CHAPTER 1
INTRODUCTION

Many organizations rush to advertise their business and operations on WWW. Each organization designs and develops their information independently, catering to the needs of their own departments. As a consequence, various types of information sources within an organization, ranging from structured, semi-structured to unstructured information sources, are set up on different physical hosts, operating systems, or DBMSs. These different information sources are called the heterogeneous information sources (hereafter HIS). The structured information sources are usually constructed from predefined schemas; the semi-structured information sources, in most cases, are constructed from similar configurations, but in weaker form of predefined schemas; yet the unstructured information sources are constructed without any restrictions. Examples of such information sources are structured database files, semi-structured XML (Extensible Markup Language) documents (W3C, 2000), and unstructured plain text, respectively.

Due to the increase of demands in accessing and integrating the information to support global applications and decision making requirements, organizations have to solve the heterogeneity problems of the HIS. These problems precipitate from the difficulties of accessing different systems, where the obtained information can be inconsistent and contradictory. The lack of awareness of the structure of data, the location of the required data, query language, and the use of different terms belonging to the same domain in each proprietary system can also lead to cross-system data sharing and exchanging problems, i.e., the heterogeneity problems. These heterogeneity problems can be classified into different levels as follows:

(1) *The system heterogeneity* which includes hardware and software heterogeneity;

(2) *The query language heterogeneity* in which different languages are used to manipulate data represented by different data models, e.g., SQL is used for manipulating the relational data model, whereas OQL is used for manipulating the object-oriented database model;

(3) *The structural heterogeneity* which occurs in four possible ways, namely, *type conflicts*, *dependency conflicts*, *key conflicts*, and *behavioral conflicts* (Batini, Lenzirini, and Navathe, 1986; Özsu, and Valduriez, 1999).

- *Type conflicts* caused by the abstraction level in data modeling, e.g., an entity of one database is designed as an attribute in another database,
- *Dependency conflicts* resulting from different relationship modes (e.g., one-to-one versus many-to-many) are used to represent the same thing in different schemas,
- *Key conflicts* caused by different candidate keys are available and different primary keys are selected in different schemas, and
- *Behavioral conflicts* implied by the modeling mechanism, for example, deleting the last item from one database may cause the deletion of the containing entity, i.e., deletion of the last employee causes the dissolution of the department;

(4) *The semantic heterogeneity* which occurs when there is a disagreement on the meaning, interpretation, or intended use of the same or related data (Sheth, and Larson, 1990). The semantic heterogeneity can be classified into four types, namely, *naming conflicts*, *data type conflicts*, *scaling conflicts*, and *generalization conflicts*.

- *Naming conflicts*, encompassing two different kinds of conflict, namely, synonym and homonym conflicts. Synonym conflicts are concerned with semantically equivalent concepts (i.e., entities) or properties (i.e., attributes) defined by different names. For example, concept `Staff` in data source 1 and concept `Instructor` in data source 2 are synonyms since they both refer to the same fact. Homonym conflicts, on the other hand, are concerned with semantically unrelated concepts or properties defined by the same name. For example, attribute `Name` in data source 1 refers to customer name, whereas attribute `Name` in data source 2 refers to company name;
- *Data Type conflicts*, concerning semantically equivalent properties that are defined by different data types. For example, the attribute `staff_id` in data

source 1 that is synonymous with attribute `instructor_id` in data source 2 is defined as string, whereas the attribute `instructor_id` is defined as integer;

- *Scaling conflicts*, concerning semantically equivalent properties that are defined by different scales (or units of measure). For example, the unit type of attribute `salary` in data source 1 is $US, whereas the `salary` in data source 2 is $AUS; and
- *Generalization conflicts*, concerning semantically related concepts that are defined in different systems where the concepts in one system subsume the concepts in another system. For example, concept `Staff` in data source 1 subsumes concept `Instructor` in data source 2 since concept `Instructor` is a subconcept of `Staff`.

This research proposed the Semantic Information Gathering Approach (SIGA) system for accessing and integrating the HIS on the WWW. A reference architecture of SIGA is designed based on layered-architecture, along with a metadata dictionary which is a core component of the reference architecture for resolving semantic heterogeneity. The metadata dictionary is designed to be a repository for storing conceptual level and physical level data descriptions. These data descriptions are subsequently used by agents to access and retrieve real data from the underlying physical sources. As a consequence, the application of metadata dictionary also provides greater benefits to solving data replication in distributed environment.

According to Gruber (Gruber, 1993), the ontology is introduced as an "explicit specification of a conceptualization." This notion provides a common understanding of a domain for communicating between people and application systems (Fensel, 2001). As such, the underlying ontology principles are extensively employed in the design of the metadata dictionary. This research, however, merely focuses on modeling and designing the domain ontology, which is the fundamental building block of the metadata dictionary on the basis of a bottom-up design approach (Castano, Antonellis and Vimercati, 2001; Özsu, and Valduriez, 1999; Vet and Mars, 1998). The domain ontology components are defined in terms of object-oriented framework and set theory.

In order to render maximal interoperability of the metadata dictionary for practical purpose across heterogeneous systems, XML is chosen as a language for expressing the metadata dictionary contents. This research describes the design and construction of XML-DTD from domain ontology components in a web-based environment. The metadata dictionary contents are systematically implemented by means of flexible XML data model. As a consequence, the querying process has been effectively employed based on this metadata dictionary.

1.1 The Objectives

This research aims to overcome semantic heterogeneity problem so as to achieve semantic interoperability and cooperation when accessing and integrating data from heterogeneous information sources (HIS). The following objectives are proposed:

(1) To develop a reference architecture for accessing and integrating the HIS in a web-based environment,

(2) To develop a metadata dictionary which is an essential component used to solve semantic heterogeneity, and

(3) To develop a global query model for users in accessing the HIS through the metadata dictionary.

1.2 Procedure and Outline

In order to achieve the aforementioned objectives, the following tasks will be addressed: the related technologies are investigated to be incorporated with the proposed approach, namely, XML technology, ontology-based approach, agent technology, information integration techniques, and architectures of the heterogeneous information sources as described in Chapter 2. The reference architecture is proposed for integrating the heterogeneous information sources in a web-based environment, including the agent architecture as presented in Chapter 3. A methodology for modeling the metadata dictionary is devised to represent metadata dictionary components. The metadata dictionary management is also provided in Chapter 4 as a means to manage the metadata dictionary contents when adding or dropping the physical information sources. The technique for representing the metadata dictionary components with the help of object-oriented paradigm

and set theory is illustrated in Chapter 5. A method for structuring the metadata dictionary components to the XML-based metadata dictionary which is a web application architecture that supports system-wide interoperability is presented in Chapter 6. This research also demonstrates through a practical case study, how to model and apply the proposed metadata dictionary to a practical use, hence eliminate the semantic heterogeneity as described in Chapter 7. The querying process in accessing and integrating the heterogeneous information sources in accordance with the metadata dictionary structure are also illustrated in Chapter 8. The query validation of the returning results is also provided in order to ensure that the relevant data is returned to the user. To compare the advantages and disadvantages of the proposed approach with the other approaches, the current related works are explored to classify the advantages and drawbacks of each approach as described in Chapter 9, and the conclusions and recommendations of this work will be given in Chapter 10.

1.3 Benefits of the Dissertation

This work contributes to both theory and practice of the HIS in many aspects as follows:

(1) A reference architecture of the HIS and agent model,

(2) A domain ontology model based on object-oriented and set theory which is an abstract representation of the proposed metadata dictionary structure,

(3) A metadata dictionary which provides a mapping mechanism to associate user's requests posed at the conceptual level with the physical level, allowing direct access to stored information without loss of general query formulation,

(4) A formal definition of domain ontology which gives rise to the construction of metadata dictionary for use in real world implementation,

(5) An implementation paradigm that demonstrates flexibility and interoperability across heterogeneous system, and

(6) A systematic approach for global transaction decomposition to matching physical source sub-transactions.

Additional benefits precipitated from this work encompass the transformation from theoretical foundation to actual implementation by means of XML technology. This is a significant connection that bridges the gap between theory and practice. Information

querying and retrieving from heterogeneous information sources can be realistically accomplished through XML document that is constructed from ontology-based metadata dictionary. As a consequence, the semantic heterogeneity is eliminated.