行政院國家科學委員會專題研究計畫 成果報告

智慧型資訊整合於異質資料倉儲和資料探勘之模型、架構、績效評估-應用本體論、母型架構、和學名結構(第3年)
研究成果報告(完整版)

計畫類別：個別型
計畫編號：NSC 96-2416-H-004-017-MY3
執行期間：98年08月01日至99年12月31日
執行單位：國立政治大學會計研究所

計畫主持人：譚家蘭

計畫參與人員：學士級-專任助理人員：張芷霖
ADING-兼任助理人員：李子佑
碩士班研究生-兼任助理人員：曾國哲
碩士班研究生-兼任助理人員：賈雅雯
博士班研究生-兼任助理人員：賴彥昌
博士班研究生-兼任助理人員：杜逸寧
博士班研究生-兼任助理人員：詠凱婷

報告附件：國外研究心得報告
出席國際會議研究心得報告及發表論文

處理方式：本計畫可公開查詢

中華民國100年03月31日
Abstract ......................................................................................................................... V
1. Introduction ................................................................................................................ 6
   1.1 Research Motivation ........................................................................................ 6
   1.2 Research Problem ............................................................................................ 6
   1.3 Research Approach .......................................................................................... 7
2. Literature Review ....................................................................................................... 9
   2.1 XML Query Capability .................................................................................... 9
   2.2 XML Benchmarks .......................................................................................... 10
      2.2.1 XMark ................................................................................................. 10
      2.2.2 XMach ................................................................................................. 10
      2.2.3 XOO7.................................................................................................. 10
   2.3 XML Benchmarks Comparison ..................................................................... 11
   2.4 Ontology ........................................................................................................ 13
      2.4.1 Ontology and Intelligent Information Integration ............................... 14
      2.4.2 Ontology and Benchmark ................................................................... 14
3. Research Method ..................................................................................................... 15
   3.1 Research Approach ........................................................................................ 15
      3.1.1 XML Data Model ................................................................................ 17
      3.1.2 XML Query Model ............................................................................. 19
      3.1.3 Exact Match ........................................................................................ 20
      3.1.4 Joins .................................................................................................... 21
      3.1.5 Regular Path Expressions ................................................................... 21
      3.1.6 Document Construction ...................................................................... 22
      3.1.7 Ordered Access ................................................................................... 23
      3.1.8 Sorting ................................................................................................. 23
      3.1.9 Missing Elements ................................................................................ 24
      3.1.10 Text Search ........................................................................................ 24
      3.1.11 Data-type Cast ................................................................................... 25
      3.1.12 Function Application ......................................................................... 25
   3.2 Ontology Data Model .................................................................................... 26
   3.3 Ontology Query Model .................................................................................. 27
   3.4 Test Database Generation ............................................................................... 29
   3.5 Experimental Design ...................................................................................... 32
圖目錄

Figure 1: Research Model ................................................................. 16
Figure 2: Graphic Representation of XML Data Model ..................... 18
Figure 3: XML Query Model ............................................................. 20
Figure 4: Ontology Query Model ..................................................... 28
Figure 5: Data Loader ................................................................. 33
Figure 6: Query Generator ............................................................. 34
Figure 7: Query Generator - Input ................................................ 34
Figure 8: Query Generator – Select Standard Query Type ............... 35
Figure 9: Query Generator – Set Standard Query Complexity .......... 35
Figure 10: Result Collector ....................................................... 36
Table 1: Comparison of Benchmarks over Workload Characteristics ......................... 12
Table 2: Comparison of Benchmarks over Query Functionalities.............................. 13
Table 3: Complexity Factors Listing............................................................................ 25
Abstract

Benchmarks are vital tools in the performance measurement and evaluation of computer hardware and software systems. Standard benchmarks such as the TREC, TPC, SPEC, SAP, Oracle, Microsoft, IBM, Wisconsin, AS^3AP, OO1, OO7, XOO7 benchmarks have been used to assess the system performance. These benchmarks are domain-specific in that they model typical applications and tie to a problem domain. Test results from these benchmarks are estimates of possible system performance for certain pre-determined problem types. When the user domain differs from the standard problem domain or when the application workload is divergent from the standard workload, they do not provide an accurate way to measure the system performance of the user problem domain. System performance of the actual problem domain in terms of data and transactions may vary significantly from the standard benchmarks. In this research, we address the issue of domain boundness and workload boundness which results in the ir-representative and ir-reproducible performance reading. We tackle the issue by proposing a domain-independent and workload-independent benchmark method which is developed from the perspective of the user requirements. We present a user-driven workload model to develop a benchmark in a process of workload requirements representation, transformation, and generation. We aim to create a more generalized and precise evaluation method which derives test suites from the actual user domain and application. The benchmark method comprises three main components. They are a high-level workload specification scheme, a translator of the scheme, and a set of generators to generate the test database and the test suite. The specification scheme is used to formalize the workload requirements. The translator is used to transform the specification. The generator is used to produce the test database and the test workload. In web search, the generic constructs are main common carriers we adopt to capture and compose the workload requirements. We determine the requirements via the analysis of literature study. In this study, we have conducted ten baseline experiments to validate the feasibility and validity of the benchmark method. An experimental prototype is built to execute these experiments. Experimental results demonstrate that the method is capable of modeling the standard benchmarks as well as more general benchmark requirements.

Keywords: XML, Ontology, Intelligent Information Integration, Generic Construct, Benchmark, Workload Model, Performance Measurement and Evaluation
1. Introduction

1.1 Research Motivation

A benchmark is a standard by which something can be measured or judged. A computer system benchmark is a set of executable instructions to be enforced in controlled experiments to compare two or more computer hardware and software systems. Hence, benchmarking is the process of evaluating different hardware systems or reviewing different software systems on the same or different hardware platforms. A web search service benchmark is therefore a standard set of executable instructions which are used to measure and compare the relative and quantitative performance of two or more systems through the execution of controlled experiments. Benchmark data such as throughput, jobs per time unit, response time, time per job unit, price and performance ratio, and other measures serve to predict price and performance and help us to procure systems, plan capacity, uncover bottlenecks, and govern information resources for various user, developer, and management groups (Can et al., 2004) (David et al., 2001) (Anon et al., 1985).

Examples are the TREC, TPC, SPEC, SAP, Oracle, Microsoft, IBM, Wisconsin, AS3AP, OO1, OO7, XOO7 standard benchmarks that have been used to assess the system performance. These benchmarks are domain-specific in that they model typical applications and tie to a problem domain. Test results from these benchmarks are estimates of possible system performance for certain pre-determined problem types. When the user domain differs from the standard problem domain or when the application workload is divergent from the standard workload, they do not provide an accurate way to measure the system performance of the user problem domain. System performance of the actual problem domain in terms of data and transactions may vary significantly from the standard benchmarks. Performance measurement and evaluation is crucial in the development and advance of web search technology. A more open and generic benchmark method is needed to provide a more representative and reproducible workload model and performance profile (Jansen et al., 2006) (Richard 2006) (Vaughan 2004) (Kraaij et al., 2002).

1.2 Research Problem

Domain boundness and workload boundness are the research problem we try to tackle in this research. As described above, standard benchmarks model certain application types in a pre-determined problem domain. They represent a fixed problem set presented to the proposed system. When the user domain differs from the standard domain or when the user workload deviates from the standard
workload, the test results vary significantly in the real setting and under the actual application context. Users cannot reproduce the test results and predict the performance. The reason is because benchmark results are highly dependent upon the real workload and the actual application. The standard test workload cannot represent the real workload and the test suite cannot accommodate the application requirement. Standard benchmarks cannot measure the effects of the user problem on the target system nor generate the realistic and meaningful test results (Stephen 2002).

In this research, we address the issue by proposing a domain-independent and workload-independent benchmark method which is developed from the perspective of the user requirements. We propose to develop a more generalized and more precise performance evaluation method from the perspective of the common carriers of workload requirements. We create a user-driven approach which models the benchmark development in a process of workload requirements representation, transformation, and generation.

1.3 Research Approach

Benchmarks can be synthetic or empirical. Synthetic benchmarks model the typical applications in a problem domain and create the synthetic workload. Empirical benchmarks utilize the real data and tests. Though real workloads are ideal tests, the costs of re-implementation of the actual systems usually outweigh the benefits obtained. Synthetic benchmarks are therefore the common approach chosen by developers and managers. Further, benchmark experiments are composed of the experimental factors and the performance metrics. Experimental factors represent the variables which can affect the performance of the systems. Performance metrics are the quantitative measurements to be collected and observed in the benchmark experiments. They represent the set of independent variables and dependent variables to be modeled and formulated in the benchmark.

A workload is the amount of work assigned to or performed by a worker or unit of workers in a given time period. The workload is the amount of work assigned to or performed by a system in a given period of time. The loads are best described by the amount of work, the rate at which the work is created, and the characteristics, distribution, and content of the work. Conventionally, workload modeling and characterization start with the domain survey, observation, and data collection, and continue with a study of the main components and their characteristics. In general, the workload components consist of the data, operations, and control.

In specific, workload analysis
involves the data analysis and the operation analysis. We analyze the size of the data, the number of records, the length of records, the types of attributes, the value distributions and correlations, the keys and indexing, the hit ratios, the selectivity factors. We investigate the complexity of operations, the correlation of operation, the data input into the operation, the attributes and objects used by the operation, the result size, and the output mode. These are further examined with the control analysis of the duration of test, the number of user, the order of test, the number of repetition, the frequency and distribution of test, and the performance metrics.

In the web search context, we develop a benchmark method that comprises a workload requirements specification scheme, a scheme translator, and a set of benchmark generators. We adopt the common carrier of generic constructs. We analyze the key web search algorithms and formulate the generic constructs. The generic constructs describe the page structure and the query structure of web search that is not tied to a per-determined search engine.

**Workload Specification Scheme**

The workload specification scheme is designed to model the application requirements. It is a high-level generic construct concept to describe requirements concerning data, operation, and control. A generic construct is the basic unit of operand. An operation is the basic unit of operator. The collection of a generic construct and an operation formulate a workload unit. Each workload unit becomes a building block to compose a larger workload unit.

**Scheme Translator**

The scheme translator is created with a set of lexical rules and a set of syntactical rules to translate the workload specification. It performs the code generation and produces three output specifications. One is the data specification. The other is the operation specification. Another is the control specification.

**Data Generator**

The data generator is made up of a set of data generation procedures which are used to create the test database according to the data distribution specification.

**Operation Generator**

The operation generator is made up of a set of operation generation procedures to generate the search operations. These procedures select operations, determine operation precedence, schedule arrivals, prepare input data, issue tests, handle queues,
gather and report time statistics.

**Control Generator**

The control generator is made up of a set of control generation procedures to generate the control scripts which are used to drive and supervise the experiment execution.

2. **Literature Review**

2.1 **XML Query Capability**

Benchmarking the XML data management systems should consider many factors. Designing a set of comprehensive queries to test the XML databases’ performance is an important point. XML query languages should capture the whole characteristics of a XML document, and the functionalities they provide would influence the query performance. The W3C XML Query Language working group (Chamberlin, Fankhauser, Marchiori, & Robie, 2003) list 20 XML query language “must have” functionalities, as Table shows. Some of the expected functionalities may affect the efficiency of the system significantly.

XQuery has met all of the requirements except F12 and F16, and it becomes a standard query language to test the performance of XML data management systems. Generally speaking, queries to benchmark XML databases would fall into several categories: Match, Join, Navigation, Casting, Reconstruction, and Update. Queries for Match are mainly used to test the database ability to handle simple string lookups with a fully specified path. Join queries can be divided into two parts: Join on References, and Join on Values. References are an important part of XML, because they allow richer relationships than just hierarchical structure. Queries Join on References would test if query optimizer can take advantage of references to be joined. Queries Join on Values, on the other hand, would test the database’s ability to handle large intermediate results. Differing from the former, their joins are on the basis of values. Navigation Queries investigate how well the query processor can optimize path expressions, and avoid traversing irrelevant parts of the tree. Strings are the basic data type in XML documents. Casting strings to another data type that carries more semantics is necessary. Queries for Casting challenge the ability of the database to cast different data types. Reconstruction Queries attempt to reconstruct the original document from its fragmentations stored in the databases. Update Queries try to add, delete, and modify elements in the XML document. These queries test the databases’ ability to manage XML document. Furthermore, other XML query functionalities such as sort, ordered access, text search, and
aggregation also should be captured in the benchmark query set.

2.2 XML Benchmarks

XMark, XMach-1 and XOO7 are three benchmarks available today that can be used to evaluate certain aspects of XML database systems.

2.2.1 XMark

XMark (Schmidt, Waas, Kersten, Carey, Manolescu, & Busse, 2002) is a single-user benchmark. The data model of XMark is an Internet auction site. Therefore, its database contains one big XML document with text and non-text data. XMark enriches the references in the data, like the item IDREF in an auction element and the item’s ID in an item element. The text data used are the 17000 most frequently occurring words of Shakespeare’s plays. The standard data size is 100MB with a scaling factor 1.0 and users can change the data size by 10 times from the standard data (the initial data) each time. However, it has no support for XML Schema. In operation model of XMark, 20 XQuery challenges are designed to cover the essentials of XML query processing, as Table shows. No update operations are specified in XMark.

2.2.2 XMach-1

XMach-1 (Böhme & Rahm, 2001) is a scalable multi-user benchmark. The main objective of the benchmark is to stress-test XML systems under a multi-user workload. The data model of XMach-1 is designed for B2B applications and considers text documents and catalog data. It assumes that size of the data files exchanged will be small. It provides support for DTD only and does not consider XML Schema for optimization. The operation model of XMach-1 consists of eight queries and three update operations, shown in Table.

Queries specified in XMach-1 cover typical database functionality (join, aggregation, sort) as well as information retrieval and XML-specific features (document assembly, navigation, element access). Update operations cover inserting and deleting of documents as well as changing attribute values. We find that some queries contain several query functionalities. It is hard to analyze the experiment result and ascertain which feature leads to the given performance result. Specially, XMach-1 has defined three update operations that are unique across other XML benchmarks.

2.2.3 XOO7

XOO7 (Li, Bressan, Dobbie, Lacroix, Lee, Nambiar, & Wadhwa, 2001) is an XML version of the OO7 benchmark, which was designed to test
the efficiency of object-oriented DBMS. XOO7 is a single-user based benchmark for XMLMS that focuses on the query processing aspect of XML. The data model of XOO7 comes from the OO7 benchmark by mapping the OO7 schema and data set to XML. No specific application domain is modeled by the data of XOO7. It is based on a generic description of complex objects using component-of relationships. XOO7 also proposes three different databases of varying size: small, medium, and large. It supports DTD only. In operation model, XOO7 provides relational, document and navigational queries that are specific and critical for XML database applications. These queries test the primitive features and each query covers only a few features. Table displays the queries adopted in XOO7. XOO7 contains large amount of queries, each query covers only a few features. Comparing to the other two benchmarks, XOO7 has certainly the highest ratio which stresses its data-centric focus. However, we can find that some queries are focus on the same functionality. Similar to XMark, no update operation is specified in XOO7.

2.3 XML Benchmarks Comparison

A comparison of key features of these main XML benchmarks against this research is described in Table 1. The key features include application focus, evaluation scope, database and workload characteristics.
Table 1: Comparison of Benchmarks over Workload Characteristics

<table>
<thead>
<tr>
<th>Feature</th>
<th>XMark</th>
<th>XMach-1</th>
<th>XOO7</th>
<th>This Research</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evaluation Scope</td>
<td>Query Processor</td>
<td>DBMS</td>
<td>Query Processor</td>
<td>Heterogeneous Information Integration</td>
</tr>
<tr>
<td>Application Domain</td>
<td>E-Commerce</td>
<td>E-Commerce</td>
<td>Generic</td>
<td>Generic</td>
</tr>
<tr>
<td><strong>Data Model</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Documents</td>
<td>Single-document</td>
<td>Multi-documents</td>
<td>Multi-documents</td>
<td>Multi-documents</td>
</tr>
<tr>
<td>Scalability of Document Number</td>
<td>1</td>
<td>$10^4$~$10^7$</td>
<td>Unlimited</td>
<td>Various</td>
</tr>
<tr>
<td>Scalability of Document Size</td>
<td>10MB~10GB</td>
<td>16KB</td>
<td>Unknown</td>
<td>Various</td>
</tr>
<tr>
<td>Data Heterogeneity</td>
<td>XML document only</td>
<td>XML document only</td>
<td>XML document only</td>
<td>Heterogeneous data sources</td>
</tr>
<tr>
<td>Nodes/KB</td>
<td>18</td>
<td>10</td>
<td>67</td>
<td>Various</td>
</tr>
</tbody>
</table>

**Operation Model**

<table>
<thead>
<tr>
<th>Queries</th>
<th>20</th>
<th>8</th>
<th>23</th>
<th>14</th>
</tr>
</thead>
<tbody>
<tr>
<td>Update Operation</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 2 groups queries of each benchmark by query functionality. Compared to other XML benchmarks, XMark provides a concise and comprehensive set of queries. However, it does not provide update operations to manipulate XML documents. XMach-1 only defines a small number of XML queries that cover multiple functions and update operations for which system performance is determined. XOO7 maps the original queries of OO7 into XML, and adds some XML specific queries. In general, XMach-1, XMark, and XOO7 cover only a subset of the XML query requirements. In this research, we attempt to propose a generic workload model. In order to cover the whole functionalities of XML query processing, we combine queries of these three XML benchmarks and integrate them into ten types of queries. In particular, the intelligent information integration system is generally used for query data, not provide data manipulation functions. Therefore, the query model in this research does not support update operations.
### Table 2: Comparison of Benchmarks over Query Functionalities

<table>
<thead>
<tr>
<th>Query Functionality</th>
<th>XMak</th>
<th>XMach-1</th>
<th>XOO7</th>
<th>This Research</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exact Match</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Joins</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Join on Reference</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Join on Value</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Regular Path Expressions</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Full Sub-path</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Unknown Sub-path</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Document Construction</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Structure Preserving</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Structure Transforming</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Ordered Access</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Sorting</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>By String</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>By Non-string</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Missing Elements</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Text Search</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Data-type Cast</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Function Application</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Update Operation</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

### 2.4 Ontology

Ontologies play an important role for integration as a way of formally defined terms for communication. They aim at capturing domain knowledge in a generic way and provide a commonly agreed understanding of a domain, which may be reused, shared, and operationalized across applications and groups.

A good ontology should represent the domain specific knowledge explicitly. The question is how do we know an ontology is good? The answer is the ontology benchmark. There are plenty of benchmark studies in other fields like database or compilers. However, there are no specific benchmarks studies or tools for evaluating ontology-based applications. In fact, there is still no guideline to evaluate ontologies and related technologies.

In this section, we introduce the role of ontologies in intelligent information integration first. And then we discuss a major inference task which
is the main operation of an ontology benchmark. Finally, the ontology related benchmark works are reviewed.

2.4.1 Ontology and Intelligent Information Integration

Traditional integration approaches use inexpressive models of database schemas or XML trees to integrate heterogeneous data sources. This would cause many semantic heterogeneity problems. Ontologies provide much richer modeling means with classes and properties organized into is-a hierarchy and enriched with axioms and relations processable with inference (Maier, Aguado, Bernaras, Laresgoiti, Pedinaci, Pena, & Smithers, 2003). Almost all ontology-based integration approaches ontologies are used for the explicit description of the information source semantics. With respect to the integration of data sources, they can be used for the identification and association of semantically corresponding information concepts. Some approaches use ontologies not only for content explication, but also either as a global query model or for the verification of the (user-defined or system-generated) integration description (Wache, Vögele, Visser, Stuckenschmidt, Schuster, Neumann, & Hübner, 2001). Ontologies are usually expressed in a logic-based language, so that fine, accurate, consistent, sound, and meaningful distinctions can be made among the classes, properties, and relations. Therefore, ontologies not only have the expressiveness needed in order to model the data in the sources, but their reasoning ability can help in the selection of the sources that are relevant for a query of interest, as well as to specify the extraction process.

2.4.2 Ontology and Benchmark

To the best of our knowledge, the benchmark presented here is the first one for ontology-based intelligent information integration. The ontology benchmark model in this research differs from database benchmarks, such as Wisconsin benchmark, OO7 benchmark, and BUCKY benchmark. They are all DBMS-oriented and storage benchmarks, and there is no inference ability included. In this research, the ontology workload model is applied to an intelligent information integration system, and we focus on the inference ability of the ontology. Ontology and XML are often found together and are often confused. XML is a standard for marking up - adding additional information, called metadata - to documents. The purpose of XML is to tag textual information with additional structure that enables it to be “understood” and exchanged by programs. However, XML tags still require humans to interpret their meanings. Therefore, XML benchmarks only focus on structural and syntactic evaluation of systems, and they have no
semantics. On the other hand, ontology benchmark is devoted to capture the semantic expressions in the system. Thus, ontology and XML are complementary technologies: ontology provides the meaning for XML standards; XML provides a valuable medium for information exchange between programs that share the same ontology.

As mentioned above, there is still no guideline for evaluation of ontology-based application. Horrocks and Patel-Schneider (1998) benchmark description logic systems, or so-called knowledge bases. Description logics (DLs) are a family of knowledge representation languages that can be used to represent the knowledge of an application domain in a structured and formally well-understood way. Description logic systems provide their users with various inference capabilities that deduce implicit knowledge from the explicitly represented knowledge. Horrocks and Patel-Schneider try to evaluate the reasoning algorithms in description logics. Terminological part (Tbox) is a set of axioms describing the structure of domain. Assertional part (Abox) is a set of axioms describing concrete situation (Horrocks, 2002). They are related to this research. In an intelligent information integration system, the ontology can be viewed as the Tbox, and the heterogeneous data can be viewed as Abox. However, the logic described is only a subset of the ontology languages, such as DAML+OIL and OWL. DAML+OIL and OWL can be seen to be equivalent to a very expressive description logic. They provide more constructors and allow more axioms than description logic. Therefore, the inference services of ontology are more complex than traditional description logic systems.

3. Research Method

3.1 Research Approach

The literatures studied would help us to identify the important performance factors for XML and ontology processing. We analyze the XML-specific and ontology-specific requirements in more details to justify the design of the benchmark. The research structure is shown in Figure 1. The benchmark study consists of two benchmark workload models, the XML benchmark workload model and the ontology benchmark workload model. Both of them consist of the data model and query model according to the generic constructs and constraints requirements. Next, the control model is created before the generic workload model to be generated and executed so as to measure and evaluate the systems.
Figure 1: Research Model
In this research, we focus on intelligent information sources integrated in XML and ontology. The benchmark model we propose would capture most features of the released XML-based and ontology-based specifications.

Developing a benchmark requires the definition of the test workload model first. In this research, we provide a benchmark workload model that combines XML and ontology in intelligent information integration. In XML workload model, the data model describes a generic XML data model and the operation model defines a comprehensive set of test queries that covers the major aspects of XML query processing. In ontology workload model, the data model describes the major ontology component, and the operation model defines some important criteria to query the ontology. The control model defines the variables that used to set up the benchmark environment.

3.1.1 XML Data Model

XML is a hierarchical data format for information exchange on the Web. An XML document consists of nested elements that contain data or other elements. The boundaries of these elements are either delimited by start-tags and end-tags, or, for empty elements, by empty-element tags. The text between start-tags and end-tags is the content of the element. Each element has a type, identified by name, sometimes called its “generic identifier” (GI), and may have a set of attribute specifications. Each attribute specification has a name and a value (Bray, Paoli, Sperberg-McQueen, & Maler, 2000). XML documents may comply with a Document Type Definition (DTD) or a XML Schema. DTD has traditionally been the most common method for describing the structure of XML document. But DTD lacks enough expressive power to properly describe highly structured data. XML Schemas are an XML language for describing and constraining the content of XML documents. It provides a richer and more powerful means for defining the data. Therefore, XML schema becomes the most common method for defining and validating highly structured XML documents rapidly.

In the XQuery and XPath data model, XML documents are modeled as an ordered tree. The tree contains seven distinct kinds of nodes: document, element, attribute, text, namespace, processing instruction, and comment. In this research, for simplicity, we only consider document, element, attribute, and text nodes. The data model is a node-labeled, directed graph, in which each node has a unique identity shown in Figure 2. Document order is defined for all the nodes in the document and corresponds to the order in which the
first character of each node occurs in the XML document.

- **Document nodes**: The document node is a virtual node pointing to the root element of an XML document. The document element in a XML document is a child of the document node.

- **Element nodes**: Every element in the document is an element node. Element nodes have zero or more children that can be element nodes or text nodes.

- **Attribute nodes**: Each element node has an associated set of attribute nodes. Note that the element node that owns this attribute is called its “parent” even though an attribute node is not a “child” of its parent element. An attribute node has an attribute name and an attribute value. Attribute nodes have no child nodes. If more than one attribute of an element node exists, the document order among the attributes is not distinguished. This is because there is no order among XML attributes.

- **Text nodes**: A text node must have only one parent and have no child nodes. A text node cannot contain an empty string as its content.

Figure 2: Graphic Representation of XML Data Model
3.1.2 XML Query Model

The query model is specified in generic constructs. We further identify key factors that influence the complexity of each query. This would help users to evaluate performance of the system with increasing complex queries.

The query model we defined can be classified into ten categories. Each of them challenges different aspects of XML processing. Besides, users can specify queries according to their requirements which is called “user-driven query”. Figure 3 shows the XML query model. The following will describe each category briefly, and express each query in generic constructs. In each query, the generic term is written in italics. Then we illustrate them in XQuery. We use E1, E2 etc. to denote a certain element, and A1, A2 etc. to denote a certain attribute. The number of them does not indicate their order in a XML document, just for representing convenience. Finally, the complexity factors will be discussed.
3.1.3 Exact Match

This type of queries specifies a full path expression. One main concept of XQuery is the use of path expressions for selecting nodes. The length of the path expression depends on the levels of predicates being queried in XML documents. This is the simplest query type. We can use this type of queries to establish a simple “metric” comparing performance of the following queries. It tests the database ability to handle simple string lookups with a fully specified path.
Generic constructs are: **Given a full path expression, find elements E1 that have an attribute A1 in a certain value X.** XQuery expression is:

```
FOR $a IN input()/SUBPATH/E1[@A1 = "X"]
RETURN $a
```

### 3.1.4 Joins

References are an integral part of XML identifying the relationship between related data. With using of reference, richer relationships can be represented than just hierarchical element structures. The system must be able to combine separate information together using joins. Horizontal traversals are defined in this type of queries. Joins can be on the basis of references and values. References are specified in the DTD and may be optimized with logical OIDs for example. The system should make use of the cardinalities of the sets to be joined. Joins based on values test the database’s ability to handle large (intermediate) results.

#### Join on Reference

Generic constructs are: **Find element E1 by the reference attribute A1 of E2.**

```
FOR $a IN input()//E1
  $b IN input()//E2
WHERE $a/@A2 = $b/@A1
RETURN $a
```

#### Join on Value

Generic constructs are: **This time reference is based on join of the data values. Find element E1 whose attribute A1 is equal to the attribute A2 of E2.**

```
FOR $a IN input()//E1
  $b IN input()//E2
WHERE $a/@A1 = $b/@A2
RETURN $a
```

### 3.1.5 Regular Path Expressions

Regular path expressions are a basic building block of almost every XML language including XPath, XQuery, and XSLT. The system should
be capable of optimizing path expressions and reducing traversals of irrelevant parts of the tree. We often use wildcards in regular path expressions and the system should realize that it is not necessary to traverse the complete document tree to execute such expressions. This type of queries tries to quantify the costs of long path traversals that do not include wildcards, and the costs of path traversals that include wildcards.

**Full Sub-path**
Generic constructs are: *Find element E1 with a long path expression*. XQuery expression is:

```
FOR $a IN input()/SUBPATH/E1
RETURN $a
```

**Unknown Sub-path**
Generic constructs are: *Find element E1 with a regular path expression include wildcards*. XQuery expression is:

```
FOR $a IN input()//E1
RETURN $a
```

### 3.1.6 Document Construction

Structure is very important to XML documents. But XML documents storing in relational DBMSs often need to be broken down. Reconstructing the original document is a big challenge to systems. We might retrieve fragments of original documents with original structures. But sometimes we may want to construct document fragments with new structures. These queries tests for the ability of the system to reconstruct portions of the original XML document.

**Structure Preserving**
Generic constructs are: *Return a XML document constructed by element E1 and its sub-element E2. Retrieve E2 of E1 that has an attribute A1 equal to a certain value X*. XQuery expression is:

```
FOR $a IN input()//E1[@A1 = X]
RETURN <$a> $a/E2 </$a>
```

**Structure Transforming**
Generic constructs are: *Construct a new XML document. Find element E1 with an attribute A1 equal to a certain value X, and select several sub-element of E1 to construct a new XML document*. XQuery expression is:
3.1.7 Ordered Access

Order of elements is important in XML documents. Because documents will sometimes be fragmented when they are stored on disk, it is important that the order of these fragments in the original document is preserved. The system should be able to preserve these intrinsic orders. This type of queries attempts to test how efficient the system handle queries with order constraints.

Generic constructs are: Find element E1 with attribute A1 in certain value X, and return the first sub-element E2 of E1. XQuery expression is:

```xquery
FOR $a IN input()//E1[@A1 = X]
RETURN $a/E2[1]
```

3.1.8 Sorting

The order by clause is the only facility provided by XQuery for specifying an order other than document order. In XML documents, the generic data type of element content is string, but users may cast the string type to other types. Therefore, the system should be able to sort values both in string and in non-string data types. This type of queries tests whether the system can do sorting efficiently.

**By String**

Generic constructs are: List sub-element E3 of element E1 sorted by sub-element E2. XQuery expression is:

```xquery
FOR $a IN input()//E1
ORDER BY $a//E2
RETURN $a/E3
```
By Non-string

Generic constructs are: List sub-element \(E3\) of element \(E1\) sorted by sub-element \(E2\).

XQuery expression is:

```
FOR $a IN input()//E1
ORDER BY $a//E2
RETURN $a/E3
```

3.1.9 Missing Elements

In XML, schemas are more flexible and may have a number of irregularities. Queries in this type are to test how well the system knows to deal with the semi-structured aspect of XML data, especially elements that are declared optional in the schemas.

Generic constructs are: Find element \(E1\) whose sub-element \(E2\) has NULL value.

XQuery expression is:

```
FOR $a IN input()//E1
WHERE EMPTY($a/E2/text())
RETURN $a
```

3.1.10 Text Search

Text search plays a very important part in XML document systems. This type of queries conducts a full-text search in the form of keyword search. They will challenge the textual nature of XML documents.

Generic constructs are: Find element \(E1\) whose sub-element \(E2\) contains a specific text \(Y\).

XQuery expression is:

```
FOR $a IN input()//E1
WHERE CONTAINS ($a/E2, "Y")
RETURN $a
```

3.1.11 Data-type Cast

Strings are the generic data type in XML documents. But we often need to cast strings to another data type that carries more semantics. These queries challenge the system’s ability to transform between data types.
Generic constructs are: Find element $E1$ with a constraint that contain operations need to transform data value of sub-element $E2$ to other data-type. Retrieve element $E1$ whose sub-element $E2$ is bigger than a certain number $X$. XQuery expression is:

```xml
FOR $a$ IN input()//E1
WHERE $a/E2 > X$
RETURN $a$
```

3.1.12 Function Application

The following query challenges the system with aggregate functions such as count, avg, max, min and sum. Generic constructs are: Group element $E1$ by sub-element $E2$, and calculate the total number of elements for each group. XQuery expression is:

```xml
FOR $a$ IN DISTINCT-VALUES (input()//E1/E2)
LET $b = i nput()//E1[E2 = $a$]
RETURN count($b$)
```

Table 3 summarizes the complexity factors of each query type.

<table>
<thead>
<tr>
<th>Complexity</th>
<th>Low</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exact Match</td>
<td>Shallow Path Expression</td>
<td>Deep Path Expression</td>
</tr>
<tr>
<td>Joins</td>
<td>Two-way Join</td>
<td>N-way Join</td>
</tr>
<tr>
<td></td>
<td>Small Result Size</td>
<td>Large Result Size</td>
</tr>
<tr>
<td>Regular Path Expressions</td>
<td>Shallow Path Expression</td>
<td>Deep Path Expression</td>
</tr>
<tr>
<td></td>
<td>Few Unknown Elements</td>
<td>Many Unknown Elements</td>
</tr>
<tr>
<td>Document Construction</td>
<td>Simple Original Structure</td>
<td>Complex Original Structure</td>
</tr>
<tr>
<td></td>
<td>Simple Output Structure</td>
<td>Complex Output Structure</td>
</tr>
<tr>
<td>Ordered Access</td>
<td>No Index</td>
<td>With Index</td>
</tr>
<tr>
<td>Sorting</td>
<td>Few Qualified Tuples</td>
<td>Many Qualified Tuples</td>
</tr>
<tr>
<td></td>
<td>Single Condition</td>
<td>Multiple Condition</td>
</tr>
<tr>
<td>Missing Elements</td>
<td>Few Generated Tuples</td>
<td>Many Generated Tuples</td>
</tr>
<tr>
<td>Text Search</td>
<td>One Text</td>
<td>Multiple Text</td>
</tr>
<tr>
<td></td>
<td>Few Generated Tuples</td>
<td>Many Generated Tuples</td>
</tr>
<tr>
<td>Data-type Cast</td>
<td>Few Generated Tuples</td>
<td>Many Generated Tuples</td>
</tr>
</tbody>
</table>
3.2 Ontology Data Model

An ontology defines the terms used to describe and represent an area of knowledge. Ontologies are used by people, databases, and applications that need to share domain information. Ontologies include computer-usable definitions of basic concepts in the domain and the relationships among them. An ontology may take a variety of forms, but necessarily it will include a vocabulary of terms, and some specification of their meaning. This includes definitions and an indication of how concepts are inter-related which collectively impose a structure on the domain and constrain the possible interpretations of terms (Uschold, King, Moralee, & Zorgios, 1998). Generally speaking, an ontology consists of the following main constructs (Stevens, Goble, & Bechhofer, 2000; Weißenberg & Gartmann, 2003).

- **Facts** represent explicit knowledge, consisting of:
  - **Classes or concepts** are generalizations of instances. Concepts are the focus of most ontologies. A concept is a representation for a conceptual grouping of similar terms. A concept can have subconcepts that represent concepts that are more specific than the superconcept. Concepts fall into two kinds of (1) primitive concepts are those which only have necessary conditions (in terms of their properties) for membership of the class. (2) Defined concepts are those whose description is both necessary and sufficient for a thing to be a member of the class.
  - **Properties** can be subdivided into scalar attributes and non-scalar relations. The property can be defined to be a specialization (subproperty) of an existing property. An attribute is a property of a concept that refers to a datatype (integer, string, float, boolean etc.). An example of an attribute is “has-name” related to a string. A relation is a property of a concept that refers to another concept. Specialization / Generalization are one of the standard relations. For instance, “is a kind of” defines a relation that may be applied to the concepts “Enzyme” and “Protein”.
  - **Instances** represent individual entities and are connected by type-of relation to at least one
class; some authors only consider facts about instances as real facts. Strictly speaking, an ontology should not contain any instances, because it is supposed to be a conceptualization of the domain. The combination of an ontology with associated instances is what is known as a knowledge base. However, deciding whether something is a concept of an instance is difficult, and often depends on the application.

- Axioms are rules used to add semantics and to infer knowledge from facts. In contrast to facts, they represent implicit knowledge about concepts and relations, e.g., whether a relation is transitive or symmetric.

### 3.3 Ontology Query Model

Initially, ontologies are introduced as an “explicit specification of a conceptualization”. In an intelligent information integration system, ontologies can be used to establish common vocabularies and semantic interpretations of terms from information sources. With respect to the integration of data sources, they can be used for the identification and association of semantically corresponding information concepts. People can share and exchange information in a semantically consistent way.

Using ontology basic components described in the previous section, users can define their own ontology in any application domain. Then we conduct a series of tests to see how the system handles such ontologies. The operation model in the ontology workload model is a set of queries, and the answers are generated by inferring from the ontology. The queries we present here are representative for different application domains. We conclude the reasoning tasks and construct six basic reasoning queries for the ontology benchmark. Figure 4 shows the ontology query model in this research.
- **Concept Subsumption Queries:** checks if one concept is a subconcept of another. Generic constructs are: *Given concepts C and D, determine if C is a subconcept of D with respect to ontology O.*

- **Concept Hierarchy Queries:** determines the concepts that immediate subsume or are subsumed by a given concept. Generic constructs are: *Given a concept C return all/most-specific superconcepts of C and/or all/most-general subconcepts of C.*

- **Concept Consistency Queries:** checks for (in)consistency of concept definitions. Generic constructs are: *Given a concept C, determine if the definition of C is generally satisfiable (consistent).*

- **Instance Checking Queries:** given a partial description of an individual (instance) and a concept description, finds whether the concept describes the instance. Generic constructs are: *Given a concept C, determine whether a given individual A is an instance of C.*

- **Instance Retrieval Queries:** finds all instances that are described by a given concept. Generic constructs are: *Given a concept C, determine all the individuals in ontology O that are instances of C.*

- **Instance Realization Queries:** given a partial description of an instance, finds the most specific concepts that describe it. Generic constructs are: *Given an individual A, determine all the concepts in ontology O that A is an instance of.*
When querying the intelligent information integration system, the reasoning service may not be so straightforward. We need to evaluate the ontology with increasing complexity. When formulating the complex benchmark queries, several factors should be taken into account (Guo, Heflin, & Pan, 2003). More complex queries may be formulated according to these factors. These allow the system be evaluated under different level of workloads.

- **Input size**: This is measured as the proportion of the class instances involved in the query to the total class instances in the benchmark data.
- **Selectivity**: This is measured as the estimated proportion of the class instances involved in the query that satisfy the query criteria.
- **Complexity**: We use the number of classes and properties that are involved in the query as an indication of complexity.
- **Hierarchy information assumed**: This considers whether information of class hierarchy or property hierarchy is required to achieve the complete answer. Besides, the depth and width of class hierarchies should also been considered.

### 3.4 Test Database Generation

In order to evaluate the performance of the intelligent information integration system, we must define the test database. The workload consists of a test operation and a test database. The test database identifies what data must be loaded into the data sources, as well as the volume of the test data. Information integration system data sources are disparate and heterogeneous. Information comes from various sources (including structured, semi-structured and unstructured sources) and formats (such as database tables, XML files, PDF files, streaming media, internal documents, and Web pages). For this research, the data sources can be divided into three kinds: relational databases, object-oriented databases, and Web pages. For each data source, we must analyze the actual data and extract statistical data. Data analysis characterizes data in terms of the size of the database, the number of records, the length of records, the types of fields, and the value distributions.

- **Determine data values**: A number of data types are supported in this research, including long integer number, double precision floating point number, decimal number, money, datetime, fixed-length and variable-length character strings. We must conduct extensive studies to characterize each data source with several distribution parameters. Frequency distributions are
computed and standard probability distributions are fit to the data in order to generate the value of test data. Data values are created with these common data distributions such as exponential, normal, discrete, rotating, zipfian \(^2\) or uniform distribution.

- **Determine scaling factors:** After determining the value of the test data, we must define how much data should be generated, i.e. defining the database scaling factor. Generally speaking, the logical size for the test database used for the benchmark is at least equal to the logical size of physical memory on the host(s). For this research, we refer to the AS³AP benchmark standard.

- **Open data source:** We must determine the test data of the open data source on the Web, but this is problematic. There is in excess of 10 billion pages on the Web, which include HTML files, text documents, PDF files, Microsoft Office documents and other similar data files. We cannot possibly download every page from the Web much less adequate sample size. Even the most comprehensive search engine currently indexes just a small fraction of the entire Web.

As such, it is important to carefully select the so-called “important” pages, so that the fraction of the Web that is visited becomes more meaningful. In order to select these important pages, we can use several metrics for prioritizing them. For any given web page, we must define its importance using the following methods (Arasu, Cho, Garcia-Molina, Paepcke, & Raghavan, 2001):

- **Interest-driven.** The goal is to obtain pages of interest to a particular user or set of users. Important pages are those that match user interest. One particular way to define this notion is through what we call a *driving* query. For any given query, the importance of a page is defined by the “textual similarity” between the page and the driving query. Assuming that query represents the user’s interest, this metric shows how relevant the page is. Another interest-driven approach is based on a hierarchy of topics. Interest is defined by a topic, and we attempt to guess the page topics that will be visited by analyzing the link structure that leads to the candidate pages.

- **Popularity-driven.** Page importance depends on how popular a page is. For instance, one way to define popularity is to use a page’s backlink count. Intuitively, a page that is linked to by many pages is more important than one that is seldom referenced.

- **Location-driven.** The importance
of a page is a function of its location, not its contents. For example, URLs ending with “.com” may be deemed more useful than URLs with other endings, or URLs containing the string “home” may be of more interest than other URLs. Another location metric that is sometimes used considers URLs with fewer slashes more useful than those with more slashes.

(3) Control Model

The control model defines the environment setup variables to execute the experiment. A common set of parameters including the steady state, the test mode, the test duration, the test sequence and the number of repetitions should be specified as follows.

- **Steady State**: The benchmark test must be executed in a steady state, in order to return the sustained system performance.

- **Test Mode**: There are three kinds of test mode, that is, cold mode, warm mode, and hot mode. In cold mode, there is no data in the cache. The system cannot retrieve data from the cache directly. Therefore, the performance in cold mode is usually slower than the other two modes. In warm mode, the data is left in the cache from the prior query. Because of that, the test response time decreases. In hot mode, a query is executed in cold mode first, and then is executed with the cache data for several times. The average response time is computed.

- **Test Duration**: Test duration means time intervals of the benchmark. Each interval must begin after the system has reached steady state and is long enough to generate the reproducible throughput. Each interval must extend uninterruptedly for a period of time.

- **Test Sequence**: Test sequence indicates the order of the queries to be executed.

- **Number of Repetitions**: Number of repetitions means execution repeated times.

(4) Performance Metrics

Performance metrics can be divided into two types, i.e., the speed-specific metrics and the relevance-specific metrics. The former consists of the metrics of response time and throughput. The latter has the metrics of relative recall and precision.

- **Response time**: Response time refers to the time interval between when a request is made and when the response is received by the requester.

- **Throughput**: Throughput refers to the number of operations completed by the system per unit time.

- **Recall and precision**: Recall and
precision refer to two important measures of evaluation of information retrieval. However, it is very difficult, if not impossible, to directly apply these measurements to the evaluation of Web information retrieval systems due to the unique nature of the Web. There is no proper method of calculating absolute recall of search engines as it is impossible to know the total number of relevant in huge databases. The relative recall value is defined in (Clarke & Willet, 1997).

3.5 Experimental Design

We use the university campus as the test scenario of the benchmark prototype. It describes universities and departments and the activities that occur at them. The global schema for the intelligent information integration prototype system is shown on Appendix A. For the Tamino XML server, we have defined six schemas: Department, Faculty, Student, Grad Student, Course, and Publication. These schema details are shown on Appendix C.

3.6 Prototype Functions

3.6.1 Data Loader

In the data loader, users can define the logical database size. The default logical database sizes for each data source are uniform and users can modify them individually. The data sources and data schema in the prototype are described in the previous section. Furthermore, in the SQL Server users can adjust the number of tuples loaded into each relation. Similarly, users can alter the number of instances generated for each schema in Tamino XML server as shown in Figure 5.
3.6.2 Query Generator

The query generator can be divided into two parts: pure XML query and XML query combined with ontology reasoning, as shown in Figure 6. Users can evaluate the pure XML processing performance of intelligent information integration systems or combined ontology reasoning services. In each of them, users can input the queries they want to test or simply select the standard query types predefined in the query selector. The queries are generated according to the XML query model. In open query input, users can specify several queries according to the global schema and their requirements. All specified queries will be executed in one test, and users can delete any one in the open query list as shown in Figure 7-Figure 9. In the standard query selector, after choosing the standard test queries, users can specify the complexity of each query further. It would help users to evaluate the system performance at different complexity levels. The complexity of query is determined by the complexity factor.
Figure 6: Query Generator

Figure 7: Query Generator - Input
3.6.3 Scheduler

According to the control model, several parameters should be set to execute the benchmark. The parameters we implement in the prototype are test sequence and number of repetitions. Both in the open query input and in the standard query selector, once the test
query set has been determined, users can set up the executed sequence and the repetitions of each query in the scheduler.

3.6.4 Result Collector

The result collector shows the test results of the queries we specified. The test results can be divided into three parts: the total execution result, the XML query test result, and test result of each data source. If the test query needs to combine ontology reasoning service, the ontology query test result will be shown in the test result. In the total execution result, the total query response time and throughput are illustrated. In the XML query test result, it shows the query response time and throughput of the XML query processing as shown Figure 10. In the ontology query test result, it adds two extra performance metrics: recall and precision. This can help users evaluate the quality of answer entailed by the ontology. Finally, the result collector lists the query response time and throughput of each data sources.

![Figure 10: Result Collector](image-url)
4. Research Implications and Concluding Remarks

4.1 Research Implications

In this research, we have accomplished four main tasks. First, an analysis framework of web search and benchmark literature to lay the basis of generic construct development is developed. Our aim is to collect all related literature on the classic web search algorithms and the benchmark methods. We collected and compiled the key web search algorithms and the benchmark methods summarized to be representative. Secondly, a set of heuristics to formulate the generic constructs of web search algorithms are presented. Generic constructs are extracted from the main web search algorithms and the benchmark methods. We analyzed the algorithms and find the essential constructs. For instance, PageRank is based on inlinks and outlinks of the page so these become the key components of the algorithm where the web page” is a generic construct and the “tag” is the operation of the generic constructs. Thirdly, a more representative and reproducible workload model of web search is created. The generic constructs of a web page is extracted into the page model. The generic constructs of the search types are extracted into the query model. Designed as such, this benchmark meets the desired characteristics of scalability, portability and simplicity. Fourthly, a computer-assisted benchmarking process is implemented in a prototype system. The prototype system is designed to help prove the feasibility and validity of the research method.

In this research, we have described a detailed approach to model workload requirements from the user's perspective. This results in a more realistic environment of workload representation, transformation, and generation. We have delineated the main components of the method. They include the workload specification scheme, the scheme translator, and the data and operation generators.

The method is domain-representative and workload-representative because we model from the user problem domain and characterize from the user application. The benchmark method is scalable because we can scale up or down the problem size and the problem complexity by changing the data definition and the operation definition via specification. It is reproducible because we use a high-level specification scheme to describe the general workload requirements. The method enables a custom benchmark where users can control the execution through requirements specification instead of manual manipulation.

In the new benchmark method, we have presented a common carrier
concept to capture and compose the user requirements into three carrier components of model. They are the data component model, the operation component model, and the control component model. Web search experiment requires a page model similar to the object model to abstract web as a directed labeled graph in which the nodes model objects and the outgoing edges of an object model the attributes of the object. Designed as such, the benchmark conforms to the desirable characteristics of relevance and rigorousness.

There are several limitations in this research due to the time and resource constraints. We did not verify all performance indicators through the prototype system. The validity of this research can be further improved. The data generator of the prototype system is primitive.

- Due to the infinity of the Internet, we cannot precisely verify all performance variables. Thus, the validity of this research is limited which can be further improved.
- In the prototype, the query generator developed is primitive. So far, it depends on the extent of functions supported by the web search service APIs.
- The experiments are mainly the basic and synthetic tests. Thus, the comprehensiveness and completeness of experiments can be enhanced.

The future research will continue to augment the experimental prototype in order to accommodate a larger set of data, more complicated operations, more data distribution types, and a wider collection of performance metrics. We plan to develop an expert system to analyze benchmark results, pinpoint performance bottlenecks, provide possible reasons for the test results and advise on the actions to take. In addition, we will further quantify the advantages of the method in the form of metrics on cost and quality. In the future, we further enhance the method to provide users and managers the means to diagnose and detect the strength and weakness of each benchmark.

- Add more advanced generic constructs: Web search benchmark and development for new algorithms is a continuing effort. Continuously collecting the new web search-related literature can help find more advanced generic constructs. Adding more advanced generic constructs to a workload model can advance the generality.
- Enhance the complexity of tests: The experiments we have performed only include the baseline test suites of algorithms. In order to completely verify the workload model, we need to test more new
algorithms. Due to the limitations of the web search APIs, we have only designed ten simple tests to be applied with the APIs. If more web search APIs can be available, we can perform more complicated tests in the future. Another direction is to provide a comparison of experimental results with those of different search engines besides Google and Yahoo.

- Enhance the features of the prototype: The prototype system can be expanded to include the rest of the features of the research method in the future.

4.2 Concluding Remarks

In this research, we have developed the XML and ontology benchmark workload model in intelligent information integration, and built a workload generation prototype. We have reviewed the XML and ontology related literature to motivate the design of the workload model. The objective of this research is to develop a workload model to test whether the intelligent information integration system under EB environment can overcome the diverse formats of content and derive meaning from this content. In order to apply the workload model to different scenarios easier, it is designed in generic constructs. Finally, we validate the research model through the prototype implementation.

- Enhancing the ontology query model. The development of an ontological standard presents many opportunities and challenges. New reasoning tasks may arise in the future. Retrieval (instances of a concept) and realization (most specific class of instance) may not be sufficient. In order to make the ontology query model more comprehensive, further study to keep track of ontology progression is needed.

- Improving the complexity factors of the XML query model. The complexity factors we analyze in the XML query model are still too rough. Each query type can be analyzed more carefully to refine the query model.

- Implementing various data distributions. In this research, only uniform distribution is implemented. It cannot evaluate performance under different distributions. Implementation of diverse data distributions will become a user requirement.

- Applying the workload model to other applications. Ontology and XML are complementary technologies, and there are other applications that can apply. In this research, we assume the intelligent information
integration system is used on Intranets, such as enterprise information integration (EII), electronic business (EB), and enterprise application integration (EAI). There are other applications between enterprises that may need to integrate heterogeneous information, such as business-to-business integration (B2Bi), collaborative commerce (C-Commerce), and electronic commerce (EC). We can modify the workload model of this research to create other benchmarks that are based on XML and ontology with different characteristics.
5. References


Shanmugasundaram, Shekita, Kiernan, Krishnamurthy, Viglas,


6. 計畫成果自評

資訊科技和網際網路的蓬勃發展，異質資訊整合在企業電子化與電子商務環境中已是一項普遍存在且重要的議題。在缺乏整合的情形下，個別地存取異質資訊來源，已造成資訊的混雜和錯誤及浪費，尤其不能提供即時管理決策分析給企業主管。本研究發展出一項能夠整合不同資料模型，以及這資料模型中衍生出的語意，依照延伸標記語言與本體論以及學名式的資料結構進行建模，並達到負載量模型具有可攜性和延展性的趨近一般化的績效評估方法。此資料模型結合本體和延伸標籤語言的查詢方法，提高對底層異質資訊來的結構互動性和語意互動。基於學名結構方法，發展單一的標準轉換模式和交換模式，促成分散地的資料倉儲間形成多對多的系統以對模式，輔以多維度中介資料管理功能，可形成網路上通用並能兼顧效率與品質的離型系統。對於在電子商務環境中，企業對於大量的資料和資訊，倚賴有效系統工具來協助進行萃取、分析和預測的重要商業智慧利器。本研究的部分研究成果已發表在 *Information & Management*, *Information Processing & Management*, *Information Sciences*, *Expert Systems with Applications* 與國科會 A 級國際學術研討會論文（附錄）。
（附錄）論文著述：

A. 期刊論文


43. Seng*, Jia-Lang (1998)，A hybrid design of cost accounting information systems，政大學報，第76期，6月，329-344。NSC87-3011-P-301-001-。(TSSCI Extended)

B. 研討會論文


43. Seng*, Jia-Lang (2002), A Web-Based Data Warehouse Planning, Global Perspective of Information Technology Management, IRMA.
44. Seng*, Jia-Lang and Han, S. (2002), A Representation of Distributed DBMS Benchmark Workload Model, Information Systems Evaluation Management, IRMA.
High-Tech Industry, National Science Council Annual Conference in Management, Taipei, Taiwan, October.


59. Seng*, Jia-Lang, (2001), A Web-based DW Planning- A Large Scale Project Management Information Systems in Taiwan, Global Perspective of Information Technology Management, IRMA.


75. Seng*, Jia-Lang and Tsai (1999), A structured transformation approach to
legacy information systems, The 32nd Hawaii International Conference on Systems Sciences (HICSS), Hawaii USA.


development, The 2nd Annual Americans Conference on Information Systems (AMCIS), Phoenix, Arizona, USA.


102. Seng*, Jia-Lang (1993), Extended Abstract, Pre-ICSE Workshop, Baltimore, Maryland, USA.


104. Seng*, Jia-Lang (1993), A mathematical model of requirements analysis, Pre-SIGMOD Workshop, Washington, D.C., USA.


C. 專書及專書論文
8. Seng*, Jia-Lang, 1994, Requirements-Driven Database Systems Benchmark Methodology, University of Maryland at College Park, USA.
國科會補助專題研究計畫項下出席國際學術會議心得報告

日期：99年12月30日

<table>
<thead>
<tr>
<th>計畫編號</th>
<th>NSC 96 - 2416 - H - 004 - 017 - MY3</th>
</tr>
</thead>
<tbody>
<tr>
<td>計畫名稱</td>
<td>智慧型資訊整合於異質資料倉儲和資料探勘之模型、架構、與績效評估 - 應用本體論、母型綱要、和學名結構</td>
</tr>
<tr>
<td>出國人員姓名</td>
<td>諶家蘭</td>
</tr>
<tr>
<td>服務機構及職稱</td>
<td>政治大學會計系教授兼系主任</td>
</tr>
<tr>
<td>會議時間</td>
<td>99年12月16日至99年12月18日</td>
</tr>
<tr>
<td>會議地點</td>
<td>香港科大</td>
</tr>
<tr>
<td>會議名稱</td>
<td>(中文)香港科大學計研究學術研討會</td>
</tr>
<tr>
<td>(英文)2010 HKUST Accounting Research Symposium</td>
<td></td>
</tr>
<tr>
<td>發表論文題目</td>
<td>(中文)</td>
</tr>
<tr>
<td>(英文)</td>
<td></td>
</tr>
</tbody>
</table>

一、參加會議經過

香港科大是亞洲首屈一指的科技與商管教研機構，結合世界頂尖研究學者，屢獲全球教學、研究殊榮、以及國家最先進計畫資源等，提供最先進的見解和國際視野。香港科大會計與資訊研討會在會計界中為重要的學術研討會，2010香港科大會計與資訊研究研討會邀請各領域中的會計與資訊研究學者參加對於會計議題的討論與發表論文。

本次會議為期三天 (12/16 ~ 12/18)，16號晚上舉行會前會議，提供與會者一個互動交流的機會。17~18號二天為正式大會。會議程主要包括開幕式、專題演講、分組報告與討論、座談等，大會與其它座談會議程摘錄如表1。

專題演講「Challenges and Opportunities in Disclosure Research」主講者Philip Berger P.D.的研究大量的發表在會計與財務期刊，曾獲得多項的研究獎，研究領域主要在財務報告和企業融資，包括 (1) the valuation consequences of diversification strategies; (2) the use of financial statement data to value real options; (3) the impact of managerial entrenchment on corporate finance and financial reporting decisions; (4) the effect of improved financial statement disclosure on capital markets; (5) the motives for choosing opaque versus transparent financial reporting practices and (6) factors influencing R&D investment and financing decisions including the use of off-balance-sheet financing.
二、與會心得

參加本次為期三天的會議，受益良多，各與會專家、學者對會計與資訊議題的精彩討論與見解，使本人在會計與資訊議題上的學術研究與發展有更廣的視野與學習，並透過參與會議，可以與更多優秀學者進行學術交流與研究合作，是發掘重要研究議題、瞭解國際學術脈動、與拓展研究合作團隊的最佳機會與管道。

三、考察參觀活動（無是項活動者略）

無

四、建議

中國大陸在科技與財經上有長期的進步並與國際接軌。感謝國科會支持使得本人能夠進行研究與參與研討會和訪問，希望國科會能繼續給予老師在學術研究上的大力支持。

五、攜回資料名稱及內容

會議論文集

六、其他
HKUST ACCOUNTING RESEARCH SYMPOSIUM

DECEMBER 17-18, 2010

December 16, 2010 (Thursday)
19:00 – 22:00 Welcome Reception at UC Bistro

Day 1: December 17, 2010 (Friday)
Venue: Lung Kai Lam Theatre (Rm 6602, Classroom 5)
(Lifts 31-32, 6th Floor)
08:30 – 08:45 Tea and Refreshments

SESSION 1
Moderator: Allen Huang, HKUST

08:45 – 09:00 Dean’s Welcome and Remarks from the Conference Organizing Committee
Leonard Cheng, Dean, HKUST
Kevin Chen, Head of Accounting Department, HKUST

09:00 – 10:00 Challenges and Opportunities in Disclosure Research
Keynote speech by Professor Philip Berger, University of Chicago

10:00 – 10:15 Tea Break

10:15 – 11:15 Differences of Opinion, Short-Sale Constraints, and Voluntary Disclosures
by Sanjeev Bhojraj, Yan Li, and Holly Yang
Presenter: Holly Yang, Wharton School, University of Pennsylvania
Discussant: Vicki Tang, Georgetown University

11:15 – 12:15 A Unified Framework of Management Earnings Forecasts: Voluntary, Disclose or Abstain, and Opportunistic Incentives
by Edward Li, Charles Wasley, and Jerold Zimmerman
Presenter: Charles Wasley, University of Rochester
Discussant: Rebecca Hann, University of Maryland

12:15 – 14:00 Lunch at Multi-Purpose Hall (next to Lung Kai Lam Theatre)
SESSION 2
Moderator: Shiheng Wang, HKUST

14:00 – 15:00  The Effect of Regulator Oversight on Firms’ Information Environment: Securities and Exchange Commission Comment Letters
by Reining Chen and Rick Johnston
Presenter: Rick Johnston, Purdue University
Discussant: Clive Lennox, Nanyang Technological University

15:00 – 16:00  Lending, Lying, and Costly Auditing
by John Kareken and Jack Stecher
Presenter: Jack Stecher, Carnegie Mellon University
Discussant: Kirill Novoselov, HKUST

15:00 – 16:00  The Effect of Regulator Oversight on Firms’ Information Environment: Securities and Exchange Commission Comment Letters
by Reining Chen and Rick Johnston
Presenter: Rick Johnston, Purdue University
Discussant: Clive Lennox, Nanyang Technological University

15:00 – 16:00  Lending, Lying, and Costly Auditing
by John Kareken and Jack Stecher
Presenter: Jack Stecher, Carnegie Mellon University
Discussant: Kirill Novoselov, HKUST

16:00 – 16:15 Tea Break

16:15 – 17:15  The Debt Market Relevance of Management Earnings Forecasts: Evidence from Before and During the Credit Crisis
by Lakshmanan Shivakumar, Oktay Urcan, Florin P. Vasvari, and Li Zhang
Presenter: Oktay Urcan, London Business School
Discussant: David Reeb, Temple University

Day 2: December 18, 2010 (Saturday)
Venue: Lung Kai Lam Theatre (Rm 6602, Classroom 5)
(Lifts 31-32, 6th Floor)

08:30 – 08:45 Tea and Refreshments

SESSION 3
Moderator: Haifeng You, HKUST

08:45 – 09:45  Loan Loss Reserves, Regulatory Capital, and Bank Failures: Evidence from the 2008-2009 Economic Crisis
by Jeffrey Ng and Sugata Roychowdhury
Presenter: Sugata Roychowdhury, Boston College
Discussant: Jeff Yu, Southern Methodist University

09:45 – 10:45  Are Analysts’ Cash Flow Forecasts Naïve Extensions of Their Own Earnings Forecasts?
by Andrew C. Call, Shuping Chen, and Yen H. Tong
Presenter: Shuping Chen, University of Texas at Austin
Discussant: Zhaoyang Gu, University of Minnesota
10:45 – 11:00  Tea Break

11:00 – 12:00  Informativeness of Text in Analyst Reports: A Naïve Bayes Machine Learning Approach
by Allen Huang, Amy Zang and Rong Zheng
Presenter: Allen Huang, HKUST
Discussant: Qiang Cheng, University of Wisconsin-Madison

12:00 – 14:00  Lunch at China Garden Restaurant (G/F)
國科會補助專題研究計畫項下出席國際學術會議心得報告

計畫編號：NSC 96 - 2416 - H - 004 - 017 - MY3

計畫名稱：智慧型資訊整合於異質資料倉儲和資料探勘之模型、架構、與績效評估 - 應用本體論、母型綱要、和學名結構

出國人員姓名：譚家蘭

服務機構及職稱：政治大學會計系教授兼系主任

會議時間：99年12月2日至99年12月9日

會議名稱：
(中文)2010中國會計財務研究國際研討會
(英文)2010 China Accounting and Finance Review International Symposium
(中文)上海復旦大學管理學院學專題演講
(英文)Seminar Series, School of Management, Fudan University, Shanghai, China

發表論文題目：
(中文)探討以XBRL為基礎之財務資訊整合
(英文)Exploring an XBRL-assisted Approach to Integrate Financial Information

一、參加會議經過

本次受邀參加上海「2010中國會計財務研究國際研討會」，本次會議主題為「如何開展以中國為基礎的財務及會計資訊研究」，探討中國會計資訊、財務資訊、經濟發展以及企業改革方面等的
最新議題;並受邀上海復旦大學管理學院會計研究學術專題演講，主題為「探討以 XBRL 為基礎的財務資訊整合」。

（1）香港理工大學會計及金融學院與上海財經大學會計學院及上海財經大學會計與財務研究院於 2010 年 12 月 4 日至 5 日，在上海聯合舉辦“2010 年中國會計與財務研究國際研討會”。本次研討會近百名來自中國內地、香港和海外的專家、學者齊聚一堂，針對中國會計資訊、財務資訊、經濟發展及企業改革方面的議題進行廣泛與深入的討論。大會議程主要包括開幕式、專題學術報告、特別專題演講、分組報告與討論、座談等，大會與其它座談會議程摘錄如表 1。與會期間，香港理工大學副校長徐林倩麗教授和上海財經大學副校長孫錚教授就會議主題作了開幕致辭及五位著名學者（依演講順序）辜飛南教授、徐林倩麗教授、柯濱教授、顧朝陽教授及李真教授作了精彩的主题演講。

（2）XBRL（eXtensible Business Reporting Language，可延伸商業報導語言）以 XML（eXtensible Markup Language，可延伸標記語言）技術為基礎，藉由國際標準之訂定，建構全球企業資訊供應鏈，以方便各階段參與者得以更有效率的方式取得、交換與分析比較企業的各項資訊，可解決日益繁雜之資訊揭露問題。中國是最早強制要求採用 XBRL 之國家，一開始是由上海證券交易所於 2003 年推動自願申報計劃，當時約有 50 家公司參與；此後，在 2004 年時申報第 1 季季報期間即已全面推廣採用 XBRL 申報財務報告摘要，而在 2005 年起也開始要求上市公司採用 XBRL 申報財務報告全文，目前已有超過 800 家上市公司採行。上海復旦大學管理學院會計研究院以探討 XBRL 為基礎的財務資訊整合為議題，進行學術專題演講。本次發表論文「探討以 XBRL 為基礎的財務資訊整合」，研發 2 個系統軟體 XBRL FR 轉換軟體、XBRL GR 轉換軟體，以解決財務會計與審計問題。自 2005 年開始，主管機關欲解決異質財務報表的問題並與國際接軌，本研究以人工智慧方式以整合不同格式與不同準則之財務報表。首先將最常用的微軟 Office 系列格式轉換成 XML，再加上台灣主管機關的分類標準，開發出系統軟體。利用人工智慧進行資料判讀，轉換標準格式，第一個系統 XBRL FR 轉換軟體為對外針對財務報表層次，第二個系統 XBRL GR 轉換軟體對內針對財務交易層次。

二、與會心得

本次參加「2010 中國會計與財務研究國際研討會」與「上海復旦大學管理學院會計研究學術專題演講」受益良多，與會期間，各地專家、學者們參與踴躍，在研討會中針對中國會計資訊、財務資訊、經濟發展及企業改革方面的議題進行廣泛與深入的討論及在學術演講中，針對本人所發表的理論模型亦提出建議並廣泛交換意見。

三、考察參觀活動(無是項活動者略)

無

四、建議

中國大陸在科技與財經上有長期的進步並與國際接軌。感謝國科會支持使得本人能夠進行研究與參與研討會和訪問，希望國科會能繼續給予老師在學術研究上的大力支持。

五、攜回資料名稱及內容

會議論文集

六、其他
初探以 XBRL 为基础之财务信息整合

黏凯婷
國立政治大学会计系博士班
台北市、台湾

谌家兰*
國立政治大学会计系特聘教授兼系主任
國立政治大学会计系大陆学术交流委员会召集人
台北市、台湾

魏长风
资深工程师
國立政治大学 XBRL 研究推广中心
台北市、台湾
Exploring an XBRL-assisted Approach to Integrate Financial Information

K. T. Nan
Ph.D. Candidate
Department and Graduate School of Accounting
College of Commerce
National Chengchi University
Taipei, Taiwan 116

Jia-Lang Seng*, Ph.D.
Distinguished Professor and Chair
Department and Graduate School of Accounting
Taiwan-China Research Exchange Committee
College of Commerce
National Chengchi University
Taipei, Taiwan 116
seng@nccu.edu.tw

C. F. Wei
Senior System Engineer
XBRL Research Center
National Chengchi University
Taipei, Taiwan 116

*corresponding author
Abstract

Integrated financial information presents the substantial and financial condition of an enterprise. The provision of complete information to report is important for managers, investors, and supervisory commissions. Currently, most consolidated financial reports are compiled manually which is ad hoc, subjective, time-consuming, and error-prone. In order to resolve the inadequacies and difficulties of integrated financial information, information technology and integrated financial information are integrated in this paper to propose three computer-aided integrated financial information compilation approaches. A field study and on-site interviews have been conducted with large-scale ERP software companies and independent consultants to probe into problems associated with integrated financial information compilation and systematization. Based on the field study interviews and analysis results, three categories of computer-aided integrated financial information methods are described in this paper. They are the conventional mapping table approach which is used within homogeneous ERP systems, the XBRL FR-based approach which supports heterogeneous ERP systems and adopts the emerging international standard, and the XBRL GL-based approach which transforms the compilation to comprehensively re-engineer the processes in an automatic manner. Detailed steps of each approach are discussed to illustrate the feasibility and validity of the research method.

Keywords: Extensible Business Reporting Language, Integrated Financial Information, Consolidated Financial Information, Enterprise Resources Planning System, Heterogeneous Information Systems, Mapping Table, XBRL FR, XBRL GL
1. Introduction

1.1 Research Motivation

The purpose of ‘integrated financial information’ provisioned by an enterprise is to provide report users with financial and non-financial information related to the substantial and financial condition of the enterprise. In accounting, substantial and financial conditions are prioritized according to legal aspects; thus, the concept of integrated financial information. Due to considerations of business operations, the enterprise frequently operates as two or more entities (e.g. parent and subsidiary companies operations). Therefore, if financial report users merely read financial reports from a single enterprise, they would fail to gain an accurate insight into the overall economic activities occurring in the business group (Hartgraves et al., 2002). As a result, the parent company needs to compile integrated financial information to integrate the business conditions of all its affiliate companies and present as a single financial entity. Therefore, report users are able to clearly understand the overall financial conditions of the financial entity, its business performance, and changes in the financial conditions (Maines et al., 2003).

Integrated financial information involves multiple enterprise entities. The transactions that take place in these report-compiling entities are often complex. When compiling integrated financial information, the confirmation of the bodies of integrated financial information and the internal transaction write-offs in affiliate companies are two important issues. In the global layout of enterprises, the parent and subsidiary companies are distributed worldwide, and different currencies are used to compute unit prices. Reports based on accounting standards of different nations have made it much more difficult for enterprises to compile financial reports.

At present, the ERP (Enterprise Resource Planning) system is a system designed to record daily business activities and is used extensively by enterprises. Financial reports of a single enterprise may also be prepared through the ERP systems. However, the ERP systems have not been able to compile consolidated financial reports of the affiliate companies of an enterprise. Currently, most consolidated financial reports are prepared manually. This type of method not only demands a huge input of labor, it is also prone to human errors that are difficult to detect (Pinker, 2005). This results in prolonged information time and increased incidence of errors. The timeliness and accuracy of the integrated financial information therefore diminish and make it less useful.

XBRL (eXtensible Business Reporting Language) is an XML (eXtensible Business Language) language used to describe the financial conditions of an
enterprise. In other words, it is the language that expresses and presents the financial reports of the enterprise. XBRL not only provides a set of standard languages that facilitates the conversion of financial report data between systems, it also promotes the information transparency of financial reports and enhances added values of financial reports of the enterprise (Bovée, 2002; Hodge et al., 2004). The standard financial reporting language of XBRL provides a common language framework that makes the information transformation more efficient and effective (Pinsker and Li, 2008).

XBRL-U.S. was established by American Institute of Certified Public Accountants (AICPA) in October, 1999. XBRL International was also established in July, 2000. With active promotions undertaken by Securities and Exchange Commission (SEC), the adoption of XBRL enterprises is growing rapidly. Many the ERP systems have also added XBRL support. The XBRL2.0 and 2.1 Taxonomy of U.S. and International Accounting Standards were released in 2000 and 2003, respectively. Enterprises are able to prepare XBRL instance documents based on the released XBRL standard taxonomy package. Besides the U.S.A., fifteen other countries, including Canada, U.K., Australia, New Zealand, Ireland, Netherlands, Germany, South Korea, Denmark, Sweden, Belgium, France, and South Africa, as well as International Accounting Standards Board (IASB), have also set up XBRL organizations. In addition, XBRL organizations are now undergoing active planning in over 20 countries, including China and Taiwan. They are now ready to promote XBRL instance document documentation works. Thus, financial reports of enterprises expressed and presented through XBRL’ has become a global trend (Glen, 2008, Bovée et al., 2005).

In view of this, the framework and method of computer-aided collection and compilation would be incorporated and integrated into financial information in this paper. Through a field study and analysis, the research problems related to the computer-aided compilation of integrated financial information encountered by enterprises are addressed with one conventional alternative and two XBRL-centric approaches to tackle the ad hoc, subjective, time-consuming, error-prone manual handling of integrated financial information.

1.2 Research Issue

Following the outbreak of Enron case in the U.S.A., nations around the world began valuing the transparency and quality of financial information (Pinsker and Li, 2008). IAS also released No. 27 Accounting Research Bulletin (ARB)-Consolidated Financial Statement in 2003 to regulate the integrated financial information
compilations. Integrated financial information is crucial for report users; however, the financial information compilation process is subjective, complex, and tedious. Therefore, the advancement of information technology be adopted and applied in this study. This allows enterprises to quickly and accurately compile consolidated financial reports. In this paper, several techniques of computer-aided approach would be proposed and the advantages and disadvantages of each be analyzed.

Based on the literature review and field study, this study has collected and compiled eight types of problems related to integrated financial information preparation and process. They are (1) problems related to accounts mapping, (2) problems related to internal transaction write-off, (3) problems related to shareholding ratio, (4) problems related to accounting inconsistency, (5) problems related to lack of flexible adjustment of integrated financial information, (6) problems related to foreign currency conversion, (7) problems related to disclosure information collection, and (8) problems related to integrated financial information audit.

Based on the literature review and field study, XBRL provides a common data format for sharing business reporting information across applications. XBRL format has an impact on how business data are collected, controlled, analyzed, and reported to external users. The source of XBRL-related cost savings has increased the processing capability and decreased the data redundancy. Other identified cost savings are in increased efficiency (less data redundancy) and decreased costs of bookkeeping. Therefore, a field study and analysis is conducted to propose the XBRL-centric approach to prepare the integrated financial information.

1.3 Research Approach

In this paper, through a field study and on-site interviews with experts and enterprises, the research issues encountered in enterprises’ financial information compilations are investigated. Further analysis is then conducted to find a suitable computerized method to solve the problems and provide the framework and technique to prepare integrated financial information.

In this paper, the research investigates the possible problems arising from compiling financial information, as well as classifying several categories of financial information compilation problems, and followed by analyzing possible problems related to individual financial information computerization. A questionnaire survey is then designed based on the problems proposed in order to conduct a field study and on-site interviews. The study intends to provide the real world conditions to be better understood and the novel ideas to be explored to outreach. The experts at the
interviews with the ERP software companies include the CEO, senior engineers, business consultants, and chief accountants who provide the comprehensive insight on the enterprise’s requirements for the computer-aided integrated financial information compilation systems. Based on the field study and on-site interview results, the computer-aided integrated financial information compilation system framework and technique is proposed. The methods under each framework are expounded, and the advantages and disadvantages and future development trends are analyzed finally.

The paper is organized into four sections. The first section introduces the research motivation, research issue, and research approach. The second section discusses the literature review and the features and functions of XBRL language. In addition, the problems related to integrated financial information compilations are analyzed. In the third section, the research method, the design of the field study and interview, and the interview results are presented. Finally, in the fourth section, the problems of the enterprise’s integrated financial information compilations are analyzed. A set of computer-aided XBRL-centric techniques are proposed.

2. Literature Review

2.1 eXtensible Business Reporting Language

American Institute of Certified Public Accountants (AICPA), Big Five, Information Technology Companies, and other organizations began promoting XBRL (eXtensible Business Reporting Language) since 1998. AICPA has also pointed out that XBRL-oriented electronically-based financial reporting will be the main language used in future financial statements. Currently, many countries around the world have already adopted or are preparing to adopt XBRL as the main method to express financial statements. XBRL International further reported that XBRL might become the common language used in all business reports on the Internet (Bovee, 2002; Debreceny and Gray, 2001).

Based on the well-developed XML technique, XBRL provides enterprises a way to supply financial information without changing the existing accounting standards (Hannon, 2002). Following the advantages of reporting business information in XBRL format:

1. Accounting information is reported based on the standard taxonomy among enterprises; thus, the problem of dissimilar accounting accounts among enterprises is resolved. Users can easily understand the reported accounting information.

2. XBRL not only processes numerical data, it also processes text data. Therefore, a
standard XBRL document not only provides conventional financial statement information, it also includes non-financial, operational, and diversified information that is different in nature. Reporting financial information in the XBRL format could enhance the information transparency of enterprises (Glen, 2008).

3. The information contained in the XBRL documents can be downloaded directly by computers. Therefore, the repeated manual data entering works can be spared. As a result, reporting the financial information in the XBRL format can greatly eliminate inefficient processes of data inputs and errors from manual inputs.

The XBRL technique released by XBRL International is adopted worldwide. According to XBRL specifications, the framework of three development stages must be abided by when preparing a standard XBRL document.

(1) Setup the XBRL specification. The specifications are set up by XBRL International. XBRL specification is the basis of all XBRL documents, and it defines the form and hierarchical structure of XBRL taxonomies and XBRL instance documents. All taxonomies and XBRL instance documents should be set up in compliance with the specifications. With the specifications, the elements and data types such as duration, share type, and currency type can be pointed toward the accounts. The specifications also can define the link base elements and special elements that link toward the respective accounting terms. For example, the subsidiary department’s elements may be used to express the financial conditions of the subsidiary departments. It is noticed that the specifications are only set up by XBRL International. Nations select the specification version to abide by at the time of setting up the taxonomies or producing the instance documents. In this way, the XBRL labels are internationally unified and easy to interflow.

(2) Setup the taxonomy. The XBRL taxonomy is set up by the accounting authorities in different countries. The authorities set up the XBRL taxonomies in accordance with the XBRL specifications and the generally accepted accounting principles in the country. Taxonomy is similar to the chart of accounts of financial statements. It provides a standard set of accounts name and attribute to compile the XBRL instance documents. The taxonomy is not set up by XBRL International; instead, it is set up by accounting authorities in different countries because the GAAP adopted is not quite the same between the countries. As the accounting principles vary, the name and classification of accounts also vary. Therefore, the authorities must set up a proper taxonomy depending on the local accounting principles.

(3) Compiling a XBRL instance document. Firms may compile the XBRL instance
documents comply with the standard specification set by the XBRL International and the taxonomies set by the authority of the country.

2.2 Integrated Financial Information

Business groups diversify their layouts through mergers and investments. However, the complex investments make the financial information of the firm less transparent and make it difficult to see through the actual financial information of the enterprise for investors, creditors, and authorities. Moreover, under the global layout, firms are relatively more difficult to make overall operating strategies. In 2001, the outbreak of the Enron Case in the U.S.A. drew concerns over the importance of integrated financial information.

The concept of the integrated financial information is to break the legal confines of firms and provide the overall financial conditions of the business group. All of the affiliate companies are viewed as a group like a single financial entity. Report users can therefore clearly understand the overall financial conditions, business performance, and changes in the financial conditions of the financial entity. Based on the concept of integrated financial information, the internal transactions (such as the long-term investment, investment income or loss, purchase and sales, credits and debts, accounts receivable/payable, revenue and expense) between the parent company and the subsidiary companies must be written off. Therefore, the actions taken to improve the appearance of the financial report would be uncovered and firms can not window the real financial conditions of the business group through internal transactions. With the consolidated financial report, the report users can have a better ‘view’ to see through the overall business conditions and in turn make correct decisions. Authorities also can detect the firm which is in violation of laws or regulations with consolidated financial reports.

Currently, countries such as the U.S.A., Singapore, and Hong Kong have long demanded companies to disclose their quarterly consolidated financial reports. The financial reports of individual companies merely provide information that complements the consolidated financial reports. On the other hand, many countries around the world are also headed in the same direction. The compilation of consolidated financial reports has become a routine work for companies. Thus, the greatest impact and challenge for companies lie in how to compile consolidated financial reports accurately in a given short period of time by law and further compile the integrated financial information for internal management requirements.

Therefore, the computer-aided integrated financial information system is
considered an important tool for enterprises at present. It not only helps companies cut down the labor requirement for integrated financial information, but also greatly reduce the errors in manual compile works. Most important of all, through the computer-aided consolidated financial report system, companies can compile accurate consolidated financial reports in the given short period of time. Moreover, through flexible system designs, various types of management used reports can be prepared. The computer-aided integrated financial information system not only effectively and accurately provides the consolidated financial reports, it also helps the company cut down the reporting costs and obtain up-to-date management information (Pinsker and Li, 2008).

2.3 The Related Problems

In this view, possible problems encountered in compiling consolidated financial reports are examined in this paper. The problems related to the integrated financial information are divided into eight types. Each problem below is analyzed based on the problems related to the compilation integrated financial information and the computerization of it.

1. Problems related to accounts mapping. The accounting accounts used by integrated financial information entities may be different. The accounts should be mapped in the process of consolidated financial reports compilation. At present, the inconsistent accounts are mapped manually. In case when the computer-aided integrated financial information compilation system is to be used, the accounts of the companies must first be standardized. Then, the accounts of the affiliate companies can be converted into a consistent basis before accurate consolidated financial reports are compiled. With XBRL technique, the technique of mapping can map the accounts of different enterprises to the pre-set XBRL standard taxonomy package of consolidated financial reports. Therefore, the consolidated financial reports can be compiled automatically (Li, 2003).

2. Problems related to internal transaction write-off. In order to truthfully reflect the business conditions of a company, the internal transactions of the company must be write-off in accordance with GAAP. However, internal transactions inside a business group are often complex. In compiling consolidated financial reports, much attention has to pay to the internal transaction write-offs and the recognition of incomes (losses). Recently, the computations of internal transaction and write-off are made manually. Also, the internal transaction cut-off time and recording time are judged by accountants. If the computer-aided integrated financial information system is to be used, the user must first enter the
internal transaction data that need to be written-off, shareholding ratio between
the parent and subsidiary companies, and the changes of shareholding ratio. Since
internal transaction write-offs involve many professional judgments, it cannot be
replaced by the computer-aided integrated financial information system. Li (2003)
established a ‘standard formula’ to build the taxonomies and implement the
prototype system. The features of the standard structure of the XBRL taxonomy
facilitate the representation of the standard equations for compiling consolidated
financial reports. Since the elements are named uniform, the amounts of the
consolidated financial report accounts can be automatically computed by
XBRL-formula. However, the XBRL is used mainly to provide a uniform format
for financial statement presentations; the calculation functions are limited to
simple arithmetic calculations. Therefore, it is suggested to carry out the amount
calculations outside the XBRL technique and use the calculation function of
XBRL on the auto-articulation.

(3) Problems related to the changes of shareholding ratio. The shares held by parent
company may change often. The investment income and minority interest cannot
be correctly recognized unless details on shareholding period and shareholding
change data are obtained. Currently, investment income and minority interest
under manual operations rely on accountants. It is a repetitive work and the
conditions of the shareholding changes are so complex that errors often occur,
and may damage the accuracy of consolidated financial reports. When the
computer-aided integrated financial information system is used, once the user
supplies the shareholding period and shareholding change data, the system
automatically compute the correct investment income and the minority interest
that should be recognized in accordance with the accounting principles.

(4) Problems related to the accounting policy inconsistency. According to the
regulations in GAAP, accounting information contained in the integrated
financial information provided by the subsidiary companies has to conform to the
processing accounting method of the parent company. The subsidiary companies
that adopt different accounting policies are therefore should convert their
financial statements into accounting policies adopted by the parent company.
This makes integrated financial information much more difficult. If
computer-aided integrated financial information system is used, the original
transaction records must be stored in the computer system. With the transaction
data, the system can automatically compute the amounts that should be reported
in individual firms’ financial report and in consolidated financial report based on
the different accounting principles. The financial reports are converted by XBRL
automatic calculation in accordance with the rules.

(5) Problems related to lack of flexible adjustment of integrated financial information framework. Other than consolidated financial reports regulated by law, an enterprise also needs to compile consolidated financial reports for management purposes. Through integrated financial information of different consolidated hierarchies and different levels, the enterprises are able to make management decisions more efficiently. Recently, consolidated financial reports of different hierarchies and different levels are also compiled manually. It is a time and labor consuming operation. Through a flexible system design, the computer-aided system can access information in the database and compile all sorts of consolidated financial reports through different hierarchy and different level settings. The enterprises therefore obtain various types of consolidated financial reports required by law or for management use with the computer-aided system.

(6) Problems related to foreign currency conversion. Integrated financial information entities may use different currency units. All foreign currencies must first be converted into functional currencies before consolidating. Presently, accountants use Excel in compilations and they have to enter the exchange rates that are corresponding to different subsidiaries. With the computer-aided system, once the exchange table set up the amounts would convert automatically. There is no need for exchange rate entries each time or manual computations. As a result, repetitive works can be greatly reduced.

(7) Problems related to disclosure information collection. Disclosure information normally is not found in the ERP systems. Therefore, disclosure information of the parent company and subsidiary companies cannot be collected systematically (Callaghan et al., 2008). At present, integrated financial information compilations separately collect disclosure information manually. In order to compile consolidated financial reports by computer-aided system, a uniform disclosure information collection platform must first be set up. In addition, the proper disclosure format in consolidated financial reports must be set up. The system then can automatically produce the disclosure information of consolidated financial reports based on the information collected.

(8) Problems related to integrated financial information audit. For conventional manual compilers, automatic financial report compilation contains little audit trails (Coppers and Lybrand, 2002). This makes it more difficult for auditors during audits. At present, auditors normally perform computerized accounting information audit through the 3 methods below: 1. Auditors may test system
validity by means of pre-designed constructive data; 2. Auditors may randomly inspect a certain transaction in the system; and 3. Auditors may use the pre-designed ERP system in inspecting whether the ERP system results are correct (Sirikulvadhana, 2002). In terms of consolidated financial reports, since the financial statements of individual companies are duly audited and certified by CPA, the auditing of financial statement information of the said companies are not included as part of the discussion scope. In view of integrated financial information auditors, the focus lies in whether enterprises compile consolidated financial reports in accordance with GAAP. Therefore, the integrated financial information compilations through the computer-aided system must be able to create the same documents which are provide when manually compiling in order to provide auditors with necessary audit trails. Chen (2003) set up XAUL, eXtensible Auditing Language to confirm the system framework of automatic audit, and develop a data exchange and analysis platform between the CPA and the client.

<table>
<thead>
<tr>
<th>Table 1 Main problems in compiling integrated financial information.</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Problems related to accounts mapping.</td>
</tr>
<tr>
<td>(2) Problems related to internal transaction write-off.</td>
</tr>
<tr>
<td>(3) Problems related to the changes of shareholding ratio.</td>
</tr>
<tr>
<td>(4) Problems related to the accounting policy inconsistency.</td>
</tr>
<tr>
<td>(5) Problems related to lack of flexible adjustment of integrated financial information framework.</td>
</tr>
<tr>
<td>(6) Problems related to foreign currency conversion.</td>
</tr>
<tr>
<td>(7) Problems related to disclosure information</td>
</tr>
</tbody>
</table>
collection. systematically (Callaghan et al., 2008). The information presently is collected manually.

(8) Problems related to integrated financial information audit. The automatic financial report compilation contains relatively little audit trails (Coppers and Lybrand, 2002). This makes it more difficult for auditors during audits.

<table>
<thead>
<tr>
<th>3. Research Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1 The Field Study</td>
</tr>
</tbody>
</table>

In this section, the study presents the research approach it adopts to address the research issue. Based on the literature review and compiled categories of issues, the study develops a survey questionnaire to conduct a field study and on-site interviews with ERP software companies to gather and analyze the existing approaches to prepare the integrated financial information. The questionnaire is based on the problems we survey in the literature review. Each represents one main category of problems the enterprises face in consolidating financial reports. We analyze the interview results and survey documents collected to develop three main computer-aided approaches to resolve the inadequacies and difficulties in the conventional, ad hoc, subjective, manual, and error-prone approach. Detailed steps are presented to illustrate the positive feasibility and validity of these XBRL-centric approaches.

First the study analyzes the reasons of each problem category based on the field study.

(1) Problems related to accounts mapping. The interview is conducted to find out the problem of incompatible ERP systems. The interview results show that companies are often inclined toward adopting uniform accounts. The interviewed companies expressed that in terms of the present compatible ERP systems design, when loading different company accounts, new site ids are given so that the site ids of all accounts differ. When compiling reports, the account of each report column is first defined before performing grouping. In other words, the account mapping action takes place only during report compilation. However, the present technology is not applicable with incompatible ERP systems.

(2) Problems related to internal transaction write-off. In order to truthfully express the business conditions of a corporation as a whole, the internal transactions inside the business group must be written off based on the GAAP. The interviewed companies expressed that this problem has not been resolved in integrated financial information compilations. One possible way to solve the
problem is to set up a virtual body and record the processes of the integrated financial information write-offs. The advantage of doing so is that the original account records of the companies would not be altered and write-off records can be kept. However, the problems faced by the interviewed companies are that internal transaction write-offs involve professional judgments that the system cannot deal with. It is a bottleneck during system designs. In view of this problem, it is suggested that since professional judgments in accounting cannot be substituted by the system, forms and lists may be designed and system users may be required to provide information needed in making professional judgments in advance. So that the system only needs to assist in systemic calculations outside the confines of professional judgments.

(3) Problems related to the changes of shareholding ratio. The shareholding of the parent company may change often. When compiling consolidated financial reports, the investment income and minority interest cannot be correctly recognized unless details on shareholding period and shareholding change data are obtained. The interviewed companies expressed that currently the problem has not been solved in integrated financial information compilations. It is suggested that the shareholding change data should be stored in ERP databases; it allows the ERP system to automatically compute the amounts of investment income and minority interest.

(4) Problems related to the accounting policy inconsistency. According to regulations in GAAP, accounting information contained in the integrated financial information provided by the subsidiary companies has to conform to the accounting method of the parent company. The interviewed companies expressed that currently the problem has not been solved in consolidated financial account compilation. One possible way to solve the problem is to set up a virtual body to re-compute and record the processes of integrated financial information write-offs. The advantage of doing so is that the original account records of the companies would not be altered and write-off records can be kept. However, the interviewed companies expressed that if the system is to automatically convert the amounts of different accounting standards, the problem would turn to the acquisition of original transaction data.

The XBRL technique can be utilized in the setup of taxonomies and the description of conversion logics of accounting methods. Then, conversions are assisted with the use of the calculation function of the integrated financial information system. However, this method cannot fundamentally solve the problem of raw data collection. In addition, the XBRL GL technique is also
suggested. By using the XBRL GL technique, the enterprises’ transaction data about general ledger and transaction entries is stored and the problem of raw data collection also solved. However, since XBRL GL is a newly developed technique, it may change in the future, and this risk must be considered by enterprises.

(5) Problems related to lack of flexible adjustment of integrated financial information framework. Other than consolidated financial reports regulated by law, an enterprise also needs to compile consolidated financial reports for management purposes. Currently, the interviewed companies first define the different company groups for the ERP system to provide consolidated financial reports and management reports of the different target groups.

(6) Problems related to foreign currency conversion. There may be multiple foreign currency units in the integrated financial information entities. All the foreign currencies must first be converted into functional currencies before consolidation. Currently, the interviewed companies expressed that the problem has not been resolved in integrated financial information compilation.

One possible way to solve the problem is to set up the data of exchange rate to allow the ERP systems calculate automatically. However, the interviewed companies expressed that the system is unable to make judgments on the functional currencies. Moreover, the accounting principles for foreign currency conversions are complex and are considered the greatest limitation in the company’s program design. It is suggested in this paper that since functional currency related judgments involve numerous non-quantified data, which the system is unable to produce by logical rules, it is suggested that the functional currency judgments be determined by user settings. To perfectly complying with the accounting principle of foreign currency conversion is a crucial part of foreign currency conversion. If the exchange rate is erroneous, the consolidated financial reports cannot be presented appropriately. This problem also exists in manual operations. It is suggested that the regulations in the accounting principles should first be clarified and the logic rules of the foreign currency conversion are defined in the rule base to facilitate the conversion processes.

(7) Problems related to disclosure information collection. Disclosure information is normally not found in the ERP systems. Therefore, when compiling consolidated financial reports, all disclosure information of the parent company and subsidiary companies cannot be collected systematically. The interviewed companies expressed that currently, the problem has not been solved in integrated financial information compilation. At present, disclosure information is still processed
manually.

It is suggested to structuralize the disclosure information in this research. For instance, when preparing individual financial reports in the ERP system, N numbers of frequently disclosed disclosure information may be designed. The user first selects the disclosure item information and the system then produces uniform disclosures. The structuralized disclosures can be used to produce the disclosure information of consolidated financial reports. The greatest advantage of XBRL lies in its ability to process text data. Therefore, in view of companies that have already adopted XBRL, it is suggested in this paper that a standard taxonomy for disclosure information be set up so that individual companies may produce uniform XBRL formatted disclosure information based on the standard taxonomy. Subsequently, XBRL FR integrated financial information would read the XBRL formatted disclosure information of the individual companies. The feature of structuralized XBRL would automatically search the disclosure information required by consolidated financial reports. Meanwhile, through the disclosure format of the integrated financial information standard, the data column reference of each item would be pre-set to systematically produce the disclosure information of consolidated financial reports. The setup of disclosure taxonomy may be done in the same way as the aforementioned method to set up numbers of frequently disclosed items. Furthermore, additional classifications outside the confines of the standards are also included for more flexible use on the part of the user.

(8) Problems related to integrated financial information audit. For conventional manual compilers, automatic financial report compilation contains little audit trails. This makes it more difficult for auditors during auditing processes. The interviewed companies expressed that its system records all compilation processes to create intermediate data documents. They are provided to auditors in auditing processes. Moreover, the auditors may utilize constructive data to test the system validity. When compiling computer-aided consolidated financial reports, the intermediate data can also be recorded to create relevant documents for examination by other auditors.

3.2 Mapping Table Approach

The conventional method of the account mapping table is applicable to business groups with compatible with the ERP systems and to companies that have not yet adopted XBRL technology. The structure and main components of the conventional method of the account mapping table is as shown in Fig. 3.1.
Under this method, the ERP systems of the companies must be compatible. For compiling integrated financial information, the parent company should load data from the ERP systems of the subsidiary companies into the ERP system of the parent company. In order to keep complete integrated financial information compilation processes and ensure that the raw data of the parent company would not be affected by the compilation of consolidated financial reports, the ERP system of the parent company would create a virtual body. All processes of the compilation of consolidated financial reports would be recorded in this virtual body. Once the data required for compiling consolidated financial reports is loaded, account names used in the ERP systems may differ. In the mean time, the ERP system would conduct mapping in view of accounts and save all data required for compiling consolidated financial reports in database of virtual body. In the ERP system, the presentation method, column requirements for consolidated financial reports, and reference data of each column should be defined in advance. The ERP system would automatically retrieve related data from the ERP database of integrated financial information compilations, conduct logical calculations, and resend the results to the designated columns. This way, consolidated financial reports would be automatically created from the ERP systems of the enterprises.

The advantage of utilizing the conventional method of the Account Mapping Table is that enterprises can bring in one additional compilation function of integrated financial information compilation under the present ERP system structure. There is no need to bring in new software or technology, or to conduct personnel re-training. Through the virtual body compilation method, in addition of ensuring the data of the parent company not to be affected by the compilation of consolidated financial reports, it also completely records all the processes of the integrated financial information compilation. Meanwhile, intermediate reports under manual operations can be created automatically from the system and keep complete audit trails. With computer-aided integrated financial information compilations, enterprises are spared from large manual inputs and calculations. In addition to more efficient integrated financial information compilations, possible errors from manual operations can also be reduced.
The efficiency and accuracy of integrated financial information compilations are therefore enhanced. However, this method is only applicable for companies with compatible ERP systems. This method is not applicable when the ERP systems in the integrated financial information compilation bodies of enterprise are incompatible. On the other hand, since shareholding investment transactions are frequent in large companies, the bodies of integrated financial information compilations are likely to be subject to changes. It is difficult to ensure all ERP systems adopted by the integrated financial information compilation bodies are compatible. It is essential that enterprises take this aspect into consideration before adopting the method.

3.3 XBRL FR-Based Approach

The XBRL FR-based integrated financial information compilation method is applicable to companies with different ERP systems and enterprises with XBRL technology. The structure and main components of the XBRL FR-based integrated financial information compilation method is as shown in Fig. 3.2.

![Fig.3.2 The structure of XBRL FR-based method](image)

Individual companies with XBRL technology create individual XBRL formatted financial statements from their ERP systems. Since XBRL formatted financial statements provisioned by individual companies are all formats presented under XBRL taxonomy, they are standardized and not affected by the ERP systems or different account names of the individual companies. With this method, an XBRL integrated financial information is provided. This system is capable of reading XBRL instance documents of the individual companies and storing the data in the databases. In the system, the presentation method, column requirements, and reference data of each column should be pre-defined in the system. The system would automatically retrieve related data automatically from the database and conduct logical calculations. The results are then presented and stored in designated columns to create consolidated financial reports. The XBRL FR-based integrated financial information comes with the function of storing consolidated financial reports in standard XBRL format. Enterprises that have adopted XBRL technique may create XBRL formatted consolidated financial reports from the system. There is no need to create them
through other software systems.

The advantage of utilizing the XBRL FR-based integrated financial information compilation method is that enterprises with XBRL technology can directly read financial statements of other companies and complete integrated financial information compilations to ultimately create XBRL formatted consolidated financial reports. For parent companies of integrated financial information compilations, the existing ERP system need not be changed. The problem of incompatibility among ERP systems of integrated financial information compilation bodies also does not exist. The XBRL FR-based integrated financial information system is independent from the ERP systems. It can ensure that the ERP systems of each integrated financial information compilation body would not be affected. The complete process of integrated financial information compilations can also be recorded in the system. The intermediate reports under manual operations can be created automatically from the system to keep complete audit trails. With computer-aided integrated financial information compilations, enterprises are spared from large manual inputs and calculations. In addition to more efficient integrated financial information compilations, possible errors from manual operations can also be reduced. The efficiency and accuracy of integrated financial information compilations are also enhanced. However, one limitation of this method is that XBRL FR instance contains no transaction entries, ledger, and general ledger data. User interventions are required for data requirements of consolidated financial reports.

3.4 XBRL GL-Based Approach

The XBRL GL-based integrated financial information compilation method is applicable to business groups with different ERP systems and enterprises with XBRL GL technology. XBRL GL is the standard proposed by XBRL International in November, 2005. XBRL GL differs from the previously mentioned XBRL FR design. It directly labels general transaction data and ledgers of accounts. XBRL GL extends XBRL technique to the applications of the enterprise interior in order to ensure financial statements extend to general ledger and transaction ledger levels (Doolin and Troshani, 2004). There are two ways to compile consolidated financial reports using XBRL GL technique.
First, as shown in Fig. 3.3, XBRL GL and ERP of an enterprise are integrated. After each entity of the consolidated financial reports creates XBRL GL, the ERP system of the parent company is loaded to XBRL GL of the subsidiary companies. The ERP system of the parent company would conduct mapping of accounts and produce adjustment and write-off entries to complete the compilation of consolidated financial reports. Under this method, the problem of incompatibility in the ERP systems of enterprises could be solved by XBRL GL. Meanwhile, general ledger and transaction entry data in the ERP systems can be exchanged by XBRL GL.

Second, the XBRL GL-based integrated financial information compilation system is used to aid integrated financial information compilations. After XBRL GL is created by each financial information body, it would be directly loaded into the XBRL GL-based integrated financial information compilation system. The system would subsequently conduct account mapping and make necessary adjustments needed in integrated financial information compilations and write-off entries to complete the integrated financial information compilations. The advantage of
compiling integrated financial information with XBRL GL-based integrated financial information compilation system is that since the system is independent from the ERP systems, the data is not affected during the compilation processes. The complete integrated financial information compilation processes can also be recorded in the system. All intermediate reports under manual operations can be created automatically from the system to keep complete audit trails. Moreover, to an enterprise that has adopted XBRL GL, the XBRL GL-based integrated financial information compilation system not only assists the enterprise to compile consolidated financial reports, it also compiles financial statements for individual companies. All the enterprise needs to undertake all financial statement compilations is one single system.

Note that XBRL GL assists in integrated financial information compilations whether integrated with the ERP systems or XBRL GL-based integrated financial information consolidated system. Accounts mapping and the ratio of related internal transaction write-offs requiring user interventions as professional judgments are also needed.

4. Summary

In this paper, the study has presented three computer-aided approaches to prepare the integrated financial information. They are the conventional mapping table method, the XBRL FR-based method, and the XBRL GL-based method. Each of them has its own advantages and disadvantages. Two of them are XBRL-centric to provide the more adequate, systematic, and methodological approach to tackle the long-standing issue of ad hoc, subjective, error-prone, time-consuming, and manual handling of integrated financial information. In this research, the study has compiled the main problems of consolidating financial reporting from the current literature. The study has categorized them into eight types of issues. Each type has been thoroughly discussed through the in-depth field study and on-site interviews with the ERP software companies. Furthermore, the study has developed two XBRL-centric methods that can be integrated with the ERP systems in enterprises. Detailed steps of each are described to illustrate the feasibility and validity. Future experiments will be conducted to further demonstrate the positive effects.
References


Huang J.Y., “Using the Calculable XBRL for Accounting Standard Transformation”,

23


ORGANIZING COMMITTEE

Prof. Shimin CHEN
Prof. Xinyuan CHEN
Dr Qinglu JIN
Prof. Zengquan LI
Prof. Judy S.L. TSUI
Dr Donghui WU
Prof. Lijun XIA
Prof. Wayne W. YU

China Europe International Business School
Shanghai University of Finance and Economics
Shanghai University of Finance and Economics
The Hong Kong Polytechnic University
The Hong Kong Polytechnic University
Shanghai University of Finance and Economics
The Hong Kong Polytechnic University

(In alphabetical order by the last name)
FEATURED PRESENTATIONS

INVITED SPEAKER PROFILES
(by order of presentation)

**Ferdinand A. GUL**
Monash University Sunway Campus  The Hong Kong Polytechnic University

Ferdinand A. Gul is Professor of Accounting and Corporate Governance, Head of the Accounting and Finance Department, and the Director of Research at School of Business, Monash University, Sunway Campus, Malaysia. He is also the Chair Professor of Accounting and Corporate Governance and the Head of the School of Accounting and Finance at The Hong Kong Polytechnic University. Currently, he is on leave from PolyU. Professor Gul has wide practical and academic experience having held several senior positions in various universities in Australia and Hong Kong, and has provided consulting services on corporate governance reform to the Government of the Hong Kong Special Administrative Region. He held various visiting/honorary professorships in various universities in China, such as the Shanghai University of Finance and Economics, Sun Yat-sen University, and Xiamen University, and has published in journals including the *Journal of Accounting and Economics*, *Journal of Accounting Research*, *The Accounting Review*, and *The Journal of Financial Economics*. His research interests are in the inter-disciplinary application of contracting and agency theories to auditing, corporate finance, financial accounting and corporate governance as it relates to the capital market.

**Judy S. L. TSUI**
The Hong Kong Polytechnic University

Judy S. L. Tsui is Vice President (International and Executive Education), Director of Graduate School of Business and Chair Professor of Accounting of The Hong Kong Polytechnic University. She holds positions as Honorary Professor and Visiting Professor at several top mainland universities, and is the first professor in accounting to be awarded the Cheung Kong Chair Professorship by China’s Ministry of Education. Professor Tsui has also been appointed Visiting Scholar at MIT’s Sloan School of Management in the US. She has served on the Council of PolyU (2005, 2007-2010) and the Council of the Hong Kong Institute of Certified Public Accountants (2000-2004). Appointed by the HKSAR Government, Professor Tsui is currently a member of the University Grants Committee, the Research Grants Council, and the Hong Kong Committee for Pacific Economic Cooperation, and is an Independent Non-executive Director of CLP Holdings Limited and China Vanke Company Limited in Hong Kong and the mainland, respectively.

徐铭丽
香港理工大学

徐铭丽为香港理工大学副校长（国际事务及行政教育）暨工商管理研究院院长及会计学讲座教授。现任内地多间著名大学的荣誉教授及客座教授，同时是首位会计学教授获中国教育部聘为长江讲座教授，并曾出任美国麻省理工学院Sloan管理学院的访问学者。徐教授为香港理工大学校董会成员（2005年，2007-2010年），并曾出任香港会计师公会理事（2000-2004年）。徐教授获香港特区政府委任为大学教育资助委员会、研究资助局及太平洋经济合作香港委员会的成员。现任香港中电控股有限公司及内地方企业股份有限公司的独立非执行董事。
Oliver Zhen Li is an Associate Professor in Accountancy at the Eller College of Management, University of Arizona. Before joining Arizona, he taught at the University of Notre Dame. His research interests are in taxation, corporate finance and the capital markets. He is also interested in China-related research and has been working on a few China projects in recent years. Professor Li's research papers have been published in top journals, such as The Accounting Review, Journal of Accounting and Economics, Journal of Accounting Research, Contemporary Accounting Research, Review of Accounting Studies, Journal of Finance, and National Tax Journal. He is the president-elect of the Chinese Accounting Professors’ Association of North America for 2010-2011. Professor Li enjoys reading, hiking, travelling, and making friends.
<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:00 a.m.—</td>
<td><strong>Registration</strong></td>
<td><strong>会计学院109室</strong> Room 109, School of Accountancy</td>
</tr>
<tr>
<td>8:30 a.m.—</td>
<td>** Welcoming and Opening Addresses:**</td>
<td></td>
</tr>
<tr>
<td>9:00 a.m.—</td>
<td><strong>Featured Presentation I</strong></td>
<td><strong>会计学院109室</strong> Room 109, School of Accountancy</td>
</tr>
<tr>
<td>9:45 a.m.—</td>
<td><strong>Plenary Session I</strong></td>
<td><strong>会计学院109室</strong> Room 109, School of Accountancy</td>
</tr>
<tr>
<td>10:55 a.m.—</td>
<td><strong>Coffee Break</strong></td>
<td></td>
</tr>
<tr>
<td>11:10 a.m.—</td>
<td><strong>Plenary Session II</strong></td>
<td><strong>会计学院109室</strong> Room 109, School of Accountancy</td>
</tr>
</tbody>
</table>

**Programme**

**China Accounting and Finance Review International Symposium 2010**

**Day 1: 4 December 2010 (Saturday)**

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:00 a.m.—</td>
<td><strong>Registration</strong></td>
<td><strong>会计学院109室</strong> Room 109, School of Accountancy</td>
</tr>
<tr>
<td>8:30 a.m.—</td>
<td>** Welcoming and Opening Addresses:**</td>
<td></td>
</tr>
<tr>
<td>9:00 a.m.—</td>
<td><strong>Featured Presentation I</strong></td>
<td><strong>会计学院109室</strong> Room 109, School of Accountancy</td>
</tr>
<tr>
<td>9:45 a.m.—</td>
<td><strong>Plenary Session I</strong></td>
<td><strong>会计学院109室</strong> Room 109, School of Accountancy</td>
</tr>
<tr>
<td>10:55 a.m.—</td>
<td><strong>Coffee Break</strong></td>
<td></td>
</tr>
<tr>
<td>11:10 a.m.—</td>
<td><strong>Plenary Session II</strong></td>
<td><strong>会计学院109室</strong> Room 109, School of Accountancy</td>
</tr>
</tbody>
</table>
### Day 1: 4 December 2010 (Saturday)

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>12:20 p.m.--2:00 p.m.</td>
<td>拍照留念及午膳 Photo Taking and Lunch</td>
<td>上海財大霍建大酒店Howard Johnson Caidaplaza Shanghai</td>
</tr>
<tr>
<td>2:00 p.m.--3:10 p.m.</td>
<td>全会 III Plenary Session III</td>
<td>会计学院107室Room 107, School of Accountancy</td>
</tr>
<tr>
<td></td>
<td>主持 Chairperson: 苏锡嘉教授 中欧国际工商学院  Prof. Xijia SU China Europe International Business School</td>
<td></td>
</tr>
<tr>
<td></td>
<td>论文题目 Title: 媒体为什么会报道上市公司丑闻？</td>
<td></td>
</tr>
<tr>
<td></td>
<td>宣读者 Presenter: 刘德明 北京大学</td>
<td></td>
</tr>
<tr>
<td></td>
<td>讨论人 Discussant: 张田余博士 香港城市大学</td>
<td></td>
</tr>
<tr>
<td>3:10 p.m.--3:25 p.m.</td>
<td>小休 Coffee Break</td>
<td></td>
</tr>
<tr>
<td>3:25 p.m.--4:35 p.m.</td>
<td>全会 IV Plenary Session IV</td>
<td>会计学院107室Room 107, School of Accountancy</td>
</tr>
<tr>
<td></td>
<td>主持 Chairperson: 苏锡嘉教授 中欧国际工商学院  Prof. Xijia SU China Europe International Business School</td>
<td></td>
</tr>
<tr>
<td></td>
<td>论文题目 Title: 发展股权、隶属压力与分析师乐观性</td>
<td></td>
</tr>
<tr>
<td></td>
<td>宣读者 Presenter: 曹胜 同济大学</td>
<td></td>
</tr>
<tr>
<td></td>
<td>讨论人 Discussant: 颜朝阳博士 明尼苏达大学</td>
<td></td>
</tr>
<tr>
<td>4:35 p.m.--5:05 p.m.</td>
<td>特邀演讲 II Featured Presentation II</td>
<td>会计学院107室Room 107, School of Accountancy</td>
</tr>
<tr>
<td></td>
<td>主持 Chairperson: 苏锡嘉教授 中欧国际工商学院  Prof. Xijia SU China Europe International Business School</td>
<td></td>
</tr>
<tr>
<td></td>
<td>题目 Topic: 如何在中国顶级期刊发表研究</td>
<td></td>
</tr>
<tr>
<td></td>
<td>演讲者 Speaker: 徐林倩丽教授 香港理工大学  Prof. Judy S.L. TSUI The Hong Kong Polytechnic University</td>
<td></td>
</tr>
<tr>
<td>5:05 p.m.--6:05 p.m.</td>
<td>特邀演讲 III Featured Presentation III</td>
<td>会计学院107室Room 107, School of Accountancy</td>
</tr>
<tr>
<td></td>
<td>主持 Chairperson: 苏锡嘉教授 中欧国际工商学院  Prof. Xijia SU China Europe International Business School</td>
<td></td>
</tr>
<tr>
<td></td>
<td>题目 Topic: 如何在中国顶级期刊发表研究</td>
<td></td>
</tr>
<tr>
<td></td>
<td>演讲者 Speaker: 柯滨教授 南洋理工大学  Prof. Bin KE Nanyang Technology University</td>
<td></td>
</tr>
<tr>
<td>6:30 p.m.--8:30 p.m.</td>
<td>晚餐 Dinner</td>
<td>丰收日宁波海鲜大酒店Harvest Festival Chain Restaurant Shanghai</td>
</tr>
</tbody>
</table>
### 第二天：2010年12月5日（星期日） Day 2: 5 December 2010 (Sunday)

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:30 a.m.—</td>
<td>Ⅴ Plenary Session V</td>
<td>会计学院109室（会计学院109室）</td>
</tr>
<tr>
<td>9:40 a.m.</td>
<td>主持 Chairperson: 陈世敏教授  中欧国际工商学院 Prof. Shimin CHEN China Europe International Business School</td>
<td>Room 109, School of Accountancy</td>
</tr>
<tr>
<td></td>
<td>论文题目 Title: Should Earnings Guidance Be Regulated? Evidence from an Earnings-Based Regulatory Regime</td>
<td></td>
</tr>
<tr>
<td></td>
<td>宣读者 Presenter: Peng CHENG Xijiao Liverpool University</td>
<td></td>
</tr>
<tr>
<td></td>
<td>评论人 Discussant: Dr Oliver Zhen LI University of Arizona</td>
<td></td>
</tr>
<tr>
<td>9:40 a.m.—</td>
<td>Ⅵ Plenary Session VI</td>
<td>会计学院109室（会计学院109室）</td>
</tr>
<tr>
<td>10:50 a.m.</td>
<td>主持 Chairperson: 陈世敏教授  中欧国际工商学院 Prof. Shimin CHEN China Europe International Business School</td>
<td>Room 109, School of Accountancy</td>
</tr>
<tr>
<td></td>
<td>论文题目 Title: Political Participation and Entrepreneurial Initial Public Offerings in China</td>
<td></td>
</tr>
<tr>
<td></td>
<td>宣读者 Presenter: Xunan FENG City University of Hong Kong</td>
<td></td>
</tr>
<tr>
<td></td>
<td>评论人 Discussant: Prof. Donghua CHEN Nanjing University</td>
<td></td>
</tr>
<tr>
<td>10:50 a.m.—</td>
<td>小休 Coffee Break</td>
<td></td>
</tr>
<tr>
<td>11:05 a.m.—</td>
<td>Ⅳ Featured Presentation IV</td>
<td>会计学院109室（会计学院109室）</td>
</tr>
<tr>
<td>12:05 p.m.</td>
<td>特邀演讲 IV  Doing China Research: Some Personal Experience</td>
<td>Room 109, School of Accountancy</td>
</tr>
<tr>
<td></td>
<td>主持 Chairperson: 陈世敏教授  中欧国际工商学院 Prof. Shimin CHEN China Europe International Business School</td>
<td></td>
</tr>
<tr>
<td></td>
<td>演讲者 Speaker: 顾朝阳博士 明尼苏达大学 Dr Zhaoyang GU University of Minnesota</td>
<td></td>
</tr>
<tr>
<td>12:05 p.m.—</td>
<td>午餐 Lunch</td>
<td>上海费尔蒙大酒店 Howard Johnson Caidapiazza Shanghai</td>
</tr>
<tr>
<td>2:00 p.m.—</td>
<td>Ⅰ Concurrent Paper Session I</td>
<td>会计学院109室（会计学院109室）</td>
</tr>
<tr>
<td>2:45 p.m.</td>
<td>(1) 论文题目 Title: 新兴市场中的“投资偏见”－基于中国内地股票基金持股行为的经验研究</td>
<td>Room 109, School of Accountancy</td>
</tr>
<tr>
<td></td>
<td>宣读者 Presenter: 林树 南京大学</td>
<td></td>
</tr>
<tr>
<td></td>
<td>评论人 Discussant: 陈治鸿博士 香港城市大学</td>
<td></td>
</tr>
<tr>
<td>2:45 p.m.—</td>
<td>(2) 论文题目 Title: 产业政策与公司融资－来自中国的经验证据</td>
<td></td>
</tr>
<tr>
<td>3:30 p.m.</td>
<td>宣读者 Presenter: 新春 南京大学</td>
<td></td>
</tr>
<tr>
<td></td>
<td>评论人 Discussant: 童一鸣博士 中山大学</td>
<td></td>
</tr>
<tr>
<td>3:30 p.m.—</td>
<td>(3) 论文题目 Title: 监管者变更与执法精度</td>
<td></td>
</tr>
<tr>
<td>4:15 p.m.</td>
<td>宣读者 Presenter: 蒋德权 南京大学</td>
<td></td>
</tr>
<tr>
<td></td>
<td>评论人 Discussant: 原红旗教授 复旦大学</td>
<td></td>
</tr>
</tbody>
</table>

研讨会程序如有任何更改，将于会场内张贴告示，恕不另行通知。
Any changes to the programme will be posted in the venue without prior notice.
<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>2:00 p.m.—</td>
<td>Concurrent Paper Session II</td>
<td>Room 206, School of Accountancy</td>
</tr>
<tr>
<td>2:00 p.m.—</td>
<td>分组研讨 II</td>
<td></td>
</tr>
<tr>
<td>2:045 p.m.—</td>
<td>(1) 论文题目 Title: An Investigation of the Relationship Among Types of Financial Reporting Fraud, Ownership Structure and Audit Effectiveness: Evidence from China</td>
<td></td>
</tr>
<tr>
<td>2:45 p.m.—</td>
<td>宣读者 Presenter: Yu CHEN Texas A&amp;M International University</td>
<td></td>
</tr>
<tr>
<td>3:30 p.m.—</td>
<td>(2) 论文题目 Title: 经理人市场-薪酬契约有效性与管理层侵占－基于国有企业经理人“59岁现象”的研究</td>
<td></td>
</tr>
<tr>
<td>3:30 p.m.—</td>
<td>宣读者 Presenter: 万华林 上海立信会计学院</td>
<td></td>
</tr>
<tr>
<td>4:15 p.m.—</td>
<td>(3) 论文题目 Title: 政治关系、银行业开放与银行治理</td>
<td></td>
</tr>
<tr>
<td>4:15 p.m.—</td>
<td>宣读者 Presenter: 潘红波 武汉大学</td>
<td></td>
</tr>
<tr>
<td>4:30 p.m.—</td>
<td>小休 Coffee Break</td>
<td></td>
</tr>
<tr>
<td>5:30 p.m.—</td>
<td>Featured Presentation V</td>
<td>Room 109, School of Accountancy</td>
</tr>
<tr>
<td>5:30 p.m.—</td>
<td>特邀演讲 V</td>
<td></td>
</tr>
<tr>
<td>5:30 p.m.—</td>
<td>客题 Topic: Reporting Standards, Reporting Quality and Foreign Investment: Experience from China</td>
<td></td>
</tr>
<tr>
<td>5:30 p.m.—</td>
<td>演讲者 Speaker: 李真教授 亚利桑那大学</td>
<td></td>
</tr>
<tr>
<td>5:45 p.m.—</td>
<td>闭幕式 Closing Remarks</td>
<td></td>
</tr>
<tr>
<td>6:00 p.m.—</td>
<td>晚宴 Dinner</td>
<td>Ningguofu Chinese Restaurant</td>
</tr>
</tbody>
</table>

研讨会程序如有任何更改，将于会场内张贴告示，恕不另行通知。
Any changes to the programme will be posted in the venue without prior notice.
上海财经大学武川路校区
Wuchan Campus Map
ACKNOWLEDGEMENTS

The Symposium Organizing Committee wishes to thank the following named organizations and persons and acknowledge the support of everyone concerned with the event for their unstinting assistance:

Gold Sponsor
Hong Kong Institute of Certified Public Accountants

Silver Sponsor
CPA Australia

Bronze Sponsors
Association of Chartered Certified Accountants
GTA

Review Panel Members

Prof. Shimin CHEN
China Europe International Business School

Dr Qingchuan HOU
Shanghai University of Finance and Economics

Prof. Zengquan LI
Shanghai University of Finance and Economics

Dr Donghui WU
The Hong Kong Polytechnic University

Prof. Lijun XIA
Shanghai University of Finance and Economics

Prof. Wayne W. YU
The Hong Kong Polytechnic University

(In alphabetical order by the last name)
國科會補助計畫衍生研發成果推廣資料表

<table>
<thead>
<tr>
<th>國科會補助計畫</th>
<th>計畫名稱：智慧型資訊整合於異質資料倉儲和資料探勘之模型、架構、與績效評估-應用本體論、母型圖表、和學名結構</th>
</tr>
</thead>
<tbody>
<tr>
<td>計畫主持人：</td>
<td>諶家蘭</td>
</tr>
<tr>
<td>計畫編號：</td>
<td>96-2416-H-004-017-MY3</td>
</tr>
<tr>
<td>學門領域：</td>
<td>資訊管理</td>
</tr>
</tbody>
</table>

無研發成果推廣資料
<table>
<thead>
<tr>
<th>成果項目</th>
<th>量化</th>
<th>單位</th>
<th>備註（質化說明：如數個計畫共同成果，成果列為該期刊之封面故事...）</th>
</tr>
</thead>
<tbody>
<tr>
<td>論文著作</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>論文著作</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>國內</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>參與計畫人力</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>參與計畫人力</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>國外</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>參與計畫人力</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
其他成果
(無法以量化表達之成果如辦理學術活動，獲得獎項、重要國際合作、研究成果國際影響力及其他協助產業技術發展之具體效益事項等，請以文字敘述填列。)

<table>
<thead>
<tr>
<th>成果項目</th>
<th>量化</th>
<th>名稱或內容性質簡述</th>
</tr>
</thead>
<tbody>
<tr>
<td>测驗工具 數量性質與量性</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>課程 繼編</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>電腦及網路系統或工具</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>教材</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>舉辦之活動 競賽</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>研討會/工作坊</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>電子報、網站</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>計畫成果推廣之參與 （閱讀）人數</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>
國科會補助專題研究計畫成果報告自評表

請就研究內容與原計畫相符程度、達成預期目標情況、研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性）、是否適合在學術期刊發表或申請專利、主要發現或其他有關價值等，作一綜合評估。

1. 請就研究內容與原計畫相符程度、達成預期目標情況作一綜合評估
   ■達成目標
   □未達成目標（請說明，以 100 字為限）
   □實驗失敗
   □因故實驗中斷
   □其他原因
   說明：

2. 研究成果在學術期刊發表或申請專利等情形：
   論文：■已發表 □未發表之文稿 □撰寫中 □無
   專利：□已獲得 □申請中 ■無
   技轉：□已技轉 □洽談中 ■無
   其他：（以 100 字為限）

3. 請依學術成就、技術創新、社會影響等方面，評估研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性）（以 500 字為限）
   資訊科技和網際網路的蓬勃發展，異質資訊整合在企業電子化與電子商務環境中已是一項普遍存在且重要的議題。在缺乏整合的情形下，個別地存取異質資訊來源，已造成資訊的混雜和錯誤及浪費，尤其不能提供即時管理決策分析給企業主管。本研究發展出一項能夠整合不同資料模型，以及這些資料模式中衍生出的語意，依照延伸標記語言與本體論以及學名式的資料結構進行建模，並達到負載量模型具有可攜性和延展性的趨近一般化的績效評估方法。此資料模型結合本體和延伸標籤語言的查詢方法，提高對層級異質資料的結構互動性和語意互動。基於學名結構方法，發展單一的標準轉換模式和交換格式，促成分散地的資料倉儲間形成多對多的系統化映射模式，輔以多維度中介資料管理功能，可形成網路上通用並能兼顧效率與品質的徵型系統。對於在電子商務環境中，企業對於大量的資料和資訊，賴有效系統工具來協助進行萃取、分析和預測的重要商業智慧利器。本研究的研究成果已發表 Information & Management, Information Processing & Management, Information Sciences, Expert Systems with Applications 與國科會 A 級國際學術研討會論文。