

# Facilitating Integration of Distributed Statistical Databases Using Metadata and XML

Yaxin Bi and Joanne Lamb  
*Centre for Educational Sociology*  
*University of Edinburgh*  
*St John's Land, Holyrood Road*  
*Edinburgh, EH8 8AQ, UK*  
*Email: {Yaxin.Bi, J.M.Lamb}@ed.ac.uk*

**Abstract:** This paper describes a novel approach for integrating distributed databases. It is based on a semi-structured data framework which uses XML (eXtensible Markup Language) and metadata derived from data productions and corresponding databases. This framework serves as an "integrated dynamic content table", enabling the user to browse the relevant information and database structure required prior to query composition; it provides a robust facility for sending a valid query to distributed databases without a prior knowledge of the component schema structure; a new class of statistical applications may therefore be easily built and managed. It remedies deficiencies in querying to distributed databases that heavily relies on prior understanding of the schema structure. An initial prototype has been developed.

**Keywords:** metadata, XML, integration of statistical distributed databases

## 1. Introduction

In a distributed statistical database environment, data selection and actual statistical computation can be carried out at local databases, while data integration can bring together data produced from diverse sources, at different levels of details to produce statistical summaries. For example, the amount of beef retailed in European countries may be held in different local consumer databases and brought each country's consumer index. Then the total amount of beef retailed across the countries may be compared and summarized. Such data integration often requires associated local (data provider) and global (domain) metadata and can be treated as a process of distributed database querying, which is accomplished by SQL native operators or the external statistical operators of combining SELECT and GROUPBY. In this paper, we present an integration approach that was developed in a fourth framework project, ADDSIA (Access to Distributed Databases for Statistical Information and Analysis), and which is built on a semi-structured data framework consisting of local and global metadata coded in XML [1, 2]. Extensions of this approach will be used in a fifth framework project, MISSION (Multi-Agent Integration of Shared Statistical Information Over the [inter]Net), in which the mapping between the local metadata and public metadata standards will be dynamically established through agent

technology.

Traditional relational database systems force all data to adhere to an explicitly specified schema. A typical paradigm for querying databases is through an expressive, declarative query language – such as SQL – that relies on database schema. Beyond its use to define the structure of the data, a schema serves at least two important purposes [3]:

- A schema, in the form of tables and their attributes, enables users to understand the structure of the database and form meaningful queries over it.
- A query processor relies on the schema to devise efficient plans for computing query results.

With respect to distributed database systems composed of multiple databases, there are the several component schemas, and the mismatch encountered in the schemas is unavoidable. This leads to difficulties in composing a valid query and a corresponding query plan, and in comparing the querying results. To address such issues, most related work has focused on two major aspects: coping with discrepancies between component schemas that are semantically equivalent but structurally different, and effectively representing component schema structures. These various approaches have been proposed from different perspectives. For examples, the approach of schema integration for solving the problems of interoperability [4]; mapping logical federated schemas to component schemas in Federated Database systems [5]; the information mediation approach aimed at providing a uniform access to distributed heterogeneous databases [6, 7, 8]. All these depend heavily on a pre-defined mapping between a universal schema (global) and the component schemas (local), and may not be suitable for establishing flexible statistical summary operations.

In ADDSIA [9], it has been well recognized that information which views or schemas convey is not sufficient for automatically and even manually harmonising inconsistency between schemas and establishing statistical operations, and the assumption that the user has prior knowledge of the database can be not generalized to the practical cases. The integration of heterogeneous sources not only provides inter-operability between heterogeneous data, but should also provides an environment in which contextual information can be sought, thus aiding both processes of source exploration and query formulation, and allowing for the accomplishment of a range of sophisticated statistical operations. Thus we have developed an approach for representing underlying distributed databases based on local and global schemas derived from the sources themselves, along with associated background information such as concept definitions and classifications, etc. captured from the data productions. All of these used for representing the data are referred to as metadata or contextual information [10, 11].

A novel notion here is to construct a semi-structured framework through binding XML and metadata together to represent the structured data – distributed databases, and to establish a natural correspondence between the structures of the framework and the distributed databases. This framework not only can harmonize inconsistency across the schemas using the contextual information and the facilities of the XML DOM (Document Object Model) advanced technologies [12], but also it serves as an "integrated dynamic content table", thus enabling the user to browse the database structure and to facilitate query construction.

Therefore the process of statistical querying of distributed databases is greatly simplified from the user point of view. In addition, using agent technologies to develop an approach for dynamically mapping local schemas either to common ontologies or to user defined mappings is underway in the MISSION project.

## 2. Statistical metadata and XML

In distributed statistical databases environments, a sophisticated aggregation often involves a complicated database querying procedure. It can not directly be specified with the current SQL-like query language or predefined statistical operators, because information provided by database schema is not adequate for explicitly expressing users' computational requirements. Let us look at a following example to examine what extra information ( metadata) should be supplied to users to accomplish the statistical aggregation.

The scenario which is presented here is that of a user wishing to compare the attainment in mathematics of 16 year olds leaving school in Scotland in the years 1981 and 1991 [13]. This comparison task will involve two data tables stored in the different databases (Figure 1). As we see from Figure 1, tables of simple numbers or categorical values are, of course, meaningless. In themselves, they are not enough to carry out the statistical comparison. Some metadata – contextual information – thus turns the numbers and categorical values into statistical data. This includes headings of the tables, variables, labels, unit of measurement, time period, coverage, footnotes, sources, etc. This kind of metadata makes these tables meaningful, facilitating users to find and understand relevant data in the databases.

ID	Grade	...	ID	Grade	...
1	2	...	1	3	...
2	5	...	2	2	...
3	4	...	3	7	...
4	3	...	4	6	...
...	...	...	...	...	...

Figure 1: Two database tables containing variables of 1981(left) and 1991 (right) attainments

However, the comparison still may not be effectively undertaken, because the data stored in the two tables was collected based on two distinct classifications. In fact, the qualification issued in Scotland changed from O grades – which were awarded on a scale of 1-5 – to Standard grades – awarded on a scale of 1-7 – during this time, and the qualifications co-existed until 1996. The different classifications used in data collection results in a heterogeneity between the data of the two tables. Therefore, to make comparisons between the tables, such an inconsistency has to be harmonized, and a meaningful mapping has to be established beforehand (Figure 2). To achieve this, more information about the definitions of concepts, classifications, even quality declarations, etc., reflecting background information of data collection, has to be available to allow the user to evaluate the data and set up mappings.

Global ontology	Grade (1981)	Grade(1991)
High Pass	1-2	1-2
Pass	3	3-4

Low Pass	4-5	5-7
No Award	No Award	No Award

**Figure 2: A global ontology for the both variables of 1981 and 1991 attainments**

Obviously it is essential that the above two categories of metadata are made available to make a comparison across the tables as shown in Figure 1 in the statistical database environment. Of course, the degree of metadata support required relies on the designed capacity of a system. The growth of on-line statistical analyses and dissemination via Internet leads to a higher demand for metadata. Here we focus on metadata handling, rather than capturing, therefore broadly distinguishing between semantic metadata and structural metadata [1]. Semantic metadata provides information about the concept definition, classification, reference of data, etc. and their semantic relationships, e.g. data that describes the semantic content of a data value (like units of measure or scaling), or data that provides an additional information about its creation (derivation formula used), or mapping between classification and code lists. In contrast, structural metadata represents information that describes the organization and structure of the distributed databases, e.g., information about institution, database name, component database schema, attribute, data types of attribute, and the syntactic relationships between them. Each type of metadata derives from both domain and data provider level.

In order to effectively represent metadata, we employ XML, which has recently emerged as a new standard for data representation and exchange on the Internet [14, 15]. The basic ideas of utilising XML in our work are very simple: tags on data elements identify the meaning of the metadata, rather than, specifying how the metadata should be formatted (as in HTML), and relationships between metadata elements are provided via simple nesting and references. We have developed a Document Type Definition (DTD) of the domain-specific, in which a set of rules and a set of tags are established and conformed for XML data [1]. The DTD offers a homogeneous framework over the heterogeneous distributed databases, allowing multiple statistical data sources and the contextual information to be described, represented, and organized in the same DTD environment, constituting a semi-structured data framework.

### 3. Overview of the ADDSIA Architecture

The ADDSIA distributed database system consists of a number of local databases across the European statistical institutions [9]. It is built on a mediator architecture – the three-tier architecture – with facilities of developing Web-oriented applications. Except for the slightly different assignment of functions on each layer from the original specification described in [16], it also retains a partition of resources and services in two dimensions:

- horizontally into three layers, the client interface (client), the intermediate services modules (domain), and the distributed data sources (data provider).
- vertically into many domains with a number of supporting data sources, called data providers.

Figure 3 shows a simplified architecture of ADDSIA. In the data provider level, selection

of data is a basic function which is best performed in the local databases since one does not want to place large amount of unneeded data on the domain level or the client side for reason of both efficiency and data confidentiality. The SELECT statement of SQL and low level aggregation operations are effective instances of a function. Also some statistical operators can be placed on this level as a functional extension of native operations, allowing the data providers directly to carry out data aggregation.

The domain level is in the middle layer in the architecture – a central mediator. It is referred to as a global level relative to the data provider level. It consists of an independent semi-structured data framework that makes up the metadata encoded in XML (contextual information), and a set of mediation services and statistical operators. The semi-framework provides two fundamental functions: assisting users to explore and understand the content and structure of underlying databases; and facilitating query compositions and resolution of semantic discrepancies based on paths embedded in the hierarchical structure of the framework.

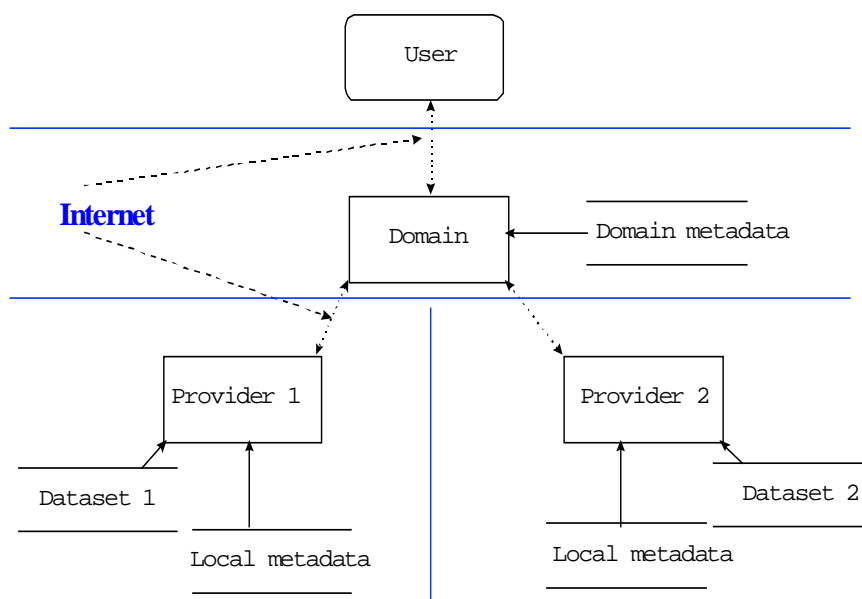


Figure 3: Architecture of ADDSIA

In the above architecture, we have applied concepts from data integration to the task of building complex Web sites that serve information derived from multiple databases [17], in terms of domain metadata and data provider metadata. In this scenario, the domain and data provider sites are declaratively defined as a graph – a graph of the content of underlying databases, over the semi-structured data framework [2]. A global query will be carried out over the data graph and then decomposed into the sub-queries over local database schemas.

#### 4. Browser Interface

To allow users rapidly to assess the content of the underlying databases through metadata e.g. find, understand and evaluate data, and construct valid queries based on the semi-structured framework, a Browser has been developed, which constitutes an important client component.

The ADDSIA Browser consists of five major parts of search engine, tree navigator, explorer for data set content, XML viewer, and DOM API interface as shown in Figure 4. Here we briefly describe two functions of them: searching and navigating.

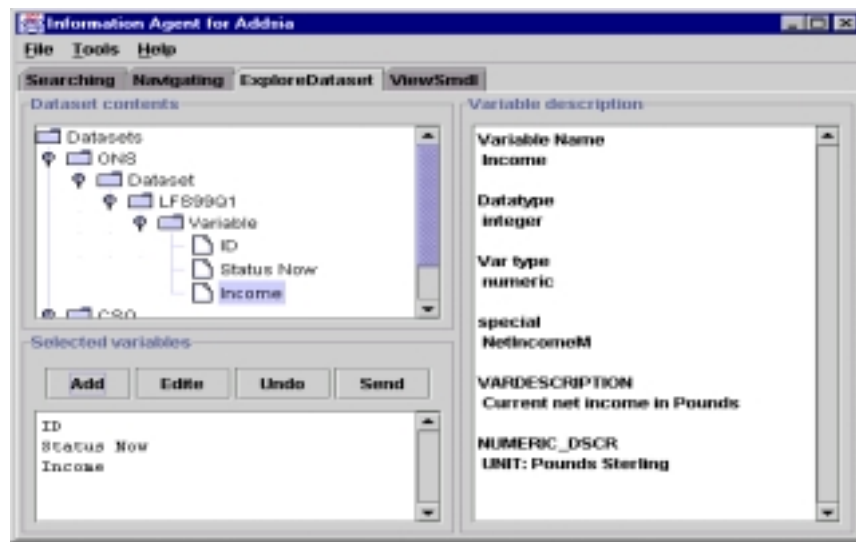


Figure 4: Browser interface

The searching function of the Browser is built on the *Isite / Isearch* search engine that has been customized for retrieving XML data. The customization includes changing the way of passing the query to the search engine (through the browser interface), identifying the structure of the tags for returning paths, rearranging the retrieved results on the screen and adding a highlighting keyword function. The search engine retrieves the relevant data based on key-word and fields defined by tags. Retrieved results may be associated with many paths, so that redundancy is unavoidable. But since the paths that XML data represent can often be restricted to be unique or valid using the context information, omitting redundancy has to be based on exact matches, rather than on similarity assessment. For example, the user may obtain many paths comprising the root, e.g. "school leavers", and the leaf node, e.g. "ID", however, if the user specifies "ONS" as a constraint to the paths, then some of the paths will be removed. Thus we need rules to justify 'path A is preferable over path B', this problem may be related to path expression optimization that remains for further investigation in MISSION.

The tree navigator provides the powerful browsing ability to obtain path expressions, however it is similar to the search in that it leaves the task of selecting a correct path expression to the user. Basically this function is built on the DOM Level 1 core API classes. In this primary interface, users have to manually navigate the hierarchical structure of XML data to obtain path expressions. An obvious advantage of the functions in this interface is that they directly return path expressions without the need for users to know much about the syntax of the query language.

As an example, Figure 4 presents a screen snapshot of the Java presentation of the Browser interface. The metadata displayed in a hierarchical structure summarizes an existing distributed database with the *School Leavers* domain across European countries. “Pluses” accompany complex objects and are used to expand or collapse sub-objects. Also, a “minus” is associated with each displayed attribute, corresponding to a unique path expression from the root. Based on the query, when the user highlights on an attribute "Income", it can be placed in the attribute box below using the button Add, and an associated path "SchoolLeavers.ONS.LFS99Q1" is automatically generated. When the user finishes attributes selection, the user clicks Send button, "SchoolLeavers.ONS.LFS99Q1" path and the attributes "Income" and "Total\_Inc" are bound together and sent to the query handler which is situated on the middle layer of the system. The query handler parses the decomposed query and send the sub-queries through the paths to the local physical databases to be executed.

## **5. Conclusion**

This paper describes a novel approach for integrating distributed heterogeneous databases, which are based on a semi-structured data framework that comprises the metadata and XML. The basic idea is to map a partition of a distributed statistical database into the hierarchical structure inherited in the semi-structured framework. In such a way, a query to the distributed database depending on an understanding of schema structure transforms into a query over an DOM tree structure and simple path expressions. However, obtaining a simple path is done through manually exploring metadata, automatically generating and optimizing simple path remains for further investigation in the MISSION project.

A distinction has been made with respect to various sources of information divided into three classes: structured, semi-structured and unstructured. Using structured data to represent and organize semi-structured and unstructured data has been exemplified by relational or object-oriented databases. The recent research on handling hybrid data also tries to take advantages from structured data management for semi-structured data. In this paper, we hold a belief that semi-structured representation could be an effective way to represent structured data, because it can turn structured data meaningful with less amount of information than unstructured data, which acts as a refinement to unstructured data and is manageable through the XML DOM advanced technologies.

We develop this framework with a comprehensive objective, that is it would be able to capture all metadata required in the ADDSIA distributed database system. Above we only illustrate an instance of its applications. In fact, all information required in queries, query

plans and semantic mappings, can derive from it. Using such an approach, semantic discrepancies between attributes could be easily resolved.

## 6. Acknowledgement

The work described is supported by the project “Access to Distributed Databases for Statistical Information and Analysis” (ESPRIT project no. 22950) which is funded as part of EUROSTAT’s “Development of Statistical Information Systems” (DOSIS) project. The authors wish to thank Prof Sally McClean who contributed to the earlier draft of this paper and Adam Taylor who provided helpful comments.

## References

- [1] Bi, Y., Murtagh, F. and McClean, S.I. Metadata and XML for Organising and Accessing Multiple Statistical Data Sources. Proceedings of ASC International Conference, Edinburgh, pp393-404, 1999.
- [2] Bi, Y., McClean, S.I., Scotney, B. Querying Distributed Statistical Databases Using Metadata and XML. 14th Conference of the International Association for Statistical Computing. The Netherlands, pp.15-16, 2000.
- [3] R. Goldman and J. Widom. DataGuides: Enabling Query Formulation and Optimization in Semi-structured Databases. Proceedings of the Twenty-Third International Conference on Very Large Data Bases, pages 436-445, Athens, Greece, August 1997.
- [4] Stonebraker, M.; Brown, P.; Herbach, M. Interoperability, distributed applications, and distributed databases: the virtual table interface. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, Sep. 1998, vol. 21 (no. 4): 25-34.
- [5] P. A. Deranley and D. J. Smith. Discovering and Using Entity Mappings in Federated Databases. Software – Practice and Experience. **29** (1), 17-42, 1999.
- [6] S. Bressan, C. H. Goh, K. Fynn, M. Jakobisiak, K. Hussein, H. Kon, T. Lee, S. Madnick, T. Pena, J. Qu, A. Shum and M. Siegel. The Context Interchange mediator prototype. Proceedings of the ACM SIGMOD International Conference on Management of Data. Pp 525-527, 1997.
- [7] Joachim Hammer, Héctor García-Molina, Svetlozar Nestorov, Ramana Yerneni, Marcus Breunig and Vasilis Vassalos. Template-based wrappers in the TSIMMIS system. Proceedings of the ACM SIGMOD International Conference on Management of Data Pages 532-535, 1997.
- [8] Ramana Yerneni, Chen Li, Hector Garcia-Molina, Jeffrey Ullman. Computing Capabilities of Mediators. SIGMOD'99, Philadelphia, PA, May 1999.
- [9] Lamb J M, et al. The ADDSIA project: Issues and Achievements. International Seminar of New Techniques & Technologies for Statistics, Italy, 1998.
- [10] V. Kashyap and A. Sheth. Semantic Heterogeneity in Global Information Systems: The Role of Metadata, Context and Ontologies. M. Papzoglou, and G. Schlageter, (Eds.), Academic Press. 1997.
- [11] Michael J. Colledge. Statistical Integration Through Metadata Management. International Statistical Review, 67(1), 1998.
- [12] Apparao V, Byrne S, et al, editors. Document Object Model (DOM) Level 1 Specification. <<http://www.w3.org/TR/PR-DOM-Level-1>>, 1998.
- [13] Lamb,J.M. and Smart,C. Simultaneous Analysis of Heterogeneous Databases on the Web: The ADDSIA Project in Recent Developments and Applications in Decision Making (eds Zanakis, S.H., Doukidis, G. and Zopounidis, C.), 2000 (forthcoming).
- [14] Bray T, Paoli J, and Sperberg-McQueen C M, editors. Extensible Markup Language (XML) 1.0. <http://www.w3.org/TR/1998/REC-xml-19980210>, 1998.
- [15] Jennifer Widom. Data Management for XML: Research Directions. Appears in IEEE Data Engineering Bulletin, Special Issue on XML, 22(3): 44-52, September 1999.
- [16] Gio Wiederhold: "Mediators in the Architecture of Future Information Systems". IEEE Computer, pages 38-49, March 1992.
- [17] Lee, T., Chams, M., Nado, R., Madnick, S., and Siegel, M. Information Integration with Attribution Support for Corporate Profile. ACM Conference on Information and Knowledge Management 1999.

- [18] B. Ludascher, Y. Papakonstantinou, P. Velikhov, V. Vianu. View Definition and DTD inference for XML. Workshop on Query Processing for Semi-structured Data and Non-Standard Data Formats . In conjunction with ICDT'99. January 13, 1999.
- [19] Dan Suciu. An Overview of Semi-structured Data. Published in SIGACT News, vol. 29, no. 4, pp. 28-38, December 1998.
- [20] Alin Deutsch Mary Fernandez Dana Florescu Alon Levy Dan Suciu. A query language for XML. In Proceedings of the Eighth International World Wide Web Conference (WWW8), 1999.