Capabilities of Native XML and Flat File Systems For Document Management

This thesis is presented to the Graduate School

in fulfillment of the requirements for

Master of Science (Information Technology)

Universiti Utara Malaysia

By

Mazlyda Abdul Rahman

JABATAN HAL EHWAL AKADEMIK
(DEPARTMENT OF ACADEMIC AFFAIRS)
UNIVERSITI UTARA MALAYSIA

PERAKUAN KERJA / TESIS
(Certification of Thesis Work)

Kami, yang bertandatangan, memperakukan bahawa
(We, the undersigned, certify that)

MAZLYDA ABDUL RAHMAN

calon untuk Ijazah
(candidate for the degree of) **SARJANA SAINS (TEKNOLOGI MAKLUMAT)**

telah mengemukakan tesis/disertasinya yang bertajuk:
(has presented his/her thesis work of the following title)

CAPABILITIES OF NATIVE XML AND FLAT FILE SYSTEMS FOR DOCUMENT MANAGEMENT

seperti yang tercatat di muka surat tajuk dan kulit tesis/disertasi
(as it appears on the title page and front cover of thesis work)

bahasa tesis/disertasi tersebut boleh diterima dari segi bentuk serta kandungan, dan liputan bidang ilmu yang memuaskan, sebagaimana yang ditunjukkan oleh calon dalam ujian lisan yang diadakan pada :
(that the thesis/dissertation is acceptable in form and content, and that a satisfactory knowledge of the field covered by the thesis was demonstrated by the candidate through an oral examination held on **1 APRIL 2004**

Pengerusi Viva *(Chairman for Viva)*	:Prof. Madya Hatim Mohamed Tahir	Tandatangan: *(Signature)*
Penilai Luar *(External Assessor)*	:Prof. Madya Dr. Ali Mamat	Tandatangan: *(Signature)*
Penilai Dalaman *(Internal Assessor)*	:Prof. Madya Md. Zahir Mat Cha	Tandatangan: *(Signature)*
Penyelia Utama *(Principal Supervisor)*	:Prof. Dr. Ku Ruhana Ku Mahamud	Tandatangan: *(Signature)*
Dekan Fakulti Teknologi Maklumat *(Dean Faculty of Information Technology)*	:Prof. Dr. Ku Ruhana Ku Mahamud	Tandatangan: *(Signature)*

Tarikh
(Date) : **1 APRIL 2004**

PERMISSION TO USE

In presenting this thesis in fulfillment of the requirements for a Master of Science in Information Technology degree from Universiti Utara Malaysia, I agree that the University Library may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by my supervisor(s) or, in their absence, by the Dean of Graduate School. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to Universiti Utara Malaysia for any scholarly use which may be made of any material from my thesis.

Requests for permission to copy or make other use of materials in this thesis, in whole or in part, should be addressed to:

Dean of Graduate School
Universiti Utara Malaysia
06010 UUM Sintok
Kedah Darul Aman

ABSTRAK

Teknologi XML merupakan antara teknologi terkini yang sedang berkembang pesat di dalam era teknologi maklumat. Kelebihan utama XML ialah ia membolehkan data-data dihantar dan ditukar di antara pelbagai sistem pengoperasian yang berlainan. XML juga mempunyai kelebihan dari segi keupayaan untuk mencipta tag yang lebih mudah difahami, mengasingkan kandungan dokumen dengan format paparan bagi memudahkan pembangunan dan penyelenggaraan sistem, membolehkan integrasi antara data dan dokumen serta keupayaannya untuk memaparkan kembali dokumen XML dalam pelbagai format. Secara asasnya, terdapat pelbagai pilihan untuk menyimpan dan mengurus dokumen XML dan dokumen bukan XML, tidak hanya terhad kepada *native XML database.* Sebaliknya, pangkalan data sedia ada seperti *flat file database, relational database* dan *object-oriented database* turut boleh digunakan. Namun, di sebalik kelebihan tersebut, wujud masalah bagaimana untuk menguruskan data dan dokumen-dokumen tersebut, serta ketiadaan garis panduan berkenaan keupayaan pangkalan data-pangkalan data tersebut menguruskan dokumen-dokumen berkenaan. Sehubungan itu, kajian ini dijalankan bagi mengkaji keupayaan *native XML* dan *flat file database* dalam menguruskan pelbagai dokumen. Setiap pangkalan data ini menggunakan pendekatan yang berbeza dalam menguruskan dokumen-dokumen tersebut. Satu sistem prototaip telah dibangunkan dan dokumen XML serta dokumen bukan XML telah digunakan sebagai data untuk menguji keupayaan teknik penyimpanan dan pencapaian tersebut. Berdasarkan keputusan yang diperolehi, satu garis panduan berkaitan keupayaan *native XML database* dan *flat file database* menguruskan pelbagai dokumen telah dicadangkan.

ABSTRACT

XML technology is becoming one of the rapidly growing technologies in this information technology era. The main advantage of XML is it could be used for data transfer and data exchange between different operating systems. It also enables creation of more meaningful tags, separates the content of a document from its presentation, which simplifies the development and maintenance, enable integration between data and documents, and the XML documents could be published in a variety of forms. Basically, to store and manage the XML and non-XML documents, not only native XML database could be used, but flat file database and the existing relational and object-oriented database could also be used as alternatives. However, besides those advantages, the problem on how to manage the documents and lack of guidelines on capabilities of those databases in managing documents aroused. Therefore, this research is undertaken to study the capabilities of native XML and flat file database in managing documents, where each database has its own different methods. A prototype system had been developed and both XML document and non-XML document had been used as the testing data for the purpose of examining the capabilities of storing and retrieving techniques in both databases. Based on the result derived, a guideline on the capabilities of storing and retrieving those documents in native XML database and flat file database was proposed.

ACKNOWLEDGEMENT

In the name of Allah, the Most Gracious and the Most Merciful.

I would like to extend my sincere gratitude and appreciation to:

My supervisor, Professor Dr. Ku Ruhana Ku Mahamud and my previous supervisor, Mr. Helmi Mohamed Hussain for their guidance, brilliant ideas and care of my progress during my study.

The ex-Dean, Professor Dr. Abu Talib Othman for taking me into the research world, sets the research environment during the initial stage of this study.

The research teams especially the 'XMLers'; Aida Liza, Khairul Adilah, Siti Salmah and Rohaya for their useful help, recommendations, and assistance throughout completing the study.

The Ministry of Science and Technology for the monetary support.

Universiti Utara Malaysia for the facilities and resources provided.

And all the individuals involved in the establishment of this study.

To Mak, Abah, Along & Abang Muaz, Abang, Ateh and Adik,
thank you for your love and the continuous support.

TABLE OF CONTENTS PAGE

List of Tables

List of Figures

List of Abbreviations

ASP	Active Server Page
B2B	Business To Business
BLOB	Binary Large Object
CDATA	Character Data
CERN	European Particle Physics Laboratory
CLOB	Character Large Object
CSV	Comma Separated Value
DAD	Data Access Definition
DBMS	Database Management System
DOCTYPE	Document Type
DOM	Document Object Modelling
DTD	Document Type Definition
EDI	Electronic Data Interchange
EDUX	EDucation with an aUthoring tool using XML
EIP	Energy Integration Platform
EWM	Enterprise Web Machines Corporation
FpML	Financial Products Markup Language
GIF	Graphics Interchange Format
GUI	Graphical User Interface
HTML	Hypertext Markup Language
IETIS	Integrated Electronic Technical Information System
iFS	Internet File System
JPEG/ JPG	Joint Photographic Experts Group
LGPL	Lesser General Public License
MPEG	Motion Picture Experts Group
NCSA	National Center for Supercomputing Applications
NXD	Native XML Database
OAGIS	Open Applications Group Interface Specification
ODBC	Object Database Connectivity
OODBMS	Object Oriented Database Management Systems
PCDATA	Parsed Character Data
PDF	Portable Document Format
PI	Processing Instruction
RADIX	Rapid Application Development In XML
RDBMS	Relational Database Management Systems
RTF	Rich Text Format
SAX	Simple API for XML
SGML	Synchronized Generalized markup Language
SMH	System Management Hub
SQL	Structured Query Language
TSE	Tamino Schema Editor
UN/CEFACT	United Nations Centre for Trade Facilitation and Electronic Business
WebML	Web Modeling Language
XFRML	XML-based Financial Reporting Markup Language
XML	eXtensible Markup Language
XQL	XML Query Language
XSL	Extensible Stylesheet Language

CHAPTER 1
INTRODUCTION

Among the first database system built upon the Structured Query Language (SQL) standard that appeared at the beginning of the 1980's were from Oracle with its Oracle Version 2 and later from IBM with its SQL/DS, as well as a host of other systems from other companies. The software of relational databases was continually refined during the 1980's. This was in part due to feedback from customers, development of systems for new industries and the increase usage of personal computers and distributed systems. Generally, relational database is a tabular database in which data is defined so that it can be reorganized and accessed in many different ways.

In mid 1980's, it had become obvious that there were several fields where relational databases were not practical due to the types of data involved. This led to research being started in object-oriented databases where users could define their own methods of access to data and how it was represented and manipulated. Furthermore, object oriented database is congruent with the data defined in object classes and subclasses. By the beginning of 1990's, the first Object Oriented Database Management Systems (OODBMS) had started to appear from companies like Objectivity.

However, problem existed when integration between different databases such as relational-to-relational, object oriented-to-object oriented and relational-to-object oriented need to be done. Codes that react as a converter between those databases need to be written so that data can be exported from one database to another. In fact, Object Database Connectivity (ODBC) was devised to try and bridge the gap between different database systems as shown in Figure 1.1.

The contents of the thesis is for internal user only

REFERENCES

Anderson, R., Birbeck, M., Kay, M., Livingstone, S., Loesgen, B., Martin, D., Mohr, S., Ozu, N., Peat, B., Pinnock, J., Stark, P., Williams, K. (2000). **Professional XML**. Wrox Press Ltd, USA.

Anonymous. XML Enables Enterprise Web Site Development, **Sun Journal,** *5*(1).

Bourret, R. (2001). XML Database Products. http://www.rpbourret.com/xml/ XMLDatabaseProds.htm#native, 10 May 2001.

Bressen, S., Lacroix, Z., Lee, M.L., Li, Y. and Nambiar, U. (2002). Current Approaches to XML Management, **IEEE Internet Computing**, August 2002, pp 43-51.

Bryan, M. (1997). An Introduction to the Extensible Markup Language (XML), http://www.sgml.u-net.com/xmlintro.htm, 14 November 2000.

Ceponkus, A. and Hoodbhoy, F. (1999). **Applied XML: A Toolkit For Programmers**. Wiley Computer Publishing, USA.

Ceri, S., Fraternali, P. and Bongio, A. ( ). Web Modeling Language (WebML): A Modeling Language for Designing Web Sites, http://www10.org/paper-sample/ html-sample.htm, 14 January 2001.

Chulskis, K. (2002). The ebXML Registry, **XML Journal**, 3(1), pp 10.

Dejesus, E. (2000). XML Enters The DBMS Arena. http://www.computerworld. com/cwi/story/0,1199,NAV47_STO53026,00.html, 28 May 2001.

Enterprise Web Machines Corporation (2002). RADIX Methodology, http://www.enterprisewebmachines.com/RADIXMethodology.pdf. 12 November 2002.

Farinetti, L., Bota, F. and Rarau, A. (2000). An Authoring Tool for Building Flexible On-line Courses Using XML, **GCA Conferences**, June 2000, http://www.infoloom. com/gcaconfs/WEB/paris2000/S09-01.HTM, 15 January 2001.

Graves, M. (2002). **Designing XML Databases**. Prentice Hall PTR, Upper Saddle River, NJ.

Halaschek, C. and Miller, J.A. (2003). Native XML Databases Today, **XML Journal**, 4(1), pp 27.

Hamscher, W. (1999). Building with XML, **GCA Conferences**, December 1999, http://www.infoloom.com/gcaconfs/WEB/philadelphia99/Hamsher.HTM, 30 January 2000.

Hay, D. (1999). OK, So What is This XML Thing?, http://www.ies.aust.com/ ~visible/ papers/xml-hay.htm, 14 November 2000.

Howard, P. (2003). The Demise of the XML Database, **IT-Director.com**, October 2003, http://www.it-director.com/article.php?articleid=11287, 14 January 2004.

IBM. DB2 XML Extender. http://www-4.ibm.com/software/data/db2/extenders/ xmlext.html, 10 November 2001.

IBM. Informix Web DataBlade Module. http://www-4.ibm.com/software/data/ Informix/blades/web/, 10 November 2001.

IBM. Object Translator. http://www.informix.com/idn-secure/webtools/ ot/, 10 November 2001.

Johnston, S.J. (2002). Database Strategies for Unstructured Content, **XML & Web Services Magazine**, 3(3), pp 18-27.

Kroenke, D.M. (1998). **Database Processing Fundamental, Design and Implementation**. Prentice Hall, Inc, New Jersey.

Kumar, S. (1999). An Introduction to XML, http://www.devshed.com/Client_Side/ XML/Introduction/page4.html, 27 December 2000.

MahirNet (2001). MahirNet Maps E-Learning, **IT Malaysia Executive**, pp 6-7.

Milbery, J. (2001). Ipedo XML Database 2.0, **XML Journal**, 2(12), pp 32.

Oppel, K. (1998). Tamino: The Power Database for the Internet, SoftwareAG, The XML Company.

Oracle (1999). Oracle 8i interMedia. http://www.oracle.com/collateral/ o8i_intermedia_fo.pdf, 10 November 2001.

Oracle (2001). What's New in Oracle9*i* *inter*Media. http://technet.oracle. com/ products/intermedia/pdf/whatsnew_wp.pdf, 10 November 2001.

Patton, T. (2002). XML Server: An Overview, **XML Journal**, 2(8), pp 17.

Quin, L., (2000). **Open Sources XML Database Toolkit**. John Wiley & Sons, Inc, Canada.

Rice, F. C. (2001). Exploring XML and Access 2002. Microsoft Corporation. http://msdn.microsoft.com/library/default.asp?url=/library/en-us/dnacc2k2/_html/ odc_acxmllnk.asp, 10 November 2001.

Rosen, D., (2000). An Extensible Model for Real-Time XML Processing. **GCA Conference**, June 2000, http://www.infoloom.com/gcaconfs/WEB/paris2000/S12-01.HTM, 15 January 2001.

Roy, J. and Ramanujan, A. (2000). Building and XML Application, **GCA Conferences**, June 2000, http://www.infoloom.com/gcaconfs/WEB/paris2000/ S02-01.HTM, 15 January 2001.

Salminen, A. and Tompa, F. (2001). Requirements for XML Document Database Systems. **First ACM Symposium on Document Engineering**, November 2001.

Simon, S. H. (2001). **Emerging Business Technology Series: XML**. McGraw-Hill, New York.

Socket, A. and Wawrzyniak, L. (2000). Experiences of an implementation, **GCA Conferences**, June 2000, pp 241, http://www.infoloom.com/gcaconfs/WEB/ paris2000/ S05-03.HTM, 15 January 2001.

Software AG. http://www.softwareag.com/tamino/, 16 January 2001.

The dbXML Group. http://www.dbxml.com/overview.xml, 10 November 2001.

Tian, F.,DeWitt, D.J.,Chen, J. and Zhang, C. (2000). The Design and Performance Evaluation of Alternative XML Storage Strategies (submitted for publication), http://citeseer.nj.nec.com/tian00design.html, 27 August 2001.

Valduriez, P. and Gardarin, G. (1989). **Analysis and Comparison of Relational Database Systems**. Addison-Wesley Publishing Company, Inc, USA.

Watson, G. (2003). Using XML Schemas and DTDs Together, **XML Journal**, 4(11), pp 32.

Whiting, R. (2002). Excelon Adds XML Content-Management Capabilities, Information Week, February 2002, http://www.informationweek.com/story/showArticle.jhtml?articleID=6501163, 12 March 2003.

Willebeek-LeMair, J. (2001). Low-Cost, Flat File XML for the Masses, **XML Conference & Exposition 2001**, December 2001, http://www.idealliance.org/ papers/xml2001/papers/html/04-05-03.html, 14 June 2002.

X-Hive Corporation. http://www.x-hive.com/, 10 November 2001.

XML Global. http://www.xmlglobal.com/prod/foundation/goxmldb.jsp, 10 November 2001.