A Framework for a Multidimensional OLAP Model using Topic Maps

Robert M. Bruckner[1], Tok Wang Ling[2], Oscar Mangisengi[2], A Min Tjoa[1]

[1]Vienna University of Technology
Institute of Software Technology
Favoritenstr. 9-11 /188, A-1040 Vienna, Austria
{bruckner, tjoa}@ifs.tuwien.ac.at

[2]National University of Singapore
School of Computing
S16, Level 5, 3 Drive 2, Singapore 117543
{lingtw, oscar}@comp.nus.edu.sg

Abstract

The goal of a data warehouse is to integrate applications at the data level. Data warehouse architectures are developing in response to our increasing data and information requirements. The traditional notion of data warehouses is evolving into a federated warehouse augmented by a set of processes and services to support integrated and consistent access to heterogeneous, decentralized warehouse systems.

The evolution of data warehouses into knowledge repositories requires an architecture supporting the information acquisition from heterogeneous sources. In this paper we propose a framework using XML Topic Maps (XTM) as a foundation for the combination of Web OLAP and data warehouse resources by integrating schema information and addressing semantic heterogeneity. This aims at providing, managing and exploiting a set of integrated data warehouses for knowledge management and decision support.

Index Terms: Data Warehouse, OLAP, XML, Topic Maps, Interoperability, Ontologies

1 Introduction

Providing data, information (data with meaning in some context), and knowledge (information that is readily accessible to its user) simultaneously becomes an active research issue in the area of information systems and intelligent information integration on the Web. This issue involves various, disparate scientific domains (e.g. knowledge management, databases, statistics). Ontologies will play an important role in supporting information exchange processes by revealing implicit assumptions [8]. An ontology provides a shared and common understanding of a domain that can be communicated between people and across application systems.

On the other hand, in the database community data warehouses (DWH) and on-line analytical processing (OLAP) [5] emerged as enabling technology for decision support systems. A combination of ontologies, OLAP warehouses, and knowledge management will become an active research area in the near future for providing distributed strategic decision support system on the Web.

The challenges in the area of information systems and intelligent information integration involve technical, syntactic, and semantic integration [7]. An issue that goes beyond syntactic integration is the mapping of the semantics of terms from different information resources, even when these terms have been expressed using the same syntactic structure: e.g. even when two applications use XML as their interchange format, how can we be sure that they use the same vocabulary, and that the words in this vocabulary have the same meaning.

The Extensible Markup Language (XML) developed by W3C evolved as a foundation for data exchange and is used to provide metadata markup. However, XML has a limited capability to describe the relationship (schemas) with respect to objects [7]. Additionally, XML has only achieved some degree of syntactic, but not semantic interoperability [1]. Topic maps emerge as a promising solution for organizing and navigating information resources on the Web, and provide a bridge between the domains of knowledge representation and information management. They are designed to provide a knowledge layer – independent of the information resources themselves – to enable the integration of information that spans multiple, disparate repositories.

Interoperability and integration of data sources are becoming more important as both, the amount of data and the number of data producers are growing. Interoperability not only has to resolve the differences in data structures; it also has to deal with *semantic heterogeneity*. Semantics is the interpretation people attribute to data according to their understanding of the world. Different interpretations of data cause semantic heterogeneity.

Although the goal of DWHs is to create a centralized and unified view of enterprise data holdings, this goal has not been fully realized [13]. Many factors contribute to this, e.g. semantic heterogeneity, terminology conflicts. However, organizations wish to share their data according to well-defined sharing agreements (e.g. Official Statistics in the European Statistical System [16]). Therefore, we need to establish federated architectures to accomplish this integration of heterogeneous systems.

In this paper we propose a framework for a multidimensional OLAP data model using topic maps for integrating the information stored in distributed data warehouses. The framework addresses the semantic integration problem by describing the mapping between topic maps of local data warehouse resources and integrating them to global topic maps. This abstraction layer hides complexity from the users of this system (analysts, knowledge workers) and simplifies the access by providing a portal view (global topic maps).

The remaining paper is organized as follows. Subsequent to a description of related work in Section 2, we will give motivating examples and describe the basic architecture in Section 3. We then explain topic map concepts in Section 4. We will assemble the framework in Section 5. Finally, we conclude and outline our future research directions and projects.

2 Related Work

The DARPA Agent Markup Language (DAML) [6] developed by DARPA aims at developing a language and tools to facilitate the concept of the *Semantic Web* [19]: the idea of having data on the Web defined and linked in a way that it can be used by machines not just for display purposes, but for automation, integration and reuse of data across various applications. The DAML language is being developed as an extension to XML and the Resource Description Framework (RDF). The latest extension step of this language (DAML + OIL - Ontology Inference Layer) provides a rich set of constructs to create ontologies and to markup information so that it is machine readable and understandable.

The Ontology Inference Layer (OIL) [7], [9] from the On-To-Knowledge Project is a proposal for such a standard way of expressing ontologies based on the use of web standards like XML schema and RDF schemas. OIL is the first ontology representation language that is properly grounded in W3C standard such as RDF/RDF-schema and XML/XML-schema. DAML and OIL are general concepts not specifically related to database or data warehouse interoperability.

In the research area of formal ontologies the authors of [8] propose the use of similarity relations in order to find out whether (and how) elements from different schemas are related (involving human supervision). They then use reasoning systems (such as PowerLoom) to merge ontologies – but they do not try to detect or resolve mismatches between ontologies.

There has already been done some research on DWH interoperability: In [15] the authors describe the usage of XML to enable interoperability of data warehouses by an additional architectural layer used for exchanging schema metadata. Distributed DWH architectures based on CORBA [10], and centralized virtual data warehouses based on CORBA and XML [2] have been proposed recently. All of these approaches propose distributed data warehouse architectures. They define a kind of restricted data and metadata interchange format using particular XML terms and RDF extensions, respectively. Basically, they achieve syntactical integration - but these concepts do not address semantic heterogeneity to enable a thorough description of mappings between federated, heterogeneous data warehouse systems.

3 Motivation

Consider a star schema [12] for an OLAP warehouse for retail information consisting of a fact table that contains tuples for each item sold in a sales transaction; and a set of dimensions (location, product, and time), which contain geographical store descriptions, product information, and time information, respectively.

```
Sales    (PID, SID, TID, qty)
Time     (TID, day, month, year)
Product  (PID, name, category, subcategory)
Location(SID, store, district, city, state,
          country)
```

Figure 1: Retail DWH

In general dimension are structured hierarchically. The hierarchy *store* → *district* → *city* → *state* → *country* characterizes the dimension *location*. Integrating this DWH with another DWH containing a location dimension and then providing a portal view (abstraction layer) for the users can lead to four conflicts we have to deal with:

a) Dimension hierarchies or levels with the same name, but different resources (identical dimension schemas). This issue will be discussed in subsection 3.1.
b) Different levels of detail for equivalent dimensions (e.g. time dimension: *month* → *year; day* → *month* → *year*)
c) Dimension levels with the same name but different meanings (like *Gregorian year*, *Chinese year*, etc.)
d) Different dimension hierarchy names but the same meaning (e.g. dimension level "*week*" in English, "*Woche*" in German).

We will address the issues b), c), and d) in subsection 3.2 (heterogeneous dimension schemas).

3.1 Distributed DWH architecture with identical dimension schemas

In the case of *equivalent* dimensions a federated DWH access is possible using some kind of communication infrastructure (e.g. CORBA, XML, RDF). From a technical viewpoint a centralized user portal distributes all queries to every DWH system in the specified Web warehouse architecture. The query results are collected and integrated. This combination process of result sets from distributed warehouses can be done in various ways in order to provide a single, integrated view for the users:

- The basic approach is to use the federated topology as an additional result dimension *dwh-source* based on the basic hierarchy: *local-dwh* → *all*. The *dwh-source* dimension is a virtual dimension that can be established at the global layer using the distribution information of the local DWHs (expressed in the hierarchy level *local-dwh*). Therefore it is possible to drill-down to a particular local DWH or to roll-up to the global view.
  Example: Consider two physically separated DWH systems with identical schemas, e.g. European Sales, Asian Sales. Then OLAP queries can be answered at the global level (by aggregating fact measures from all local DWH systems; hierarchy level: *all*) or by drilling down to a particular local DWH (hierarchy level: *local-dwh*) of the *dwh-source* dimension.
- An alternative approach is to enhance an existing dimension using the locality information from the physical distribution of the participating DWH systems.
  Example: Consider two physically separated, but identically modeled DWH systems; one for Europe and the other one for Asia (i.e. European Sales DWH and Asian Sales DWH with identical schemas). The locality information (*Europe, Asia*) is transformed into a new virtual dimension hierarchy level (*continent*) enhancing the original *location* dimension from Figure 1: *store* → *district* → *city* → *state* → *country* → ***continent.***
- Another option is to hide the distributed DWH systems architecture from the users and simulate one single (virtual) DWH environment. This can be characterized as a sort of data partitioning technique.
  Example: Consider two physically separated, but identically modeled DWH systems: the first one contains all retail data till December 2000, and the other one contains data starting from January 2001. From the user's viewpoint there should be only a single time dimension representing the whole history. The integration layer will hide the physical data partitioning.

Data warehouse systems are usually considered to be separate standalone decision-support systems tightly coupled with integrated business intelligence applications (e.g. OLAP servers). We will call this an "all-in-one" architecture, because these systems are running at a single site (or even on a single machine), are not distributed, and tightly integrate analytical applications. This makes it difficult to replace one of the components of the entire DWH system.

This traditional, all-in-one DWH architecture is a *single point of failure* - a solution that is not well suited to the high-availability needs of a mission-critical analytical environment. The above presented integration approaches enable organizations to distribute their DWH resources. This distribution of DWH resources will positively impact the availability of the analytical environment.

- Redundant computer systems with automatic failure detection and automated fail over will facilitate higher fault tolerance (hardware, network, software).
- Partitioning of data to multiple DWH systems enables smaller loading windows for every particular DWH system.
- Geographical distribution can be used to minimize the effects of disasters (fire, power outage, earthquake, etc.) to the analytical environment of an organization.
- Disaster recovery after catastrophic failures is possible at different sites in parallel.

The proposed framework in this paper supports distributed DWH systems with identical schemas. Furthermore, it addresses *semantic heterogeneity* in order to cope with heterogeneous DWH schemas as motivated in the next subsection below.

3.2 Integration of heterogeneous DWH dimension schemas

```
Sales    (PID, SID, TID, qty)
Time     (TID, day, week, year)
Product  (PID, name, category)
Location(SID, store, district, city, state, country)

Sales    (PID, SID, TID, qty)
Time2    (TID, hour, day, month, year, decade)
Product  (PID, name, category)
Location(SID, store, district, city, state, country)

Sales    (PID, SID, TID, qty)
Time3    (TID, Chinese day, Chinese month, Chinese year)
Product  (PID, name, category)
Location(SID, store, district, city, state, country)

Sales    (PID, SID, TID, qty)
Time4    (TID, Tag, Woche, Jahr)[1]
Product  (PID, name, category)
Location(SID, store, district, city, state, country)
```

Figure 2: Retail DWHs with heterogeneous time dimensions

At the beginning of Section 3 we identified three levels of heterogeneity for federated DWH schemas. The *time* dimension (in Figure 2) is a very good example to explain

[1] German dimension levels:
Tag ≡ day, Woche ≡ week, Jahr ≡ year.

why syntactic integration is not always sufficient and to motivate semantic integration.

Integrating these differently modeled DWHs using an XML/RDF based communication infrastructure is difficult, and requires following steps:

- Define the syntactic integration layer using a common, canonical data model [15], in order to express all schema objects available in different local schemas.
- Define the terms in the communication language by using e.g. XML DTDs, which provide a neutral model for describing data structures.
- Provide a mapping from the common data model to every local DWH using mediators. The local mediators are used to exchange data and interchange metadata from the local layer to the federated layer, which provides the global schema.

The main issues in traditional architectures using syntactic integration for these federated Web OLAP warehouses are how to distribute the queries to the heterogeneous local OLAP DWHs, and how to integrate information from other, differently modeled DWHs. The integration of heterogeneous DWH dimension schemas therefore requires a powerful semantic integration. In a federated (heterogeneous) DWH environment, we cannot assume that each participating local DWH system can be extended so that each is aware of the relationships among different time granularities (see examples in Figure 2). Indeed, the database management system may not have an extensible time granularity system [3]. Furthermore, some time granularities may be of interest only to some global users who draw information at the federated level utilizing the globally integrated schema. Therefore the combination of federated DWHs in order to process a query requires a unifying framework, which addresses and resolves syntactic and semantic mismatches.

Figure 3 shows an overview of the architecture that we have in mind for the integration of distributed Web data warehouse resources using topic maps. Each of the DWHs involved can model and store its data independently. This way, DWH sources with different schema-based data representations can be integrated loosely coupled. A schema integration process based on merging local topic maps will generate global topic maps.

The main difference between a mediator based approach and the proposed integration using topic map is that mediators provide particular point-to-point integration solutions, where it is difficult for DWH experts to exchange knowledge with experts from other domains (e.g. time granularities), because they have no common language. However, topic maps are *particularly designed for the representation and machine-based interpretation of semantic dependencies* as investigated by [20]. Domain experts (who are not required to know DWH concepts) can therefore organize knowledge according to semantic categories (e.g. conversions between different time granularities). This can be utilized by the proposed architecture to generate and maintain mediators.

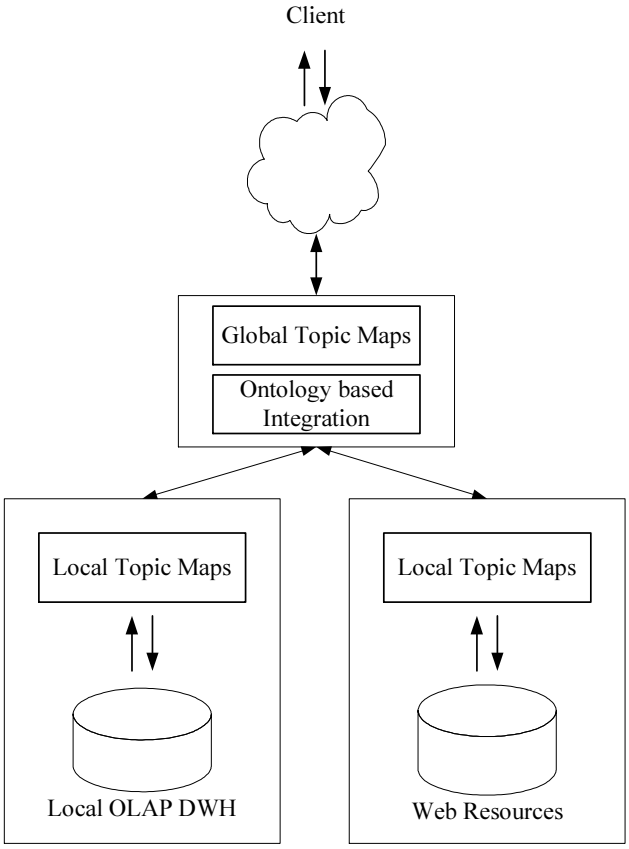

Figure 3: XTM DWH architecture

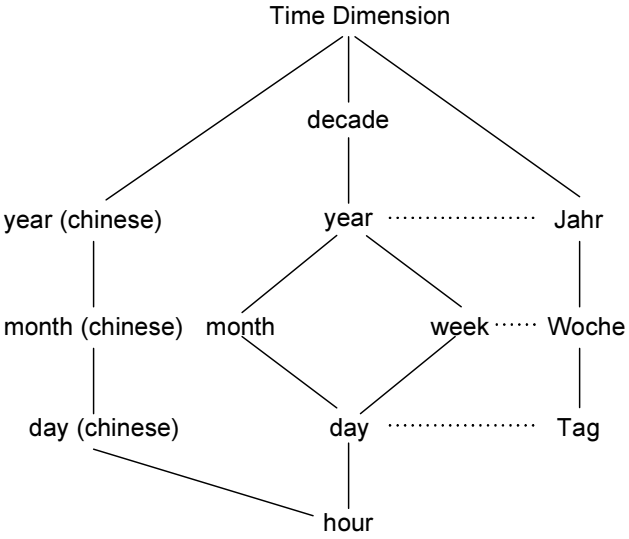

Figure 4: Globally unified time dimension

Figure 4 exemplifies the globally unified time dimension (based on the schemas of Figure 2) that is

provided to answer queries at the federated level. This graphical simplification of the global topic map can be generated automatically just from the local topic maps. In order to provide the mappings between "Jahr" and "year", "Monat" and "month", etc. external topic maps from time granularity domain experts can be integrated or done manually.

4 Topic Map Concepts

There are some approaches in order to make the *Semantic Web* vision a reality, but they are not competing approaches per se. However, they overlap in the sense that all of them seek to create so-called metadata descriptions of relationships. *Topic Maps* lean toward navigation, whereas RDF focuses on inferencing. RDF is rather centered around the resources instead of around the cross-resource knowledge structures. Moreover, it is important to note that only a topic map processor can do useful interpretations, since RDF does not deal with the semantics [1]. In order to make topic maps applicable to the Web, the XML Topic Maps (XTM) standard [11] has been proposed. There are efforts to provide interoperability for topic maps and RDF data [14], [17], where the internal representation of topic maps can be queried by an RDF query infrastructure.

Topic maps can be used as a format for the representation of multi-dimensional subject-based indices for document collections. Furthermore, they provide a format for interoperable knowledge representation, starting with the abstract layer and (optional) links to resources. One can organize knowledge according to semantic categories and aid others in navigation. This metadata is a structured view over a set of information resources that itself need not to be structured. The structure explicitly models an access structure to the knowledge contained in a collection. The XTM standard allows - by defining a syntax - to interchange the information necessary to collaboratively build and maintain indexes. Semantic networks (such as thesauri and more formal ontologies, which are more powerful than topic maps) can be merged as XML structures, and raw data (documents) can be associated with topic maps.

The semantics behind the graph representation of topic maps defined on the object layer is that of three kinds of subjects, defined as anything that can be referred in a human discourse: *topics*, *associations*, and *occurrences*.

A single topic map instance will manage many topics and topic associations. The properties of topic maps are defined as follows: A *topic* is a compound information object - represented and processed by a computer - that serves as a hub to which everything about the subject that the topic represents is connected. The *topic name* consists of three parts: the base name, the display name, and the sort name. Only the *base name* is required. Examples of topics are (base names): Austria, Tyrol, Innsbruck, etc.

A topic should have one or more *topic types*. Topic types are a typical class-instance relation and they are themselves defined as topics by the standard. Having topic types as topics the expressive power of topic maps is used to say more about the type. Examples of topic types are: country, state, city, customer, etc.

Every topic has two characteristics (or at least one of them): a topic name and an *occurrence*. A *topic occurrence* is a resource declared to contain some kind of additional information about the topic. The topic occurrence is the structure that associates resources with topics. The linked resource is typically an information object outside the topic map. Examples of occurrences are: chart of Austria, article about Tyrol, video about Innsbruck, etc.

Every occurrence belongs to one *occurrence role type*. Occurrence role types are – as topic types – themselves topics. Examples of occurrence role types are: chart, article, video, etc.

An *association* is a relationship between one or more topics, each of which plays a role as a member of that association. The *role* is a characteristic that can be assigned to it and governed by scope. *The real power of topic maps results from associations between topics*. Examples of associations are: Tyrol *is in* Austria, Innsbruck *is in* Tyrol, etc.

Each association has one *association type*. Examples of association types are: is in, is defined by, etc.

Each topic that participates in an association plays a role. An *association role type* describes the role. Examples of association role types are: state / country, city / state, city / customer. Both association types and role types are again topics.

The concept of *scope* is important to avoid ambiguities between topics and their characteristics. Any assignment of a characteristic to a topic is considered to be valid within certain limits, which may or may not be specified explicitly. The limit of validity of such an assignment is called its scope. A scope is defined in terms of themes and themes are topics. Examples of scopes are: in order to distinguish between two cities (topics) with the same base name – "Vienna" in Austria and "Vienna" in Virginia, USA – the scopes "Austria" and "USA" are assigned to the two topics having the same base name "Vienna".

Merging topic maps requires a way of establishing the identity between disparate topics from different maps. The specification of *identity attributes* on the topic elements that address the same public subject is the explicit solution the standard offers. The other solution is implicitly through the topic naming constraint, which states that any topics that have the same name in the same scope refer to the same subject.

Facets provide a mechanism for assigning property-value pairs to information resources without modifying them. A facet is a property; its values are called facet values. Example: the topic "Vienna" could have a facet "inhabitants" with the facet value "1.6 million".

If we are looking at the class-instance relationship from an object-oriented view, then there is a justifiable demand for a superclass-subclass relationship as well. However, the standard explicitly declares that such a relationship has to be user-defined. Here are the relevant quotes: The topic relationships established by the types attribute are not superclass-subclass relationships. They are only class-instance relationships. Superclass-subclass relationships between topics can be asserted by topic association links that have been user-defined for that purpose.

5 Framework of a Multidimensional OLAP Model

In this section we present a framework for a multidimensional OLAP data model. In this framework we propose the use of the XTM (XML Topic Maps) concept to model metadata of multidimensional OLAP data for the Web. We use topic maps for the foundation of interoperability. First, we will describe the local presentation of DWH schemas using topic maps (bottom-up approach). Then we will describe the integration of heterogeneous schemas in subsection 5.2. We are going to use only predefined XTM tags as proposed by the XTM Standard (`topicMap`, `topic`, `baseName`, `association`, `occurrence`, `topicRef`, etc.). Therefore it will be possible to use tools based on the XTM standard to create, generate, and maintain such XTM descriptions easily. All examples are based on schemas presented in Figure 2.

5.1 Local presentation of DWH schemas

Facts

The multidimensional data model consists of a set of dimensions and at least one fact measure. In this framework one or more facts and a set of dimensions can be modeled as one or more fact topic maps and as a set of dimension topic maps. Therefore the basic description of our example DWH will result in four XTM files (one topic map for the fact schema and three topic maps for the three dimension schemas). The *Sales* fact is modeled as the *Sales* topic map representing a fact schema. The *Sales* topic map, basically defined as the *Sales* topic, is given in Figure 5.

The *Sales* schema consists of a set of dimension identifiers and fact measures. In this model they all are defined as topics. Therefore, the *Sales* topic map (identified by a topic that contains a <baseNameString> element in the XTM representation, given in Figure 7) consists of a collection of topics (i.e., *PID*, *SID*, *TID* and *qty*) given in Figure 6a. Figure 6b shows the relationship between the topic type *PID* and its instances, which provide class/instance relationships. In addition, the topics and the topic type are scoped in the *Sales* topic.

The metadata example for describing the *Sales* topic map, its topics (i.e., product, location, time, and qty), and its instances in XML syntax is given in Figure 7. The syntax is based on the XTM 1.0 Document Type Declaration [18]. The <topicMap> element is used to define the topic map metadata. The <topic id="*sales*"> and <topic id="*PID*"> elements describe the sales schema and the *PID* attribute, respectively. The <topic id = "instanceOf(*PID*)"> element defines the instance of the *PID* topic, whereas the <scope> element defines that the instance(*PID*) topic is in the scope of the *sales* topic. The <topicRef> element provides a URI (Uniform Resource Identifier) reference to a topic. The target of a <topicRef> link must resolve to a <topic> element of a <topicMap> document that conforms to the XTM specification.

Topics
Sales

Figure 5: The *Sales* Topic Map

Topics
PID
SID
TID
qty

(a)

Topics = *instanceOf(PID)*	
TopicType	**Scope**
PID	sales

(b)

Figure 6: Topics for attributes

```
<topicMap>
<topic id="sales">
  <baseName>
    <baseNameString>Sales Schema</baseNameString>
  </baseName>
  …
</topic>

<topic id="PID"
  …
</topic>
<topic id="instanceOf(PID)">
  <instanceOf>
    <topicRef xlink:href="#PID"/>
  </instanceOf>
  <baseName>
    <scope>
      <topicRef xlink:href="#sales"/>
    </scope>
  </baseName>
</topic>
…
</topicMap>
```

Figure 7: Sales.XTM

In order to establish relationships between topics of a set of dimensions and the topic of a fact measure, *topic associations* need to be created to describe their relationship. In addition, each topic association refers to a class that is represented as topic *association role*. Therefore, an association role consists of a set of associations. The associations and association roles are given in Figure 8. The topic association *qty-SID* describes the relationship between the topic *qty* and the topic *location*, as well as *qty-PID*, *qty-TID*. The topic *fact-dim* contains a topic association role that has a set of associations, namely *qty-PID*, *qty-SID*, and *qty-TID*.

Figure 9 shows that the <topic association id="*qty-PID*"> describes the association between *qty* and *PID*, and the <topic id="*fact-dim*"> element provides information on fact and dimension relationships. This association has two members, namely *qty* and *PID*. Each member is represented by the <roleSpec> element.

Topics
qty-PID
qty-SID
qty-TID
fact-dim

Association=*qty-PID*	
Asso-role	**Scope**
fact-dim	sales

Association=*qty-SID*	
Asso-role	**Scope**
fact-dim	sales

Figure 8: Topic associations

```
<topicMap>
<topic id="qty-PID">
  <baseName>
    <baseNameString>Sales Product Relation
    </baseNameString>
  </baseName>
  …
</topic>
<topic id="fact-dim">
  …
</topic>
…
<association id="qty-PID"
  <instanceOf>
    <topicRef xlink:href="#fact-dim"/>
  </instanceOf>
  <member>
    <roleSpec>
      <topicRef xlink:href="#qty"/>
    </roleSpec>
    <topicRef xlink:href="instanceOf(qty)"/>
  </member>
  <member>
    <roleSpec>
       <topicRef xlink:href="#PID"/>
    </roleSpec>
    <topicRef xlink:href="instanceOf(PID)"/>
  </member>
</association>
…
</topicMap>
```

Figure 9: Sales.XTM with association roles

Dimensions & Dimension Hierarchies

In order to describe metadata for dimensions in this framework a dimension of the multidimensional data model is defined as a topic map that represent a dimension schema. This is used to drive OLAP operations (e.g. roll-up, drill-down) to view and analyze different levels of aggregated multidimensional data. The XTM description of the examples in this subsection (Figures 11 and 12) are similar to Figures 7 and 9, respectively. Therefore we present only the conceptual, graphical representation of the topic maps in this subsection.

Figure 10 shows an example topic map of three dimensions (*product*, *location*, and *time*) as topics.

Topics
product
location
time

Figure 10: Schema dimension topic map

The attributes of a dimension hierarchy are defined as topics. For example, the topic map for the *location* dimension consists of topics, such as *store*, *district*, *city*, *state*, and *country* given in Figure 11a. In the XTM representation a topic map is identified by a topic containing a <baseNameString> element: The topic map *location* is defined as topic (<topic id="location">) containing a baseNameString, which defines that the topic *location* is a topic map.

Each topic consists of a set of topic instances. Each instance has a topic type, for example, the topic *Vienna* refers to the topic *city* (furthermore, *city* is a topic type). The topic *city* consists of e.g. a set of cities in Austria (i.e. topic instances), such as Vienna, Salzburg, and Innsbruck. The topics are scoped in the topic map *location* shown in Figure 11b.

Topics
SID
store
district
city
state
country
all

(a)

Topics = *instanceOf(SID)*	
TopicType	**Scope**
SID	location

(b)

Figure 11: *Location* dimension: topics and its instances

In order to allow conveying some kind of relationship between topics representing a framework of a dimension hierarchy, the topic associations between topics must be defined e.g. in the topic map *location*. In addition, the topic association type within the topic association is used for the grouping of topic associations.

For instance, the dimension hierarchy of the dimension *location* is organized into *store* → *district* → *city* → *state* → *country* → *all*. The association between topics in the topic map *location* is given in Figure 12a.

In Figure 12b the association *store-district* provides the relationship between the topics store and district. The associations *store-district*, *district-city*, etc. are instances of the *rolling-up* topic association role. The *rolling-up* provides an operation name to aggregate data from low level to higher-level hierarchies. The *all* in the association *country-all* is used to define the highest level of data granularity. Each association has two topic members, which take part in a relationship between topics representing a dimension hierarchy.

Topics
store-district
district-city
city-state
state-country
country-all
rolling-up

(a)

Association= *store-district*	
Asso-role	**Scope**
rolling-up	location

Association= *district-city*	
Asso-role	**Scope**
rolling-up	location

(b)

Figure 12: *Location* dimension hierarchy: Topics, and associations

5.2 Integration of heterogeneous resources

For the integration of heterogeneous schema resources we refer to Figure 2 as the running example. Integrating a DWH with another DWH containing different time dimension schemas requires to provide a unified portal view (abstraction layer) at the federated level. This can lead to four kinds of conflicts we have to deal with. We already motivated them in Section 3, and in this section we will explain how they fit into the framework:

A) Equivalent dimension schemas in more than one DWH system (dimension levels with equivalent attribute names, but federated into different data warehouse resources). An attribute in the global topic maps will therefore have different warehouse resources (sites). For instance, the *day*, *month*, and *year* attributes map to all DWH resources where they occur.

B) *Domain mismatch*: Different levels of detail in the dimension schemas of the distributed DWH systems for equivalent dimensions (e.g. time dimension: *month* → *year*; *day* → *month* → *year*).

C) *Homonyms*: Dimension hierarchy levels with the same name, but different meanings (semantics). For instance, the Gregorian calendar is the most widely used today, but other granularity systems are still used to measure time. A *month* in the Gregorian year and a *month* in the Chinese year have different meanings, but in the global topic map they are defined as *month* and we can add semantic assumptions [3] how the Gregorian month can be derived from the Chinese one if a user wants to ask a query using this information.

D) *Synonyms*: Different names for attributes (in different dimension hierarchies) having the same semantics. For example, a *week* in a local warehouse resource equals with *woche* (German) in another warehouses, and they are defined as week in the global warehouses.

Issues, such as A) and B), can be resolved with topic maps by assuming that a topic has different resources and the usage of the *topic occurrence* concept that describes resources to which a specific topic is related.

For the mapping of the semantic conflicts in global topic maps, we define the *global-year* topic as topic-type (or as an attribute of a schema). In order to solve this kind of conflict (type C: homonym) the instances of that topic-type consist of topics, such as *global-year*, *chinese-year* and *gregorian-year*. The *global-year* topic has topic occurrences, that link to the resource given in Figure 13. The <ResourceRef> tags describe DWH resources involved and refer to the schema information of Figure 2 (e.g. *resource-2/time2* refers to the time dimension of the second DWH schema).

Topics
chinese-year
gregorian-year
global-year

Topics = *year*	
Occurrence	**ResourceRef**
gregorian-year	resource-2/time2/year
chinese-year	resource-3/time3/year

Figure 13: Linking to occurrences and resources

In Figure 14 the topic <id="*global-year*"> element defines the topic *global-year* in the global topic maps. It consists of two occurrence elements, namely the <occurrence id = "*gregorian-year*"> and the <occurrence id="*chinese-year*"> elements, for integrating federated resources. The <resourceRef> elements within the <occurrence> elements denote resources that can be linked. Example: the XTM element *<resourceRef xlink:href = "resource-2/time2/year">* refers to the year hierarchy level of the time dimension of the second DWH schema in Figure 2.

Furthermore we could manually add links to (external) resources describing a conversion algorithm between the Gregorian and the Chinese calendar. Our research project will focus on defining algorithms to do this sort of conversions on the fly. Example: a user intends to ask a query using the Gregorian calendar on the global layer. The query should be rewritten automatically for those distributed DWH systems containing a time dimensions based on the Chinese calendar.

```
<topicMap>
<topic id="gregorian-year">
  …
</topic>
<topic id="chinese-year">
  …
</topic>

<topic id="global-year">
  <occurrence>
    <instanceOf>
      <topicRef xlink:href="#gregorian-year"/>
    </instanceOf>
    <resourceRef
       xlink:href="resource-2/time2/year"/>
  </occurrence>
  <occurrence>
    <instanceOf>
      <topicRef xlink:href="#chinese-year"/>
    </instanceOf>
    <resourceRef
       xlink:href="resource-3/time3/year"/>
  </occurrence>
</topic>
</topicMap>
```

Figure 14: XTM - topic occurrences for the integration of homonyms

Basically, we solve conflicts A), B), and C) using the topic occurrence concept. A conflict of type D) - different names for attributes (in different dimension hierarchies) having the same semantics (*synonyms*) - can be resolved using the *scope* mechanism. In Figure 4 the exemplified, globally unified dimension hierarchy shows that the dimension level *day* refers to *day* in English and *tag* in German, as well as *week-woche*, and *year-jahr*. In the global topic maps the global-day is defined as a topic and has a topic type role (or as an attribute of a schema). The *day* topic consists of scopes (e.g., day, tag) and each topic scope refers to a subject identity (resource of topic) as shown in Figure 15.

Topics = *day*
Scope
day-en
tag-ge

Topics = *day-en*
SubjectIdentity
resource-1/Time

Topics = *tag-ge*
SubjectIdentity
resource-4/Time4

Figure 15: Topic scopes to resolve synonyms

The metadata example of the topic and its topic scope in XTM syntax for describing different dimension hierarchy names having the same meaning is shown in Figure 16. The <topic id="*global-day*"> element is used for the definition of the global day topic visible at the federated level. The <topic id="*day*"> element describes the day topic for users at the global level. It has two scopes denoted by the <scope> elements. Each scope has a link to its own topic ("*day-en*" is day in English, "*tag-ge*" is day in German). Furthermore, the <topic id="*day-en*"> element and its <subjectIdentity> element provide links to the information resources.

```
<topicMap>
<topic id="global-day">
  <baseName>
    <baseNameString>Global Day</baseNameString>
  </baseName>
</topic>
<topic id="day-en">
  <subjectIdentity>
    <subjectIndicatorRef
      xlink:href="resource-1/Time/day"/>
  </subjectIdentity>
</topic>
<topic id="tag-ge">
  …
</topic>

<topic id="day">
  <basename>
    <scope>
      <topicRef xlink:href="#day-en"/>
    </scope>
    <baseNameString>
      Day in English</baseNameString>
    <scope>
      <topicRef xlink:href="#tag-ge"/>
    </scope>
    <baseNameString>
      Day in German</baseNameString>
  </basename>
</topic>
</topicMap>
```

Figure 16: XTM sample for the description of synonyms

6 Conclusion and Further Research

Interoperability is very important for future data warehouse architectures. We propose a framework for a multidimensional OLAP model and a framework to achieve interoperability between heterogeneous schema models, which enable the joint querying of distributed web data warehouse (OLAP) systems.

We identified four types of semantic heterogeneity that can occur during the integration of distributed data warehouse systems and therefore motivated semantic integration at the federated level. We describe how to use topic maps to deal with these issues. Local topic maps are used as an additional layer for the representation of local schema information, whereas in the global layer topic maps acts as mediator to hide conflicts of different local schemas. This concept facilitates to achieve semantic integration.

This paper describes a co-operation project between National University of Singapore and Vienna University of Technology. The research intends to integrate and manage a set of integrated OLAP warehouses for knowledge management and decision support on the Web.

7 References

[1] L. Alschuler. Going to Extremes. *The XML Cover Page* (XML) Topic Maps, by Robin Cover. http://xml.coverpages.org/topicMaps.html

[2] A. Ammoura, O. Zaiane, and R. Goebel. Towards a Novel OLAP Interface for Distributed Data Warehouses. *Proc. of DAWAK 2001*, Springer LNCS 2114, pp. 174-185, Munich, Germany, Sept. 2001.

[3] C. Bettini, S. Jajodia, and S.X. Wang. *Time Granularities in Databases, Data Mining, and Temporal Reasoning*. Springer Verlag, Berlin, 2000.

[4] M. Biezunski, and S.R. Newcomb. XML Topic Maps: Finding Aids for the Web. *IEEE MultiMedia*, Volume 8(2), April-June, pp. 104-108, 2001.

[5] E.F. Codd, S.B. Codd, and C.T. Salley. Providing OLAP (On-Line Analytical Processing) to User-Analyst: An IT mandate. Technical Report, E.F. Codd & Associated, 1993.

[6] The DARPA Agent Markup Language Homepage. http://daml.semanticweb.org/

[7] D. Fensel, I. Horrocks, F. Van Harmelen, S. Decker, M. Erdmann, and M. Klein. OIL in a Nutshell. In: Knowledge Acquisition, Modeling, and Management, *Proc. of the 12th European Knowledge Acquisition Conference (EKAW-2000)*, R. Dieng et al. (eds.), Springer-Verlag LNAI 1937, pp. 1-16, Oct. 2000.

[8] F. Hakimpour, and A. Geppert. Resolving Semantic Heterogeneity in Schema Integration: an Ontology Based Approach. *Proc. of the Intl. Conf. On Formal Ontologies in Information Systems (FOIS-2001)*, ACM Press, Ogunquit, Maine, Oct. 2001.

[9] I. Horrocks, D. Fensel, J. Broekstra, S. Decker, M. Erdmann, C. Goble, F. van Harmelen, M. Klein, S. Staab, R. Studer, and E. Motta. *The Ontology Inference Layer OIL*.

[10] W. Hümmer, J. Albrecht, and H. Günzel. Distributed Data Warehousing Based on CORBA. Proc. of *IASTED International Conference on Applied Informatics (AI'2000)*, Innsbruck, Austria, February 2000.

[11] ISO/IEC 13250: Information technology – SGML Applications - Topic Maps. International Organization for Standardization, Genf, Schweiz, December 1999.

[12] R. Kimball. *The Data Warehouse Toolkit*, John Wiley & Sons, 1996.

[13] L. Kerschberg. Knowledge Management in Heterogeneous Data Warehouse Environments. *Proc. of DAWAK 2001*, Springer LNCS 2114, pp. 1-10, Munich, Germany, Sept. 2001.

[14] M. Lacher, and S. Decker. On the Integration of Topic Maps and RDF Data. Semantic Web Workshop at Stanford, August 2001.

[15] O. Mangisengi, J. Huber, Ch. Hawel, and W. Essmayr. A Framework for Supporting Interoperability of Data Warehouse Islands Using XML. *Proc. of DAWAK 2001*, Springer LNCS 2114, pp. 328-338, Munich, Germany, Sept. 2001.

[16] J. Mercy, and H. Sonnberger. Funding Research in Data Warehousing and Knowledge Discovery EPROS: The European Plan for Research in Official Statistics. *Proc. of DAWAK 2000*, Springer LNCS 1874, pp. 134-145, London, GB, Sept. 2000.

[17] G. Moore. RDF and TopicMaps: An Exercise in Convergence. *XML Europe 2001*, Berlin, 2001.

[18] S. Pepper, and G. Moore. XML Topic Maps (XTM) 1.0 TopicMaps.Org Specification. http://www.topicmaps.org/xtm/1.0/

[19] The Semantic Web Homepage. http://www.semanticweb.org

[20] R. Widhalm, and T.A. Mück. *Topic Maps*. To appear: Xpert.press (Springer Verlag), Berlin, 2001.