



UNIVERSITI PUTRA MALAYSIA

**A METHOD FOR MAPPING XML DTD TO RELATIONAL SCHEMAS IN THE
PRESENCE OF FUNCTIONAL DEPENDENCIES**

KAMSURIAH BT. AHMAD

FSKTM 2008 15



**A METHOD FOR MAPPING XML DTD TO RELATIONAL SCHEMAS IN
THE PRESENCE OF FUNCTIONAL DEPENDENCIES**

By

KAMSURIAH BT. AHMAD

**Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia,
in Fulfillment of the Requirements for the Degree of Doctor of Philosophy**

November 2008



Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfillment
of the requirement for the degree of Doctor of Philosophy

**A METHOD FOR MAPPING XML DTD TO RELATIONAL SCHEMAS IN
THE PRESENCE OF FUNCTIONAL DEPENDENCIES**

By

KAMSURIAH AHMAD

November 2008

Chair: Associate Professor Ali Mamat, PhD

Faculty: Computer Science and Information Technology

The eXtensible Markup Language (XML) has recently emerged as a standard for data representation and interchange on the web. As a lot of XML data in the web, now the pressure is to manage the data efficiently. Given the fact that relational databases are the most widely used technology for managing and storing XML, therefore XML needs to map to relations and this process is one that occurs frequently. There are many different ways to map and many approaches exist in the literature especially considering the flexible nesting structures that XML allows. This gives rise to the following important problem: Are some mappings ‘better’ than the others? To approach this problem, the classical relational database design through normalization technique that based on known functional dependency concept is referred. This concept is used to specify the constraints that may exist in the relations and guide the design while removing semantic data redundancies. This approach leads to a good normalized relational schema without data redundancy. To achieve a good normalized relational schema for XML, there is a need to extend the concept of functional dependency in relations to XML and use this concept as guidance for the design. Even though there exist functional dependency definitions for XML, but



these definitions are not standard yet and still having several limitation. Due to the limitations of the existing definitions, constraints in the presence of shared and local elements that exist in XML document cannot be specified. In this study a new definition of functional dependency constraints for XML is proposed that are general enough to specify constraints and to discover semantic redundancies in XML documents.

The focus of this study is on how to produce an optimal mapping approach in the presence of XML functional dependencies (XFD), keys and Data Type Definition (DTD) constraints, as a guidance to generate a good relational schema. To approach the mapping problem, three different components are explored: the mapping algorithm, functional dependency for XML, and implication process. The study of XML implication is important to imply what other dependencies that are guaranteed to hold in a relational representation of XML, given that a set of functional dependencies holds in the XML document. This leads to the needs of deriving a set of inference rules for the implication process. In the presence of DTD and user-defined XFD, other set of XFDs that are guaranteed to hold in XML can be generated using the set of inference rules. This mapping algorithm has been developed within the tool called XtoR. The quality of the mapping approach has been analyzed, and the result shows that the mapping approach (XtoR) significantly improve in terms of generating a good relational schema for XML with respect to reduce data and relation redundancy, remove dangling relations and remove association problems. The findings suggest that if one wants to use RDBMS to manage XML data, the mapping from XML document to relations must based be on functional dependency constraints.



Abstrak yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk ijazah Doktor Falsafah

**SATU KAEDAH PEMETAAN XML DTD KE SKEMA HUBUNGAN
DENGAN KEHADIRAN SANDARAN FUNGSIAN**

Oleh

KAMSURIAH AHMAD

November 2008

Pengerusi: Professor Madya Ali Mamat, PhD

Fakulti: Sains Komputer dan Teknologi Maklumat

XML (Extensible Markup Language) kini menjadi satu piawaian bagi persembahan dan perantaraan data di laman sesawang. Disebabkan semakin banyak data XML di gunakan, kini persoalan yang timbul adalah bagaimana untuk menguruskan data ini secara efektif. Disebabkan pangkalan data hubungan digunakan secara meluas untuk mengurus dan menyimpan data XML, oleh itu XML perlu dipetakan kepada skema hubungan dan proses ini berlaku agak kerap. Terdapat pelbagai cara bagaimana pemetaan boleh dilakukan dan terdapat pelbagai kaedah yang wujud berdasarkan kepada struktur XML yang fleksibel. Ini membawa kepada satu permasalahan yang penting: Adakah satu kaedah pemetaan lebih baik daripada kaedah pemetaan yang lainnya? Sebagai pendekatan kepada masalah ini, reka bentuk pangkalan data hubungan yang klasik melalui teknik penormalan berdasarkan kepada konsep sandaran fungsian dirujuk. Konsep ini diguna untuk menyatakan kekangan yang mungkin terdapat dalam data hubungan dan sebagai panduan untuk mereka bentuk data hubungan di samping menghapuskan pertindihan data semantik. Pendekatan ini membuka laluan kepada satu reka bentuk skema hubungan normal yang baik tanpa



pertindihan data. Untuk mencapai skema hubungan normal yang baik, konsep sandaran fungsian dalam data hubungan perlu diperluaskan kepada XML dan seterusnya menggunakan konsep ini sebagai panduan untuk mereka bentuk. Walaupun definisi sandaran fungsian bagi XML telah wujud tetapi definisi ini belum mencapai taraf yang piawai dan masih mengalami pelbagai kekurangan. Disebabkan kekurangan ini, kekangan tidak dapat dinyatakan sekiranya elemen-kongsian dan elemen-lokal wujud di dalam dokumen XML. Di dalam kajian ini satu definisi sandaran fungsian yang lebih umum dicadangkan untuk menyatakan kekangan dan mengesan pertindihan data semantik dalam dokumen XML.

Tumpuan kajian ini adalah mencadangkan satu kaedah pemetaan dengan kehadiran kekangan sandaran fungsian XML, kekunci dan Definisi Jenis Dokumen (DTD) sebagai panduan untuk menghasilkan satu skema data hubungan yang baik. Sebagai pendekatan kepada permasalahan ini, tiga komponen diterokai: algoritma pemetaan, sandaran fungsian bagi XML dan proses penaakulan. Kajian ke atas penaakulan XML adalah penting untuk mentaakul sandaran fungsian lain yang wujud dalam perwakilan data hubungan bagi XML, apabila diberi satu senarai sandaran fungsian. Ini membawa kepada keperluan menjana satu senarai petua taakulan. Dengan kehadiran DTD dan sandaran fungsian yang diberi oleh pengguna, sandaran fungsian lain yang dijamin menepati kekangan XML dapat dijana berdasarkan kepada petua taakulan. Kaedah pemetaan ini dibangunkan ke dalam alat pemetaan yang dipanggil XtoR. Keberkesanan cadangan kaedah pemetaan ini dianalisis dan hasil analisis ini menunjukkan XtoR mampu menghasilkan skema data hubungan yang baik bagi XML dari segi mengurangkan pertindihan data dan jadual, mengurangkan jadual tergantung dan mengurangkan masalah jadual berkait. Daripada penemuan ini, kajian

ini mencadangkan sekiranya XML dokumen ingin diuruskan oleh Sistem Pangkalan Data Hubungan, kaedah pemetaan mestilah berdasarkan kepada sandaran fungsian.

ACKNOWLEDGEMENTS

In the name of Allah, The Most Gracious, The Most Merciful. I thank Allah for granting me the perseverance and the strength I needed to complete this thesis.

In preparing this thesis, I was in contact with many people, researchers, academicians, and practitioners. They have contributed towards my understanding and thoughts. I wish to express my sincere appreciation to my main thesis supervisor Associate Professor Dr. Ali Mamat who has supported, inspired, motivated, and challenged me throughout my studies. He encouraged and helped me to stay motivated and focused throughout this lengthy period. Thanks also go to the members of my supervisory committee: Associate Professor Dr. Hamidah Ibrahim and Associate Professor Dr. Shahrul Azman Mohd Noah for their knowledgeable suggestions, comments and criticisms.

I would like to express my gratitude to JPA, by providing the scholarship, Universiti Kebangsaan Malaysia by giving me a study leave, and to FTSM by giving me a chance to further my studies.

Finally, I would like to thank my family, especially to my husband and to my five wonderful kids Aimi Dalila, Aimi Syazana, Aimi Marsya, Muhammad Adiib Suhail, and Aimi Hasya. Their loves and supports have given me the strength and confidence to complete this endeavor.



I certify that an Examination Committee has met on 10/11/2008 to conduct the final examination of **Kamsuriah Ahmad** on her **Doctor of Philosophy** thesis titled “**A Method for Mapping XML DTD to Relational Schemas in the Presence of Functional Dependencies**” in accordance with Universiti Pertanian Malaysia (Higher Degree) Act 1980 and Universiti Pertanian Malaysia (Higher Degree) Regulations 1981. The Committee recommends that the students be awarded the Doctor of Philosophy.

Members of the Examination Committee were as follows:

Name of Chairperson

Associate Professor Dr. Md. Nasir Sulaiman

Computer Science Department

Faculty of Computer Science and Information Technology

University Putra Malaysia.

Name of Examiner 1, PhD

Dr. Lily Suriani Affendy

Computer Science Department

Faculty of Computer Science and Information Technology

University Putra Malaysia.

Name of Examiner 2, PhD

Associate Professor Dr. Abdul Azim Abd. Ghani

Dean

Faculty of Computer Science and Information Technology

University Putra Malaysia.

Name of External Examiner, PhD

Y. Bhg. Professor Dr. Abdullah Embong

Faculty of Computer System and Software Engineering

Universiti Malaysia Pahang

HASANAH MOHD. GHAZALI, PhD

Professor and Deputy Dean

School of Graduate Studies

Universiti Putra Malaysia

Date:



This thesis was submitted to the Senate of Universiti Putra Malaysia and has been accepted as fulfilment of the requirement for the degree of Doctor of Philosophy. The members of the Supervisory Committee were as follows:

Ali Mamat, PhD

Associate Professor
Faculty of Science Computer and Information Technology
Universiti Putra Malaysia
(Chairman)

Hamidah Ibrahim, PhD

Associate Professor
Faculty of Science Computer and Information Technology
Universiti Putra Malaysia
(Member)

Shahrul Azman Mohd Noah, PhD

Associate Professor
Faculty of Technology and Information Science
Universiti Kebangsaan Malaysia
(Member)

HASANAH MOHD. GHAZALI, PhD

Professor and Dean
School of Graduate Studies
Universiti Putra Malaysia

Date: 15 January 2009



DECLARATION

I declare that the thesis is my original work except for quotations and citations, which have been duly acknowledged. I also declare that it has not been previously, and is not concurrently, submitted for any other degree at Universiti Putra Malaysia or at any other institution.

KAMSURIAH BT AHMAD

Date:



TABLE OF CONTENTS

	Page
ABSTRACT	ii
ABSTRAK	iv
ACKNOWLEDGEMENTS	vii
APPROVAL	viii
DECLARATION	x
LIST OF TABLES	xiv
LIST OF FIGURES	xv
LIST OF ABBREVIATIONS	xviii
CHAPTER	
1 INTRODUCTION	1
1.1 Background of Studies	2
1.2 Problem Statements	7
1.3 Motivating Examples	8
1.4 Research Questions	14
1.5 Objectives	16
1.6 Significance of Research	16
1.7 Research Methodology	18
1.8 Thesis Outline	21
2 THEORETICAL BACKGROUND: XML STRUCTURES AND CONSTRAINTS	24
2.1 Extensible Markup Language (XML)	24
2.2 Schema Language for XML	25
2.2.1 XML DTD	27
2.2.2 XML-Schema (XSD)	28
2.3 Structure Constraints in DTD	29
2.3.1 DTD Cardinality Constraints	31
2.3.2 ID and IDREF Constraints	32
2.3.3 DTD Graph	34
2.4 Semantic Constraints for XML	35
2.5 Basic Notations for XML Model	37
2.6 XML Functional Dependency vs Relational Functional Dependency	47
2.7 XFD Implication	50
2.8 Summary	52
3 LITERATURE REVIEWS	54
3.1 Management of XML Data	55
3.2 Managing XML Data in Relational Database	59
3.2.1 Model-based Approach	60
3.2.2 Structural-based Approach	64
3.2.3 Semantics-based Approach	68



3.3	Issues in Mapping from XML to Relational	74
3.4	Comparative Analysis of XFDs	76
3.5	Summary	81
4	FUNCTIONAL DEPENDENCIES AND INFERENCE RULES FOR XML	84
4.1	Path Expressions and Equality Testing	85
4.2	Functional Dependency Constraint Language for XML	86
4.3	Inference Rules for XML	91
4.3.1	Inference Rules for XML in the Presence of Keys	96
4.3.2	Inference Rules for XML in the Presence of DTD Cardinality Constraints	98
4.4	On Interaction of XML Keys, DTD Cardinality Constraints and XFDs	101
4.5	The Soundness of the Inference Rules	106
4.6	Summary	107
5	XtoR: A METHOD FOR MAPPING XML DTD TO RELATIONALSCHMAS IN THE PRESENCE OF FUNCTIONAL DEPENDENCIES	109
5.1	Relationship between Dependencies and Redundancies in XML	110
5.2	Normalized Relational Schema for XML	112
5.3	Designing Good Relational Schema for XML	113
5.4	Designing XtoR: The Mapping Method	125
5.4.1	Simplifying DTD	130
5.4.2	Creating DTD Graph	131
5.5	Constructing the Mapping Process	133
5.6	Schema Tree Construction	135
5.7	Developing XtoR: the Mapping Method	148
5.8	Computing Minimum Covers	152
5.9	Running Example	157
5.10	The Correctness of the XtoR Algorithm	160
5.11	Summary	164
6	RESULTS AND DISCUSSION	166
6.1	The Generated Schema - Result Comparison	167
6.2.	Experiment Using Dataset 1	168
6.2.1	Schema generated Using XtoR	169
6.2.2	Schema generated Using RRXS	170
6.2.3	Schema generated Using Lv&Yan	172
6.3.	Experiment Using Dataset II	175
6.3.1	Schema generated Using XtoR Algorithm	177
6.3.2	Schema generated Using RRXS Algorithm	177
6.3.3	Schema generated Using Lv&Yan Algorithm	178
6.4	Size of Database	180
6.5	Discussion	181
6.6	Summary	185



7	SUMMARY, CONCLUSION AND FUTURE WORKS	186
7.1	Summary of Research	186
7.2	Conclusion	190
7.3	Contributions	195
7.4	Recommendations For Future Research	197
7.5	Closing Notes	201
	REFERENCES	203
	APPENDICES	210
	BIODATA OF STUDENT	214
	LIST OF PUBLICATIONS	215



LIST OF TABLES

Table		Page
1.1	The input, activities and deliverables of Phase 1	19
1.2	The Input, Activities and Deliverables of Phase 2	20
1.3	The Input, Activities and Deliverables of Phase 3	20
1.4	The Input, Activities and Deliverables of Phase 4	21
3.1	The Generated Schema by Edge Algorithm	61
3.2	Table Author Generated Using Hybrid Approach	65
3.3	Table of Comparison	80
5.1	Data Redundancy in Student Table	115
6.1	Size of Database	180
6.2	The Limitations of RRXS and Lv&Yan Approaches	182



LIST OF FIGURES

Figure		Page
1.1	Trends for Data Exchange in Web Application Leading to the Problem	6
1.2	XML Document for Publication	9
1.3	XML document for Sigmod Record	11
1.4	The Research Methodology	18
1.5	Organization of Thesis	23
2.1	An Example of XML Document	25
2.2	DTD for Publication	28
2.3	Inconsistent DTD	30
2.4	DTD Graph with Shared-element	35
2.5	Simplification Step to Remove the Text Node	38
2.6	A Definition of an XML Tree	40
2.7	An XML Document Views as a Node-labeled Tree	41
2.8	The contents of $D = (E_1, E_2, A, M, N, r)$.	42
2.9	Table Item and its Values	49
3.1	Comparison of Schema Generated Based on Keys and XFDs	71
4.1	Illustration of XFD $\varphi = P: Q : X_1, \dots, X_n \rightarrow Y_1, \dots, Y_m$	88
4.2	An XML Document about Faculty	90
4.3	Downward Expansion Rule	104
4.4	The Illustration of Target-to-Context Rule	105



5.1	Table Redundancy Between Table Author and Table Author1	116
5.2	The DTD and DTD graph about Faculty	118
5.3	The Dangling Table Problem	120
5.4	Comparison Between Two SQL Statements	122
5.5	The XtoR Mapping Method	128
5.6	DTD Graph about Faculty	133
5.7	An Example of Shared-element in DTD	134
5.8	Relational Schema Design in the Presence of Shared-element in XML	138
5.9	The Reconstruction Step in the Presence of Set-element	139
5.10	The Mapping Process in the Presence of Local-element	141
5.11	Mapping to Relations in the Presence of Extended Simple-element	142
5.12	The Mapping Process in the Presence of 1:N	144
5.13	The Mapping Process in the Presence of M:N	146
5.14	The Mapping Process in the Presence of Recursive Element	147
5.15	The XtoR Algorithm	149
5.16	The Structure for Schema Tree and XFD	150
5.17	Procedure ConstructSchemaTree	151
5.18	The MinimumCover Procedure	155
5.19	The ReduceXFD Procedure	156



5.20	SchemaTree and Σ_m Constructed from DTD	158
5.21	A List of Marked XFDs in Σ_m and F	158
5.22	A Reduced List of XFD called F_m	159
6.1	Result Comparison Strategy	168
6.2	DTD Graph for Publication and its Corresponding DTD	168
6.3	XFD Constraints from Publication Document	169
6.4	The Schema Generated Using XtoR Algorithm	169
6.5	The Schema Generated Using RRXS	170
6.6	Redundant Student Table	171
6.7	The XFDs Expressed in Simple Path	173
6.8	The Schema Generated Using Lv&Yan Algorithm	173
6.9	Redundant Author Table	175
6.10	The DTD Graph for Publication and its DTD Schema	176
6.11	The XFD constrains in Sigmod Record Document.	176
6.12	The Schema Generated Using XtoR Algorithm	177
6.13	The XFDs Constraints Expressed in RRXS	177
6.14	The Schema Generated Using RRXS Algorithm	178
6.15	The XFDs in Simple Path	178
6.16	The Schema Generated Using Lv&Yan Algorithm	179



LIST OF ABBREVIATIONS

XML	Extensible Markup Language
XFD	XML Functional Dependency
XFDs	XML Functional Dependencies
DTD	Data Type Definitions
FD	Functional Dependency
FDs	Functional Dependencies
RDBMS	Relational Database Management Systems
CLOB	Character Large Object
BLOB	Binary Large Object



CHAPTER 1

INTRODUCTION

This chapter introduces the thesis. The discussion starts in Section 1.1 on the importance of Extensible Markup Language (XML) technology in data exchange environment. With the large amount of data being represented in XML on the web today, the question on how to manage this data effectively is raised. Studies (Liu et al., 2006; Fan, 2005; Kay, 2003) have shown that relational technology is still the best alternative to manage XML contents. Therefore, the need to map XML to relational schema has increased. The main problem in this context is to define what will be the best design in producing XML contents in the relational environment.

To approach this problem, the first thing that needs to be done is to define what is meant by “the best mapping method”. This unsolved puzzle, finding the best mapping for designing XML in relations, has become the motivation for the study. In Section 1.2, the existing problems in the mapping methodology are being discussed extensively and the criteria for being good design for XML in relations are also precisely defined. The motivating examples in Section 1.3 discuss the remaining issues in the existing mapping problems and this is the key to the formulation of this study. Research questions are identified and defined in Section 1.4. Objectives of the study are outlined in Section 1.5. In Section 1.6, the significance of research is clearly stated. The limitation and key assumption for the study are defined in Section 1.7. The methodology of research is broadly presented in Section 1.8. Finally, the overall organization of the thesis is described in Section 1.9.



1.1 Background Of Studies

XML technology, (Bray et al., 1998) recommended by the World Wide Web consortium, has fast become the dominant standard for data interchange and data representation on the web. It enables the storage of structured information and provides a platform-independent means to describe data. Therefore, it makes transporting data from one platform to another become easy. With these features, XML has enabled the communication between different computing systems, which was impossible or very hard to do before. XML thus provides a universal framework for the interchange of data regardless of the platforms and data models of the applications. Computing world now has a new way of implementing a distributed application systems. Nowadays, the majority of both traditional business applications and Internet based applications depend on databases management system in order to be operational (Abiteboul et al., 2000). To maintain data in a database, it must be retrieved and stored in a consistent, reliable, and efficient manner. With the large amount of data now being represented in the XML on the web, the question raised is, how to manage the data in terms of storing, updating and accessing in the same manner as it was done in database information system.

Since an XML document is a prime example of semi-structured technology, there has been an effort to use this technology to manage XML. Using semi-structured technology is indeed a viable alternative and there are considerable works in this community that focus on exploiting this approach. But the other issue that might rise is whether this is the only approach that we have. By using semi-structured database we may ignore nearly three decades of research and development in building and

maturing relational database systems, which have the commercial strength from the giant vendors. Furthermore, relational databases are famous for data management in terms of storing, updating and searching capabilities through its communication language (Structured Query Language). In view of the maturity of this technology, XML data shall adapt to the way how data has been managed in relational, therefore, need to be stored in relations. It is obvious that relational database management systems (RDBMS) will remain dominant in managing business data in the foreseeable future due to their powerful data management services (Shanmugasundaram, 1999). With this approach, XML document will be represented as a relational database and users can access the document by using the same mechanism as being used in relational database. Once they are created, the queries (including search, insert, update, delete) over the document are translated into queries over a normal relational database and the result of the queries will be translated back into XML, where all these processes will be done internally (Krishnamurthy et al., 2004; Shanmugasundaram et al., 2000).

Numerous researches focusing on the mapping process between XML documents and relational databases (Lv and Yan, 2006; Chen et al., 2003; Shanmugasundaram, 1999; Florescu and Kossman, 1999a). The main intention was to take advantage of the properties from both presentations. This is the similar problem that we would like to address in this study. However, in the mapping context, another problem arises: Given an XML document and its constraints, how to design a good relational schema to store the XML data? The issue of how to design good relational database has been the central focus in the database research. The industry has gone through the bad experience and suffers a very high maintenance cost when the database was

poorly designed. To approach this problem, the analogy of designing relational database is referred, with regard that the design is considered good if the database schema is redundancy free without anomaly problems (Elmasri and Navathe, 2006; Abiteboul et al., 1995; Batini et al., 1992). This design theory is based on the normalization technique which based on the well known functional dependencies. We believe that the study of this design technique in the context of XML is equally significant towards designing good relational schema for XML. To achieve good non-redundant relational schema for XML is important in order to avoid higher data storage cost, increased cost for data transfer, and data manipulation. Furthermore data redundancy could lead to potential update anomalies, rendering the database inconsistent. Therefore the problem that being investigated in this research is, how to extend the classical approach used in designing relational database and transform the finding to become the best mapping approach for designing XML in relations.

The notion of functional dependency (FD) plays a central role in specifying constraints and discovering redundancies in relational databases, and should play a central role in XML as well. However, it is not immediately obvious how to extend the definitions of redundancies from relations to XML because of the flexible structure of XML. Also the concept of functional dependency in relations does not immediately applies to XML. Now, the theory of functional dependencies in relational database context has matured. If we are to achieve the same functionality for XML in relations, it is essential to adapt the study of functional dependency in the context of XML. Recent studies in the context of integrity constraint for XML paying particular attention to the class of functional dependencies (Wang and Topor, 2005; Schewe, 2005; Arenas and Libkin, 2004; Vincent et al., 2004) as

renewed interest in designing XML schema in relational setting in the presence of these constraints (Lv and Yan, 2006; Chen et al., 2003; Qing et al., 2003). Figure 1.1 summarizes the current trends using XML for data exchange that leads to the needs of mapping from XML to relations in the presence of functional dependency. The problems faced during the mapping have lead to the motivation of this study.

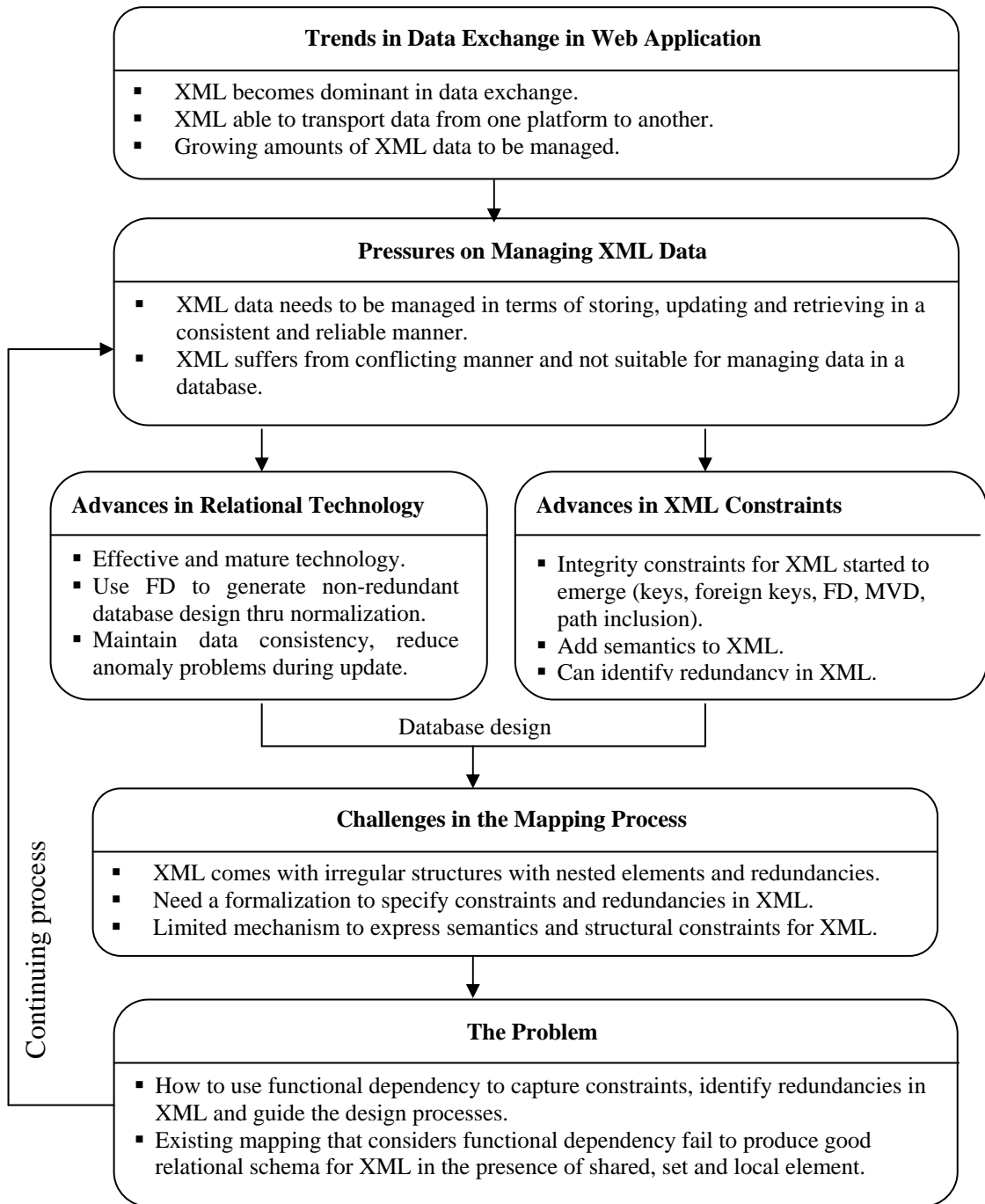


Figure 1.1: Trends for Data Exchange in Web Application Leading to the Problem