return query results to the mediator by converting results into formats as defined by the data model. In this sense, the wrapper is aware of both the common view and specific information source schema and is able to do conversion between them.

One important thing to note is that a mediation system can have multiple or multi-level mediators that work with certain sets of information sources, because there is no need to provide a global data model representing all information sources. In addition, although information mediation aims at hiding the underlying data models from end users and providing them with a way to access heterogeneous data without even knowing too much about these data, different information integration research projects have different ways to meet the above requirements based on how volatile the data models are.

## PREVIOUS WORK

### TSIMMIS

TSIMMIS (Hammer, 1995; Molina, 1995), referring to The Stanford-IBM Manager of Multiple Information Sources, is a system for information integration.

Figure 2: An OEM Object

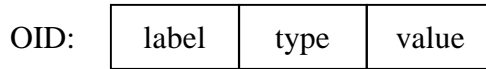
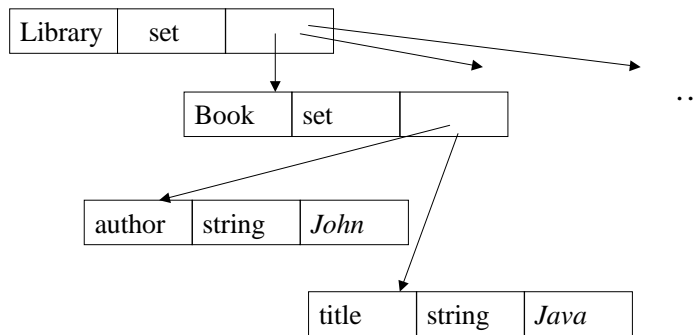


Figure 3: A collection of OEM Objects



It was a joint effort by Stanford University and IBM, and started in the mid-90s.

TSIMMIS offers a common data model and a common query language that are designed to support the integration of information from different information sources. It also offers tools for automatically generating the components that are needed to build systems for integrating information.

### Project highlights

#### (1) OEM (Object Exchange Model) as common object model

TSIMMIS uses a lightweight object model called OEM (Object Exchange Model) to convey information structure. It is a pure semi-structured model (Molina, 1997). Figure 2 shows an OEM object. Figure 3 shows a collection of OEM objects. From both figures, we can see that OEM is good at representing data and relationships in a tree structure.

#### (2) Query language

TSIMMIS provides two semantically equivalent querying languages: MSL and LOREL. Both languages are established on the OEM model. For example, to

find a book authored by “John” from the objects collection in Figure 3, we can form the following queries:

— MSL as query languages

e.g.,

<booktitle X>:-

<library { <book { <title X> <author “John”> }> }>@s1

— LOREL as query language

e.g.,

Select library.book.title

Where library.book.author = “John”

## MIX

The MIX (Mediation of Information using XML) project was a collaborative research project between the University of California at San Diego (UCSD) Database Laboratory and the Data-intensive Computing Environments (DICE) group at UCSD. The goal of this project is to study, develop, apply, and evaluate systems for querying across heterogeneous information sources using XML.

The MIX project considers all information sources as XML sources (Baru, 1999). Consequently, XML DTD (Document Type Definition) was chosen as the common data model (Papakonstantinou, 1999). Simply speaking, what DTD does is to specify some structural and data type constraints regarding how XML documents should be composed. Figure 4 shows a simple DTD example.

In MIX, XML queries are denoted in a high-level, declarative query language called XMAS. For example, XMAS allows object fusion and pattern matching on the input XML data. Figure 5 shows an XMAS query example.

*Figure 4: A DTD example*

---

```
<!ELEMENT item (prodName+,USPrice,shipDate?)
<!ATTLIST item partNum CDATA>
<!ELEMENT prodName (#PCDATA)>
<!ELEMENT USPrice (#PCDATA)>
<!ELEMENT shipDate (#PCDATA)>
```

*Figure 5: XMAS queries*


---

```

CONSTRUCT <Answer> [ <in_region name=$R>
                    <homes> [ $H ] ORDERBY $H.price </homes>
                    </in_region> ]
                    </Answer>
WHERE <home_buyer>
    <homes> $H: <home region=$R beds=$BE baths=$BA area=$A price=$P>
                <nearby_schools>
                    $$S: <school score=$SC/>
                </nearby_schools>
            </home>
        </homes>
<crimes> <yearly_crime>
        <police_service_region name=$R murder=$CM rape=$CRA/>
    </yearly_crime>
</crimes>
    </home_buyer> IN "www.sdsc.edu/MIX/MED-VIEW"
AND $BE=3 AND $BA=2 AND $A>1600 AND $P>250000 AND $P<350000
AND $SC>=70 AND $CM+$CRA=<15

```

---

## EVALUATION OF TSIMMIS AND MIX

There is no doubt that both projects made outstanding contributions to the mediation research within their scope. However, both mediation approaches have some limitations as described below.

### Handle the diversity of data models

Although there might be multiple mediators in a mediation system that handle a set of information sources, the variety of information sources can be significant in terms of different data models. Therefore, the choosing of a common mediation data model is preferably dynamic (dependent upon actual information sources) instead of static, in order to avoid losing too much information during the conversion to the common data model. In this sense, forcing the use of a specific common data model, the approach taken by TSIMMIS and MIX, is not a wise choice. For example, DTD might be a good common data model when the majority of information sources are XML sources, but this is not the case when 90% of information sources are relational databases.

## Meet different query preferences

Information users usually have different preferences for query languages. Therefore, forcing them to use a query language that is specific to the common data model may be inappropriate and error-prone. In addition, issuing queries without knowing the underlying information source structure makes it impossible to formulate appropriate and efficient queries. Users also may not be familiar with the query language.

## Other interoperability concerns

Thinking from a traditional interoperability point of view, interoperability is basically to find a common (interface definition) language for communication between server side and client side, and give users on both ends the flexibility to use their preferred programming languages. Typical examples are CORBA and Microsoft DCOM. Similarly, in data interoperability (mediation) domain, this common language should:

- (1) Allow information sources to publish their data models precisely in a neutral way for communication purposes, which is similar to publishing CORBA services using IDL.
- (2) Allow data users to choose their preferred data models and query languages, which is similar to the flexibility offered by CORBA clients.

Apparently, both OEM from TSIMMIS and DTD from XML are not able to meet the requirements of being this common language, simply because they require data users to follow a single data access method, including a single data model and query language.

In summary, the above analysis suggests a solution that will fundamentally change the design of existing solutions. We propose a meta-model based design to fulfill the above requirements.

## META-MODEL

Simply speaking, meta-data is data that describes data. For example, a UML (Unified Modeling Language) class is the meta-data for a set of objects; relational data schema is the meta-data for relational databases; XML DTD is the meta-data for XML documents. Meta-model (or meta-meta-data) is the meta-data for meta-data. In other words, meta-model is the data that describes meta-data. For example, UML class meta-model defines what a UML class is. Just like meta-data enables the exchange of data, meta-model enables the exchange of meta-data.

Based on the above discussion, it is obvious that in a database domain, meta-data can refer to data models or query languages, while meta-model defines these data models or query languages. Common Warehouse Meta-model (CWM), a new standard adopted by Object Management Group (OMG) recently, is a typical meta-model standard for defining meta-data for all aspects of data warehousing (OMG, 2001). Since CWM is still being improved and it does not contain enough meta-models needed by information mediation, we will use the generic term “mediation meta-model” in the following discussion.

### **Information mediation meta-model**

Database or data warehousing domains can have many meta-models for exchanging meta-data; for example, data models, query languages, warehouse processes, etc. Similarly, for the purpose of information mediation, we only need meta-models that can be used for exchanging:

- (1) Different data models; for example, relational data model, XML schema, record file schema, etc. As an example, CWM provides such capability in the “Resource” level (OMG, 2001).
- (2) Different query languages; for example, SQL, XQL, etc. Currently there is no standardized meta-model for this purpose.

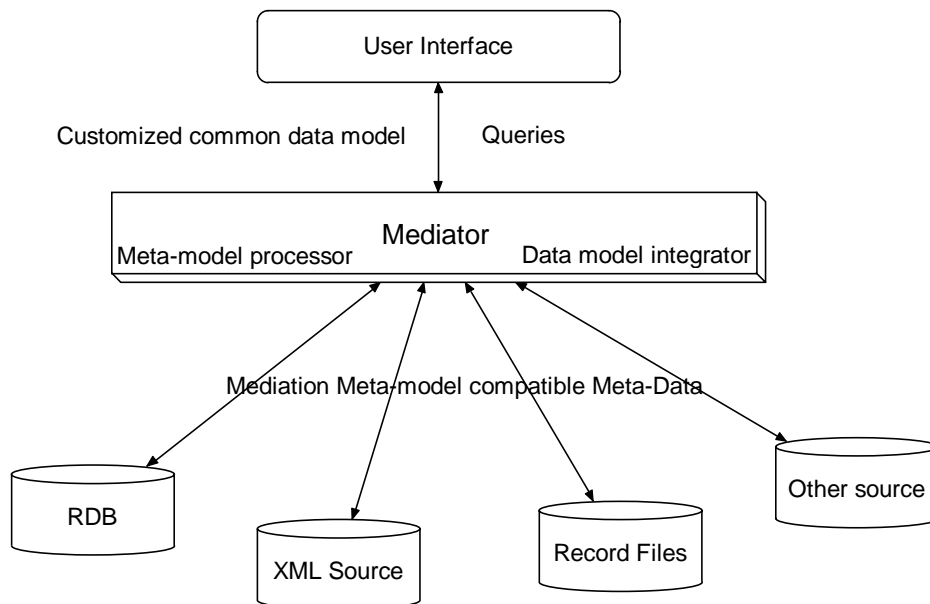
Being meta-model enabled, information sources become capable of providing understandable data models and data queries to other entities using a precise, neutral, standard, and meta-data way. The result is that a “tighter” common data model and appropriate query language could be chosen.

## **META-MODEL BASED MEDIATION**

As a consequence of the above discussion, we propose the meta-model based mediation as shown in Figure 6. This architecture assumes that underlying information sources are all “meta-model aware.” Being “meta-model aware” can eliminate the use of wrappers.

Instead of one specific OEM or DTD (meta-data), a mediation meta-model (meta-meta-data) is used as the common communication language between mediator layer and information source layer. That is, all meta-data (encoded data models) transferred over the network must conform to the mediation meta-model specification.

*Figure 6: Meta-model based mediation model*



The following describes a sample scenario. First, information sources provide mediation meta-model encoded data model to the mediator. In this way, the mediator can gather data model information from various data sources. Second, the mediator user chooses what common data model he (she) would like to use. Because the mediator has model information from all data sources, it is able to transform those data models to a common model chosen by the user. The conversion is based on pre-defined mapping rules. As an example, Tan (2000) proposed a MOF based meta-data solution for database schema integration. Finally, the user issues appropriate queries based on the common data model. The queries will be encoded again using mediation meta-model and transferred back to corresponding information sources.

In summary, instead of defining a specific common data model/query language for users, meta-model based mediation allows users to choose their preferred data models and query languages that are stronger or more appropriate for real cases. For example, if a user is working on information sources containing 90% relational databases and 10% other information sources, the user may choose to use relational data model and SQL as the common representation rather than XML or XQL.

On the other hand, handling complexity may become an issue in the meta-model based mediation approach. Similar to the mapping of programming language elements in CORBA, the mediator must maintain a set of data model mappings among different information sources. This number could be huge, with the increasing number of information sources. But in real implementations, we can reduce or customize the number of mappings to fit the specific needs.

## CONCLUSION

The increasing use of meta-model, especially in a database domain, provides a new approach to solving information mediation problems. As opposed to previous information mediation projects that used a mediation model with a pre-defined single data view and query language, this paper proposes a meta-model based mediation model that separates information source from mediator, therefore allowing users to choose their preferred data views and query languages.

## REFERENCES

- Baru, C. (1999). Xviews: XML views of relational schemas. *Proceedings of the 10<sup>th</sup> International Workshop on Database and Expert Systems Application*, 700-705.
- Baru, C., Gupta A., Ludäscher, B., Marciano, R., Papakonstantinou, Y., Velikhov, P., & Chu, V. (1999). XML-based information mediation with MIX. *Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data*, 28(2), 597-599.
- Hammer, J., Garcia-Molina, H., Ireland, K., Papakonstantinou, Y., Ullman, J., & Widom, J. (1995). Information translation, mediation, and mosaic-based browsing in the TSIMMIS System. *Proceedings of the 1995 SIGMOD International Conference on Management of Data*, 483.
- Molina, H.G., Hammer, J., Ireland, K., Papakonstantinou, Y., Ullman, J., & Widom, J. (1995). Integrating and accessing heterogeneous information sources in TSIMMIS. *Proceedings of the AAAI Symposium on Information Gathering*, 61-64.
- Molina, H.G., Papakonstantinou, G., Quass, D., Rajaraman, A., Sagiv, Y., Ullman, J., Vassalos, V., & Widom, J. (1997). The TSIMMIS approach to mediation: Data models and languages. *Journal of Intelligent Information Systems*, 2, 117-132.
- OMG, (2001). Common Warehouse Meta-model standard, <http://www.omg.org/technology/cwm>

- Papakonstantinou, Y., & Velikhov, P. (1999). Enhancing semistructured data mediators with document type definitions. *International Conference on Data Engineering (ICDE99)*.
- Poole, J., Chang, D., Tolbert, D., & Mellor, D. (2002). *Common Warehouse Meta-model - An Introduction to the Standard for Data Warehouse Integration*. Wiley Computer Publishing.
- Saelee, M., Beitzel, S., Jensen, E., Grossman, D., & Frieder, O. (2001). *Intranet mediators: A prototype*. *Proceedings of International Conference on Information Technology: Coding and computing*, 389-393
- Tan, J., Zaslavsky, A., & Bond, A. (2000). Meta object approach to database schema integration. *Proceedings of International Symposium on Distributed Objects and Applications*, 145-154.