1) Module: SEDNA XML DATABASE

Draft: 12/09/2010

1) Module Name

SEDNA XML DATABASE

2) Scope

The module introduces the use of SEDNA xml database for xml retrieval. The primary focus of the module is to describe the architecture of SEDNA database and how standard xml queries can be used to retrieve data from it.

3) Learning Objectives

At the end of the module, a student will be able to

i. Explain the basic concepts of xml retrieval and its relevance in IR systems.

ii. Explain the architecture of SEDNA.

iii. Load xml documents into the SEDNA database and perform information retrieval on the documents using xml queries.

4) 5S Characteristics of the Module

.

Ss	Examples	Objectives
Streams	XML, XQueries	Describes properties of the DL content such as encoding and language for textual material or particular forms of multimedia data
Structures	Collection, documents, Entities, Attributes, DTD, metadata.	Specifies organizational aspects of the DL content
Spaces	Xquery operations	Defines logical and presentational views of several DL components
Scenarios	Searching, Querying, updating, inserting, deleting	Details the behavior of DL services
Societies	Database administrators,	Defines managers, responsible

	database users, web users	for running DL services; actors, that use those services; and relationships among them

5) Level of Effort Required:

i. Prior to Class: 2 hours of reading

ii. In Class: 3 hours

a. 2 hours for learning the basics of xml and implementation of xml retrieval in SEDNA

b. 1 hour for class discussions, exercises, and activities

6) Relationship with other modules:

Close connections with:

a) WEKA: XML support is now available in several places in WEKA:

- Command Line
- Serialization of Experiments
- Serialization of Classifiers
- Bayesian Networks
- Tools

b) SOLR: SOLR package is also used to query content and retrieve information from xml documents.

7) Prerequisite knowledge required:

Students need to know basic concepts of UNIX commands, XML, XML database, and XQuery.

8) Introductory Remedial Instruction:

None

9) Body of Knowledge

i. What is XML?

a. XML stands for extensible Markup Language.

b. A markup language is used to provide information about a document.

c. Tags are added to the document to provide the extra information.

d. HTML tags tell a browser how to display the document.

e. XML tags give a reader some idea what some of the data means.

.

ii. Difference between XML and HTML

a) XML is not a replacement for HTML.

b) XML was designed to describe data and to focus on what data is. HTML was designed to display data and to focus on how data looks.

c) HTML is about displaying information, XML is about describing information.

.

iii. Why Is XML Important?

a) Plain Text

- Easy to edit.
- Useful for storing small amounts of data .
- Possible to efficiently store large amounts of XML data through an XML front end to a database.

b) Data Identification

- Tell you what kind of data you have
- Can be used in different ways by different applications

c) Inline Reusability

- Can be composed from separate entities
- Modularize your documents without resorting to links

d) Hierarchical

- Faster to access .
- Easier to rearrange.

.

iv. Why XML evolved?

a) 1960-1980 Infrastructure for the Internet.

b) 1986 SGML (Standard Generalized Markup Language) for defining and representing structured documents.

c) 1991 WWW and HTML introduced for the Internet.

d) 1991 Business adopts the WWW technology; huge expansion in the use of the Internet.

e) 1995 New kinds of businesses evolve, based on the connectivity of people all over the world and connectivity of applications built by various software providers (B2C, B2B).

f) Urgent need for a new, common data format for the Internet .

v. XML Building blocks

a) Element

- Delimited by angle brackets.
- Identify the nature of the content they surround.
- General format: <element> … </element>
- Empty element: </empty-Element>

b) Attribute

- Name-value pairs that occur inside start-tags after element name, like: <element attribute="value">

vi. Example of an XML Document

```
<?xml version="1.0"/>

<address>
        <name>John </name>
        <email>team4@vt.edu</email>
        <phone>212-346-1234</phone>
        <birthday>1985-03-22</birthday>
</address>
```

vii. XML Trees

a) An XML document has a single root node.

b) The tree is a general ordered tree.
    a. A parent node may have any number of children.

    b. Child nodes are ordered, and may have siblings.

c) Preorder traversals are usually used for getting information out of the tree.

viii. Document Type Definitions

a) A DTD describes the tree structure of a document and something about its data.

b) There are two data types, PCDATA and CDATA.

    a. PCDATA is parsed character data.

    b. CDATA is character data, not usually parsed.

c) A DTD determines how many times a node may appear, and how child nodes are ordered.

ix. XML Database

a) An XML database is a data persistence software system that allows data to be stored in XML format. This data can then be queried, exported and serialized into the desired format.

b) Two major classes of XML database exist:

    a. XML-enabled: these map all XML to a traditional database (such as a relational database [1]), accepting XML as input and rendering XML as output. This term implies that the database does the conversion itself (as opposed to relying on middleware).

    b. Native XML (NXD): the internal model of such databases depends on XML and uses XML documents as the fundamental unit of storage, which is, however, not necessarily stored in the form of text files.

x. **What is XQuery?**

a) XQuery is designed to be a small, easily implemental language in which the queries are easily understood.

b) It is also flexible enough to query a broad spectrum of XML information sources, including both databases and documents.

c) It is human-readable query syntax and an XML based query syntax.

xi. ABOUT SEDNA

a) Sedna is a free native XML database which provides a full range of core database services.
b) It provides persistent storage, ACID transactions, security, indices, hot backup.
c) Provides support for fine-grained XML triggers.
d) Provides sql connection from XQuery .
e) Provides XQuery external functions implemented in C.
f) Provides database security (users, roles and privileges).

xii. Sedna Architecture Overview

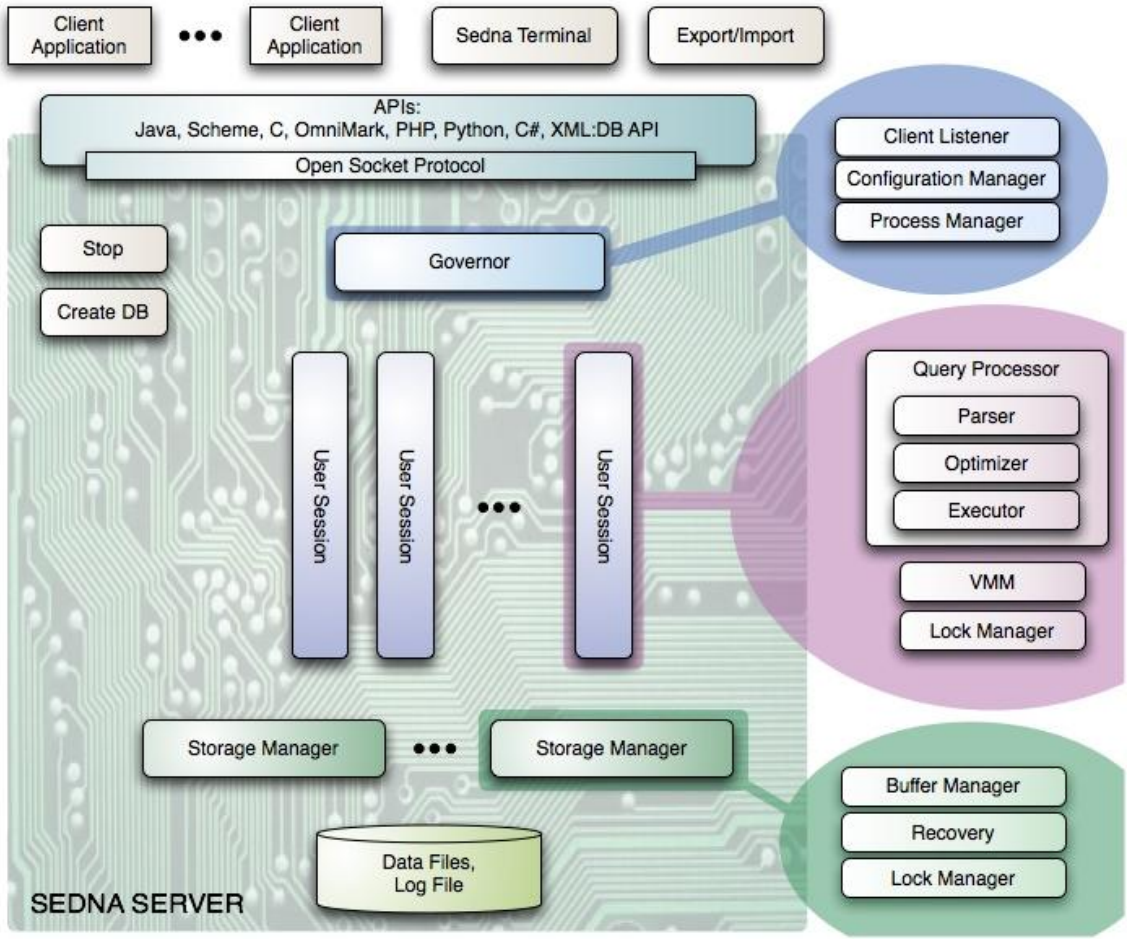


Figure 1: SEDNA Architecture Overview

xiii. Sedna Architecture: Interactions between processes

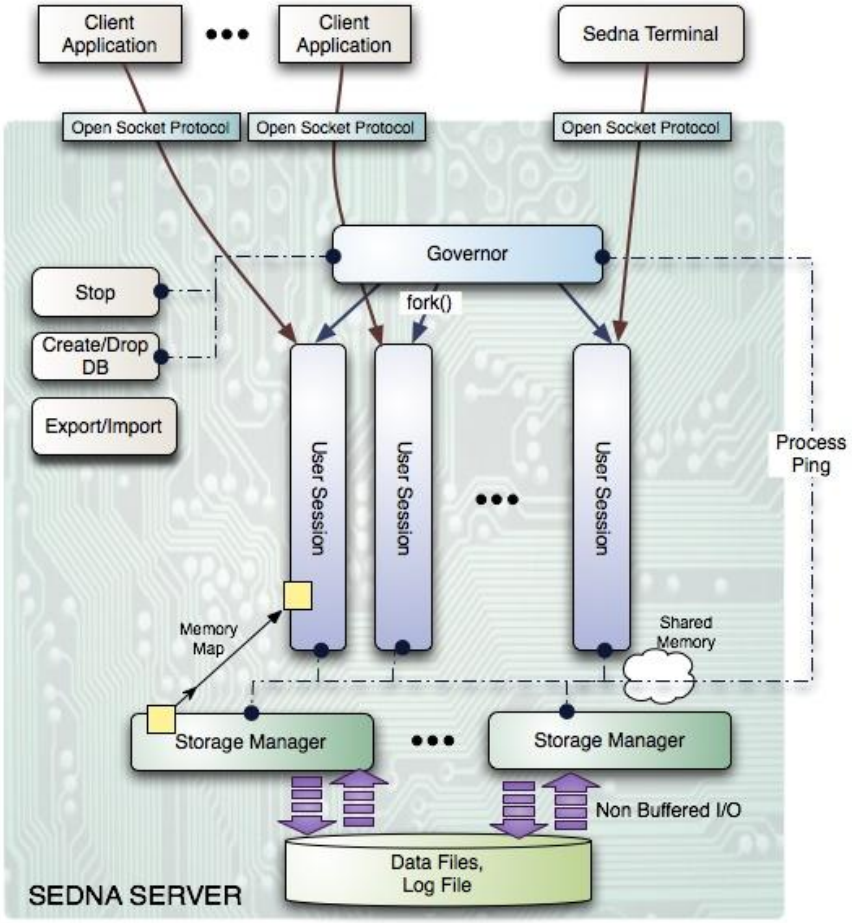


Figure 2: Sedna Architecture: Interactions between processes

xiv. Sedna Native XML Database Client/Server Protocol

a) Sedna XML Database server uses a message-based protocol for communication with clients through the TCP/IP sockets.

b) The common message structure is as follows: the first four bytes (int) is instruction, the next four bytes (int) is the length of a body in bytes; the next 'length' bytes is the body.

c) To begin a session, a client creates a connection to the server and sends a startup message. The server launches a new process that is associated with the session.

d) Queries are executed via different subprotocols depending on the type of the query and the query length. There are three types of queries: query, update, bulk load.

e) Termination can be initiated by the client (for example when it closed the session) or by the server (for example in case of an administrator-commanded database shutdown or some failure).

xv. Sedna Query Parser & Optimizing Rewriter

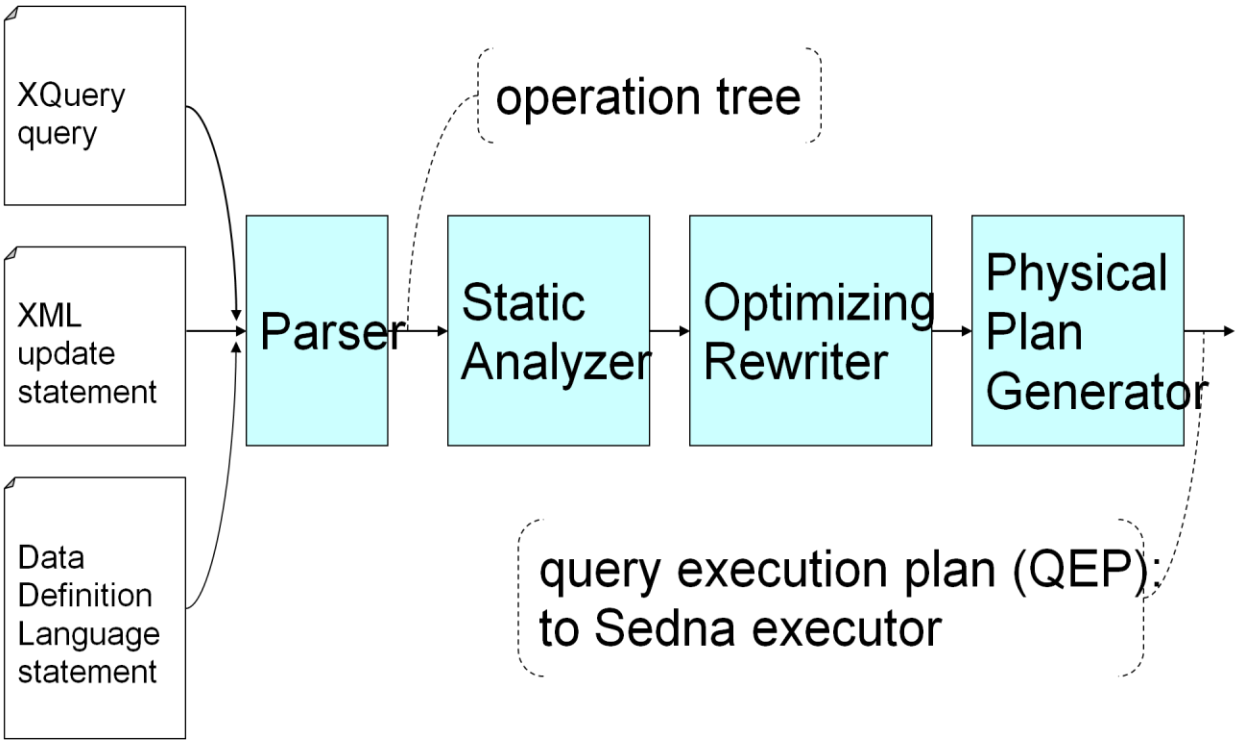


Figure 3: Sedna Query Parser & Optimizing Rewriter

xvi. Sedna API overview

a) APIs developed by our team:
   - C
   - Java
   - Scheme
   - OmniMark *(Stilo's streaming programming language used for content engineering tasks)*

b) APIs contributed by Sedna open source users:
   - Python
   - PHP
   - .Net

- XML:DB API *(standard API for XML databases, supported by other products also)*

xvii. INSTALLING SEDNA

a) The installation guide can be found at http://modis.ispras.ru/sedna/install.html#bininst

b) The following steps need to be followed.
   - Download the self extracting script from http://www.modis.ispras.ru/FTPContent/sedna/current/sedna-3.4.66-bin-linux-x86.sh .
   - Make the script executable by changing its permission using:
     chmod +x sedna-3.4.66-bin-linux-x86.sh
   - Execute the script by running the following command on the terminal in the directory in which the script was downloaded:
     ./sedna-3.4.66-bin-linux-x86.sh
   - The operating system user that is going to run Sedna must have r-w-x permissions for the following Sedna directories:
     $SEDNA_INSTALL/data
     $SEDNA_INSTALL/cfg (here $SEDNA refers to the directory in which Sedna was installed)
   - To grant the necessary permissions on linux execute the following command on the terminal.
     chown <sedna-user> cfg data
   - Sedna actually needs to run 50 sessions (default maximum) simultaneously. To do this some of the kernel parameter settings have to be extended. This can be done as under.
     - Log on as a user with root authority
     - Open up /etc/sysctl.conf in a text editor and add entries:
       kernel.sem = "250 64000 32 256"

xviii. WORKING WITH SEDNA

a) Go to INSTALL_DIR/bin (Here INSTALL_DIR refers to the directory in which Sedna was installed)

b) To start Sedna server run:
   ./se_gov
   (If Sedna is started successfully it prints "GOVERNOR has been started in the background mode“)
c) To shutdown Sedna server run:
   ./ se_stop

d) To create a database named testdb:
   ./se_cdb testdb
   (If the database is created successfully it prints "The database 'auction' has been created successfully“)

e) To run the testdb database:
   ./se_sm testdb
   (If the database is started successfully it prints "SM has been started in the background mode".)

f) To shutdown the testdb database:
   ./se_smsd testdb

xix. WORKING WITH SEDNA(RUN QUERIES)

a) The directory INSTALL_DIR/examples/commandline has an xml file named auction.xml and some sample xml queries in files named sample01.xquery to sample10.xquery.

b) The directory also contains a script which uploads an xml file into the data base named load_data.xquery.

c) Load the sample XML document into the auction database by typing the command in the directory:INSTALL_DIR/bin :
   ./se_term -file INSTALL_DIR/examples/commandline load_data.xquery testdb
d) Now you can execute the sample queries by typing the command:
   ./se_term -file INSTALL_DIR/examples/commandline /<query_name>.xquery testdb
   (where <query_name>.xquery is the name of a file with the sample query).
e) For instance, to execute sample01.xquery you should type:
   ./se_term -file INSTALL_DIR/examples/commandline /sample01.xquery testdb

xx. DEMO

a) Log in to the cloud instance 129.33.147.20 using the key provided and username idcuser.
b) Sedna has been installed in the xml directory present in /home/idcuser.

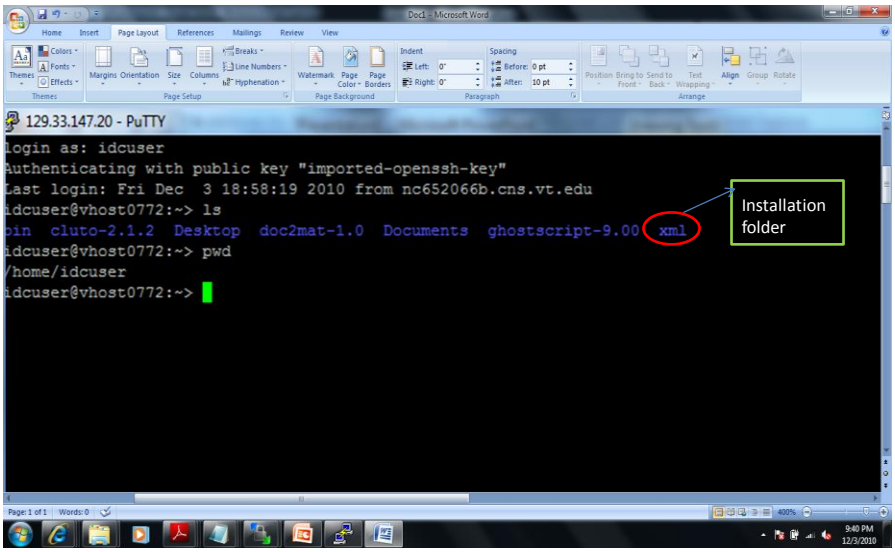


Figure 4: SEDNA DEMO part 1

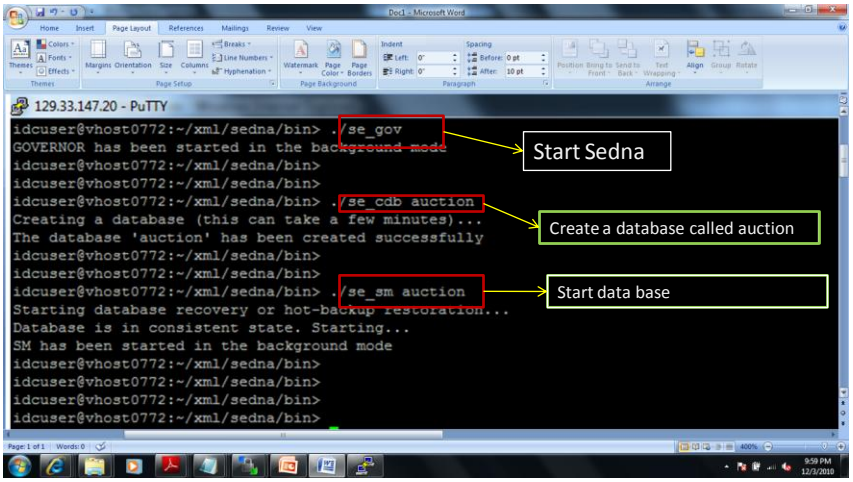


Figure 5: SEDNA DEMO part 2

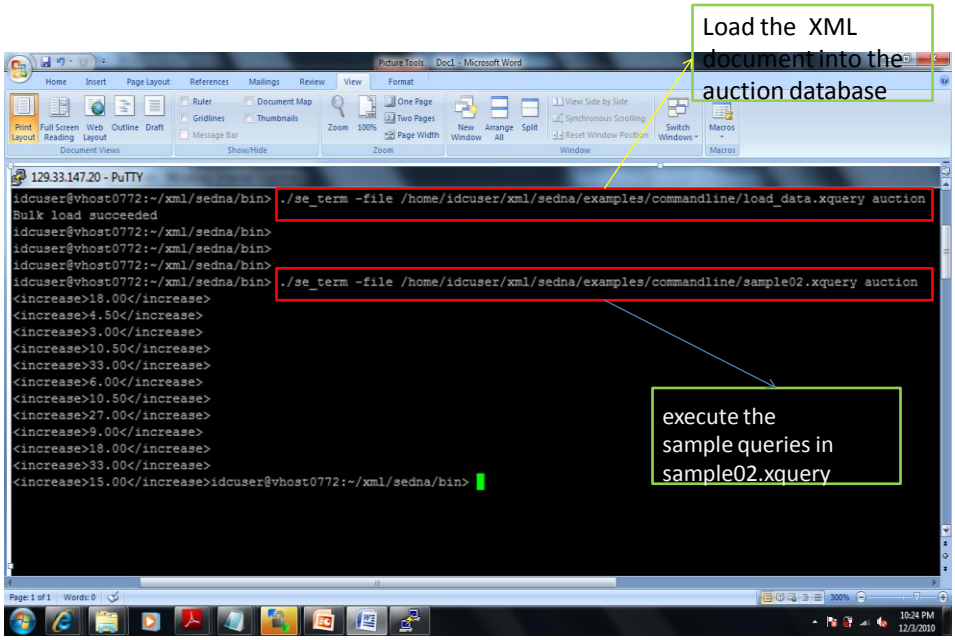


Figure 6: SEDNA DEMO part 4

xxi. DEMO:C API

a) Library files are located at the SEDNA_DIR/driver/c folder, where SEDNA_DIR is the directory Sedna is installed in.

b) Examples discussed in this document can be found in SEDNA_DIR/examples/api/c.

c) To build examples use:
gcc -ISEDNA_DIR/driver/c -osample sample.c SEDNA_DIR/driver/c/libsedna.a

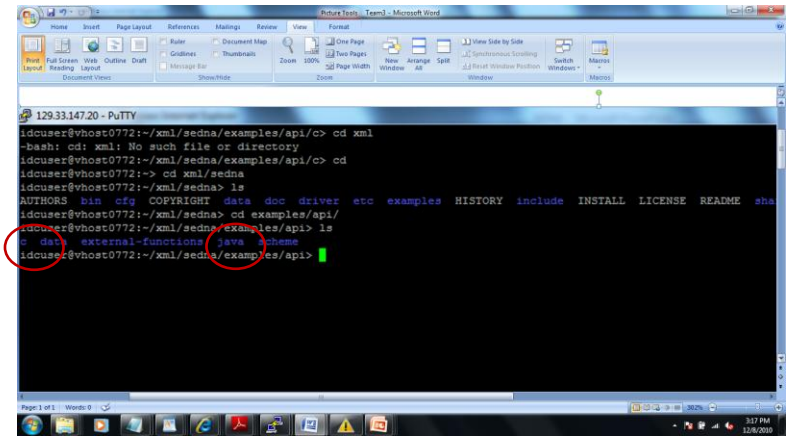


Figure 7: C API DEMO part 1

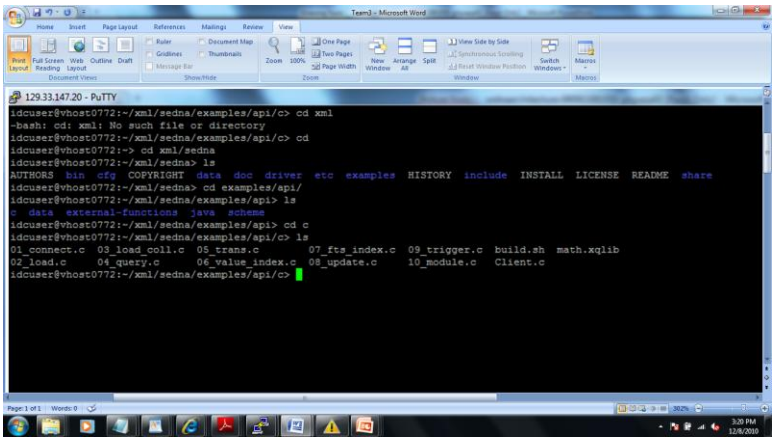


Figure 8: C API DEMO part 2

xxi. How it Works?

a) XML File

```
<site>
 <open_auctions>
  <open_auction>
   <bidder>
    ……
    <increase>15.00
    </increase>
   </bidder>
   <bidder>
    ….
    <increase>20.00
    </increase>
   </bidder>
  </open_auction>
  <open_auction>
    …..
  </open_auction>
</open_auctions>
```

</site>

b) **Xquery**

```
(:
  Return the initial increases of all open auctions.
:)
for $b in doc("auction")/site/open_auctions/open_auction
return <increase>{ $b/bidder[position()=1]/increase/text() }</increase>
```

c) Result

```
<increase>18.00</increase>
<increase>4.50</increase>
<increase>3.00</increase>
<increase>10.50</increase>
<increase>33.00</increase>
<increase>6.00</increase>
<increase>10.50</increase>
<increase>27.00</increase>
<increase>9.00</increase>
<increase>18.00</increase>
<increase>33.00</increase>
```

xxii. WikiXMLDB: Query Wikipedia with XQuery

WikiXMLDB provides a way of querying Wikipedia with Xquery.

a) Implementation:

- Entire English Wikipedia content is parsed into XML representation (total size – about 36 GB).
- Loaded into Sedna XML database.
- A query interface is provided.

b) Deploying WikiXMLDB

- WikiXMLDB demo is deployed on Amazon EC2 and runs on the virtual computer with restricted resources.

- Can be run on local computer to achieve better performance and unlimited customization.
  http://wikixmldb.org/help/deployment/

c) Uses of WikiXMLDb

- User can dissect individual articles, rip out abstracts, sections, links, info boxes and other components of Wikipedia.

- User can combine pieces of existing documents into new XML documents and convert them to web pages with XSLT for example.

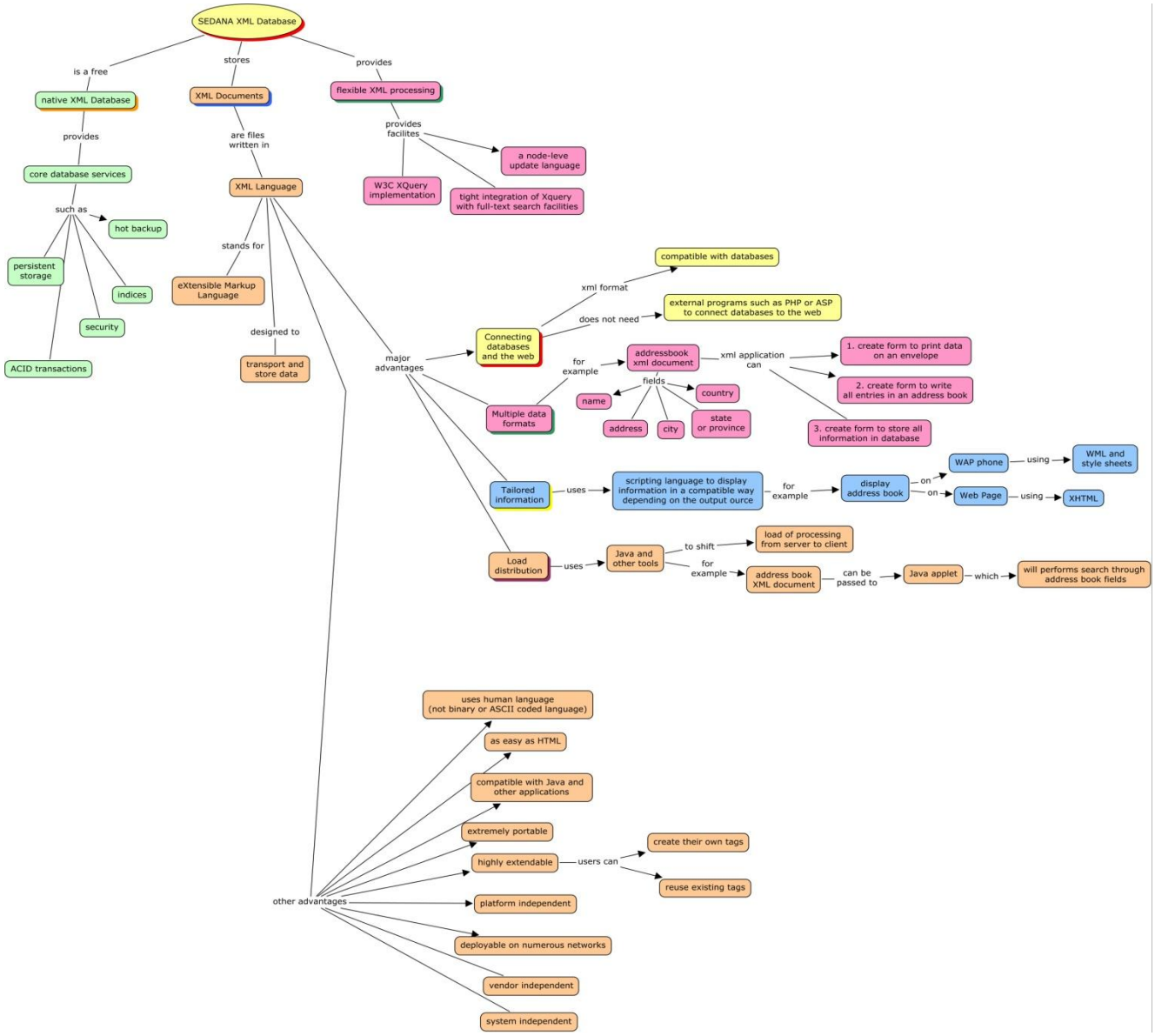
- User can enrich his content with data from Wikipedia and unlock its power for his applications.

- User can perform all these actions using W3C Xquery Language.

d) WikiXMLDB Demo

- XML representation of Wikipedia articles:
  http://wikixmldb.org/help/getting-started
- WikiXMLDB Xquery interface:
  http://wikixmldb.org/xq/

10) Concept Map



11) Exercises/Learning Activities

1. **Write Xqueries for the following scenarios at wikiXMLDB demo page:** http://wikixmldb.org/xq/
   a) Retrieve all titles of Wikipedia pages which contain the term “XML”
   b) Retrieve the categories of the Wikipedia page whose title is “Igor Kurchatov”
   c) Retrieve the contributors of all Wikipedia pages whose titles contain “XQuery”

2. Use the auction.xml file described in the demo. Write Xqueries for the following scenarios:
    a) Count the number of open auctions
    b) Count the number of bidders in the first open auction
    c) Find the date of the first bidder in the last open auction

12) Evaluation of Learning Outcomes

i. Are students familiar with the fundamental concepts, applications and techniques of XML retrieval?
ii. Are students capable of running xqueries on the SEDNA database?
This can be determined by verifying if a student can successfully load an xml document into the SEDNA database, and if he can write and execute xqueries to retrieve data from the xml document.

13) Glossary

i. **Xml**: Extensible Markup Language (XML) is a set of rules for encoding documents in machine-readable form.

ii. **Xquery:** It is a query and functional programming language that is designed to query of XML data.

iii. **Information Retrieval:** Field of computer science which studies the retrieval of information (not data) from a collection of written documents.

iv. **Database:** A database consists of an organized collection of data for one or more uses, typically in digital form.

v. **HTML:** HTML, which stands for HyperText Markup Language, is the predominant markup language for web pages.

vi. **Parser:** Parser performs the process of analyzing a text, made of a sequence of tokens (for example, words), to determine its grammatical structure with respect to a given (more or less) formal grammar.

vii. **API:** An Application Programming Interface (API) is a particular set of rules and specifications that a software program can follow to access and make use of the services and resources provided by another particular software program that implements that API.

viii. **Linux:** Linux refers to the family of Unix-like computer operating systems using the Linux kernel.

14) References

i. Weblinks:

[1] Home Page: http://modis.ispras.ru/sedna/index.html

[2] Download Page: http http://modis.ispras.ru/sedna/download.html

[3] Programmers Guide: http://modis.ispras.ru/sedna/progguide/ProgGuide.html

[4] XQuery Help: http://www.w3.org/XML/Query/

ii. Related research publications:

[5] Overview of the INitiative for the Evaluation of XML Retrieval (INEX) 2003 Saadia Malik University of DuisburgEssen, Germany Mounia Lalmas Queen Mary University of London, United Kingdom.

[6] Cheshire II at INEX: Using A Hybrid Logistic Regression and Boolean Model for XML Retrieval Ray R. Larson School of Information Management and Systems University of California Berkeley, California, USA, 947204600

[7] Content-oriented XML retrieval with HyREX Norbert Gövert Mohammad Abolhassani Norbert Fuhr Kai Großjohann University of Dortmund University of Duisburg Germany Germany

[8] A scalable architecture for XML retrieval Gabriella Kazai Thomas R¨olleke Department Computer Science Queen Mary University London.

[9] An Appropriate Unit of Retrieval Results for XML Document Retrieval ¤ Kenji Hatano Nara Institute of Science and Technology 89165 Takayama, IkomaNara 6300192, Japanhatano@is.aistnara.ac.jp Hiroko Kinutani Nara Institute of Science and Technology 89165 Masahiro Watanabe National Institute of Special Education 511 Nobi, Yokosuka Kanagawa 2390841, Japan masahiro@nise.go.jp

15) Contributors:

Authors:

CS_5604: Information Storage and Retrieval- Team 4:

i. Vijay, Sony
ii. El Meligy Abdelhamid, Sherif
iii. Malayattil, Sarosh

Reviewers:
Dr. Edward Fox