# Chapter 5. Research Discussions and Limitations

It is well known that good benchmarks drive industry and technology forward. In this chapter, the managerial and technical implications of the research result are discussed. According to the research model and prototype implementation, we summarize the research implications and limitations in the following sections.

## 5.1. Research Implications

The major discussions of this research are described as follows.

- **Not a separately designed benchmark but a combined XML and ontology benchmark in information integration systems**

As mentioned above, heterogeneous information integration on the Internet/Intranet is one of the hottest IT topics today. Both academia and industry have proposed several solutions in EB field. Recently, more and more solutions use ontology to provide meaning for XML standards, and XML to provide a valuable medium for information exchange between applications that share the same ontology. There are already separate benchmarks on XML, relational database, object-oriented database, and Web server. However, they are all independent and individual benchmarks. Heterogeneous information integration should incorporate XML and ontology methods. There is no combined XML and ontology benchmark developed for this topic. In this research, we propose an XML and ontology benchmark workload model for heterogeneous information integration under EB environment. System developers can try different query processing implementations and evaluate these alternatives with the benchmark. Thus, this benchmark can be a simple and effective method to help system developers improve system performance. The

benchmark results can also help enterprises to make purchasing decisions.

- **A generic-construct-based design approach in the formulation of workload components**

The workload model we propose in this research is developed in generic constructs. It is an extension of current XML benchmarks in terms of test database and test query and control component. Generic constructs and open standards are used to define the new workload model. The data model considers relational model, object model, and Web pages. The query model considers the typical operations in XML and ontological applications. It may evaluate pure XML query processing, or combine ontology-reasoning services. Furthermore, the data model, query model and control model can be user-driven. Users can specify each of them according to their needs and requirements. Designed as such, this method can conform to the desirable characteristics of a good benchmark design including scalability, portability, simplicity, and producibility.

It is easier for users to apply this generic workload model across different scenarios. As Table 5.1 shows, the workload model is synthetic and described in generic terms. When applying to other synthetic or empirical scenarios, one needs to map the generic terms with specific data items. However in the empirical cases, one may have a variety of query types that do not correspond to the query model we define. In such situations, we can find the major operations in a certain scenario and use mapped and synthetic queries to represent them.

Table 5.1: This Workload Model Category

| | | Data | |
|---|---|---|---|
| | | Synthetic | Empirical |
| **Query** | Synthetic | ★ | |
| | Empirical | | |

- **A systematic workload design approach**

In this research, the benchmark is a workload-based development, i.e. analysis and approach based, instead of ad hoc. In query model, we identify key factors that determine the degree of complexity of query. In data model, we define main variable that determine the level of data size and problem scope and scale. This makes the workload model systematic and methodological. We aim to provide the ability that users can diagnose and detect the strengths and weaknesses of information integration systems.

- **A pioneering workload model to test ontology used in information integration**

In present, there is no strong and standard way to benchmark ontology, especially in the heterogeneous information integration area. It is important to understand and evaluate whether an ontology can represent the meaning of real world knowledge and assist in the XML query processing.

## 5.2. Research Limitations

Due to time and resource constraints, this research may improve in several areas in the future:

- In this research, the ontology workload model is derived from a description logic benchmark. However, ontology is more expressive than description logic. New challenges may arise with wider domains to be tested. It needs further study to enhance the workload model.
- The structure of global schema and the data sources may evolve and change. Test queries must be different. This prototype at present does not support the ability to dynamically define the data structure. A wider query model is needed.
- In the prototype, the data generator at present is primitive. It assumes the distribution of data sources to be uniform. Data value distributions are assumed to be uniform as well.
- The recall and precision measurements are limited. For wider and more dynamic data sources, it may be difficult to generate the relevant results. It is hard because each domain has its own domain ontology. Expert dependency is unavoidable.