Comparison of Native XML Databases and Experimenting with INEX

Petr Kolář and Pavel Loupal

Dept. of Computer Science and Engineering
FEE, Czech Technical University
Karlovo náměstí 13, 121 35 Praha 2
Czech Republic

kolarp3@fel.cvut.cz, loupalp@fel.cvut.cz

DATESO 2006

Introduction – main goals

- summarize and compare approaches of design and architecture of native XML databases
- utilize the INEX dataset in several open source database systems (in this case only eXist and Apache Xindice)
- basic performance comparison outlined as a basis for discussion about suitability for particular database system

XML DB Products

- XML-Enabled Products
- Native XML Products
- Hybrid Products

Some of NXD

- Open-source

Product	Developer	DB Type
Berkeley DB XML	Sleepycat Software	Key-value
dbXML	dbXML Group	Proprietary
eXist	Wolfgang Meier	Relational
ozone	ozone-db.org	Object-oriented
Sedna XML DBMS	ISP RAS MODIS	Proprietary
Timber	University of Michigan (non-commercial only)	Shore, Berkeley DB
Xindice	Apache Software Foundation	Proprietary (Model-based)

- Commercial

Product	Developer	DB Type
Birdstep RDM XML	Birdstep	Object-oriented
eXtc	M/Gateway Developments Ltd.	Post-relational
Ipedo	Ipedo	Proprietary
Natix	data ex machina	File system(?)
Neocore XMS	Xpriori	Proprietary
Tamino	Software AG	Proprietary (+ODBC)
X-Hive/DB	X-Hive Corporation	Proprietary. (+JDBC)
XStreamDB Native XML Database	Bluestream Corp.	Proprietary (Model-based)
Xyleme Zone Server	Xyleme SA	Proprietary

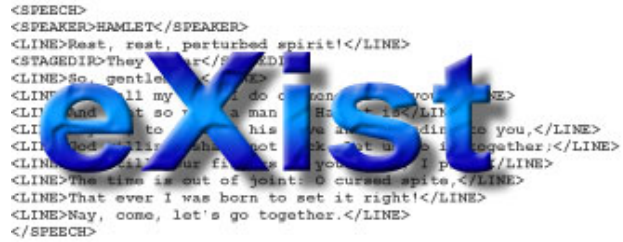
http://inex.is.informatik.uni-duisburg.de/

INEX dataset

- INitiative for the Evaluation of XML retrieval
- INEX data set (we use version 1.4) has 536MB of XML data. It is exactly 12,107 articles from 6 IEEE transactions and 12 journals from years 1995 to 2002
- In average each article contains 1,532 XML nodes
- The average depth of node is 6.9

http://exist.sourceforge.net/

eXist



- eXist XML database version 1.0-dev-20060124
  - Developed in Java, opensource
  - Supported Platforms: Platform independent
  - Data Storage: B+-trees and paged files. Document nodes are stored in a persistent DOM– No support for binary files
  - Transaction Support: No
  - Authorization:Unix like, permissions at collection and document level
  - XML Standards that are supported:
    - XPath/XQuery through Xquery engine
    - XUpdate
    - Xinclude/Xpointers
    - API: XML:DB
  - Comes with great client GUI interface
  - Types of indexes: Structural, Fulltext, Range

http://xml.apache.org/xindice/

Xindice

- Xindice XML database version 1.0 (birthday)
  - Developed in Java, opensource
  - Supported Platforms: Platform independent
  - Data Storage: Natively as indexed text files.
    - Collections as directories on file system
    - Documents in a collection as compressed text files(.tbl files); Hoffman codes.
  - No support for binary files
  - Transaction Support: No
  - Authorization:No support
  - Supported XML Standards:
    - XPath
    - XUpdate
    - AutoLinking
    - API: XML:DB, command line,
  - Unsupported XML standards: Xpointers, XQL, XQuery
  - No GUI available

Xindice vs. Exist

Feature	eXist	Xindice
Technology	Java	Java
Data storage	B+-trees and paged files. Persistent DOM	Natively as indexed text files, Hoffman codes
Binary files	No	No
Transaction Support	No	No
Authorization	Unix like, permissions at collection and document level	No Support
Supported Standards	XPath/XQuery, XUpdate, Xinclude/XPointer	XPath, XUpdate, AutoLinking
APIs	XML:DB	XML:DB, command line
Client GUI	Yes	No
Indices	Structural, Fulltext, Range	

Experiment

- We prepared set of XPath queries in following categories:
  - Selecting nodes (i.e. */article/fm/hdr/hdr1/crt/issn*)
  - Predicates (i.e. */article/bdy/sec[last() − 1]*)
  - Selecting Unknown Nodes (i.e. */*/*[@*]*)
  - Selecting Several Paths (i.e. *//article/fm/hdr* | *//article/bdy/sec*)
- We measured time needed to perform the each prepared query on Xindice and Exist on the same hardware

Results 1

No.	Query	Records retrieved	Query duration time [s] eXist	Xindice
1	$/article$	12104	1,3	230
2	$/article/fm/hdr/hdr1/crt/issn$	11666	2,2	98
3	$//issn$	11666	1,3	447
4	$/article/bdy/sec[1]$	11955	1,9	NA
5	$/article/bdy/sec[last()]$	11955	5,6	NA
6	$/article/bdy/sec[last()-1]$	11019	5,8	NA
7	$/article/bdy/sec[position()<3]$	22974	8,1	NA
8	$//sec[@type]$	868	1,0	more than 10 min
9	$//sec/p/ref[@type='bib']$	108496	81,3	NA
10	$/article/fm/hdr/hdr2/pdt[yr='1995']$	1623	2,6	NA
11	$/article/fm/hdr/hdr2/pdt[yr='1995'and mo='Spring']$	72	4,0	NA
12	$/article/*$	58472	164,3	NA
13	$/*/*[@*]$	49	352,0	NA
14	$//fig[@*]$	52857	70,6	NA
15	$//article/fm/hdr\|$ $//article/bdy/sec$	77487	8,6	NA
16	$//article/fm/hdr/hdr1\|$ $//article/fm/hdr/hdr2$	24208	3,8	NA

Results 2

- The time needed to load INEX data set into database:
  - 25 minutes for Xindice
  - 97 minutes for eXist
- The data on filesystem took:
  - 600 MB for Xindice
  - 1300 MB for eXist

- Our hardware configuration was based on a personal computer with Intel Celeron 1.7 Ghz processor, 512MB RAM and Windows XP(SP2) operating systém
- INEX XML data set in version 2003 (1.4)

Summary

- Xindice has totally failed in our experiments probably due to index malfunction (but Xindice looks like that Indexes are working)
- Most of XPath queries running over Xindice returned an empty result set.
- On the contrary, eXist showed much better behavior.
- Automatically generated structural index in eXist that is very efficient
- eXist has also an user friendly GUI for both database management and ad-hoc query processing

Conclusion 1

- The aim of our experiment was in principle not successful
- We were not able to import the INEX dataset into all proposed native XML databases
- Our results show that for further experiments we should consider only the **eXist** database
- Xindice can be used just as an example of a basic native XML database, for large data set is not usable
  - At this moment is available Xindice Version 1.1b4

Conclusion 2

- It is needed to perform further comparisons among other native XML databases
- Also, we plan to add some of non-native (or hybrid) XML databases.

The end of the poster presentation