# MAX-PLANCK-INSTITUT FÜR WISSENSCHAFTSGESCHICHTE Max Planck Institute for the History of Science

Max Planck Digital Library (MPDL)

Subproject: Content based web access

Josef Willenborg

## **Technical Specification**

Version	0.2
Author	Josef Willenborg
Created	17.09.2008
Last modified	17.09.2008
Last modified by	Josef Willenborg

## Content

1.	ARCHITECTURE AND DESIGN	3
2.	SOFTWARE	3
3.	HARDWARE	5
4	PROJECT ORGANISATION	5

### 1. Architecture and design

has to be done

#### 2. Software

For fulfilling the requirements the following software will be used:

#### Maintenance of document bases

ESciDoc could be used as the central document base system (intermediate service could be used: ???). All Create/Update/Delete-Operations of XML-documents have to trigger the same operation to the indexing software in some way (???).

#### Indexing and querying (basic system)

Functional comparision of Lucene, eXist and Oracle (other software such as Tamino, DB2, Postgres could be examined later if needed):

Requirement	Lucene	eXist	Oracle Standard Edition One
Open software	+	+	-
Price	free	free	+ (development and production: 654 Euro for 2 years)
Customizable (Sources extensible for specific needs)	++	-	-
Easy maintenance and usage	++	+	+
Powerful in functionality	+	+	++
Scalable, platforms	+	+	++
Fulltext querying for XML-docs (general)	+	+ (slower regular expressions)	++ (but no regular expressions)
Logical operators: And, or, andNot	+	+	+
Wildcard querying: * (left, middle,right in the word)	+	+	++
Wildcard querying: _ (one character), % (some characters)	-	+	+
Stemming in different languages	+	-	++
Stemming extensible with language specific dictionary	+ (Java programmable	-	++ (build in)

Writing similarity (fouzey) + + (Ngrams) + Phonetic similarity (soundex) + + Near operator + + + + + + + + + + + + + + + + + + +	Date range queries	+	+	+
Phonetic similarity (soundex)  Near operator  Near operator  + Near operator  + Case sensitive querying + (with filter) + + + + + + + + + + + + + + + + + + +	•	+	+ (Ngrams)	+
Near operator		-	-	+
Case sensitive querying Thesaurus searching Thesaurus searching for XML-documents by attributes (e.g. author, title) Structural querying of XML-documents Structural querying in one XML-document (also without index) Structural querying in one XML-document (also without index) Structural querying in one XML-document (also without index)  Content encoding in UTF8 (query and content) Query results are customizable  ++ (Java)  Content encoding in UTF8 (query and content)  Query results could be sorted by relevance Query results could be sorted alphabtically by fields (author, publication year) Datastore for specific languages Datastore content as file, URL and database cell Storing of relational data (for specific needs)  Web interfaces  H (JSP)  Build in WebServer with Servlet-Container for dynamic SQL and XQuery yeacution (also WebServices)  Maximal size of XML-document  Maximal number of XML-documents  Pathology in the file on the file system, but system hangs if it is too big (> 100 MB)  Maximal number of XML-documents  Publishing Protocol  Publishing Protocol  Publishing Protocol  Publishing Protocol  Publishing Protocol  Protocol  Publishing Protocol  Publishing Protocol  Publishing Protocol  Protocol  Publishing Protocol  Publishing Protocol  Publishing Protocol  Publishing Protocol  Protocol  Publishing Prot		+	+	+
Thesaurus searching Searching for XML-documents by attributes (e.g. author, title) Structural querying of XML-documents Structural querying in one XML-document (also without index)  Content encoding in UTF8 (query and content) Query results are customizable Query results could be sorted by relevance Query results could be sorted alphabtically by fields (author, publication year) Datastore for specific languages Datastore content as file, URL and database cell Storing of relational data (for specific needs) Web interfaces  H (JSP)  Build in WebServer with Servlet-Container for dynamic SQL and XQuery execution (also WebServices)  Maximal size of XML-document  Maximal number of XML-documents  A (H)	•	+ (with filter)	+	+
Searching for XML-documents by attributes (e.g. author, title)  Structural querying of XML-documents  Structural querying in one XML-document (also without index)  Structural querying in one XML-document (also without index)  Content encoding in UTF8 (query and content)  Query results are customizable		, , , , , , , , , , , , , , , , , , , ,	_	
attributes (e.g. author, title) Structural querying of XML- documents Structural querying in one XML- document (also without index) Structural querying in one XML- document (also without index)  Content encoding in UTF8 (query and content)  Query results are customizable  ++ (Java)  Content encoding in UTF8 (query and content)  Query results could be sorted by relevance  Query results could be sorted alphabtically by fields (author, publication year)  Datastore for specific languages  Datastore content as file, URL and database cell  Storing of relational data (for specific needs)  Web interfaces  ++ (JSP)  String of relational data (for specific needs)  Web interfaces  ++ (JSP)  ++ (JSP, SOAP with XPath/XQuery, REST, WebDAV, XMLRPC, Atom Publishing Protocol)  ++  Heoretically as big as the maximum size of a file on the file system; but system hangs if it is too big (> 100 MB)  Maximal number of XML-documents  Acquery execution  Maximal number of XML-documents  Acquery execution  Acquery execution (also  Meximal number of XML-documents  > 1.000.000  281			+	+ (within-
Structural querying of XML- documents  Structural querying in one XML- document (also without index)  Content encoding in UTF8 (query and content)  Query results are customizable  Query results could be sorted by relevance  Query results could be sorted alphabtically by fields (author, publication year)  Datastore for specific languages  Datastore content as file, URL and database cell  Storing of relational data (for specific needs)  Web interfaces  H+ (JSP)  Build in WebServer with Servlet- Container for dynamic SQL and XQuery secution (also WebServices)  Maximal size of XML-document  Maximal number of XML-documents  Programmable H+ (XQuery)  Programmable H+ (XQuery)  H+ (XPath, XQuery)  H+ (XPath, XQuery)  H+ (Java, SQL, XPath, XQuery)  H+ (YPath, XQuery)  H+ (Java, SQL, XPath, XQuery)  H+ (YPath, XQuery)  H+ (Java, SQL, XPath, XQuery)  H+ (File)  H+ (JSP, SOAP with XPath/XQuery, REST, WebDAV, XMLRPC, Atom Publishing Protocol)  H+ ++  H+ (Complex)  Heoretically as big as the maximum size of a file on the file system; but system hangs if it is too big (> 100 MB)  Maximal number of XML-documents  Maximal number of XML-documents  Programmable  H+ (XPath, XQuery)  XQuery)  H+ (XPath, XQuery)  H+ (Java, SQL, XPath, XQuery)  H+ (JSP)  H+ (File)  H+				`
Structural querying in one XML-document (also without index)   Programmable with Java and Xquery   ++ (XPath, XQuery)   ++ (XPath, XQ		_	++ (XPath	• /
Structural querying in one XML- document (also without index)  Programmable with Java and Xquery  Content encoding in UTF8 (query and content)  Query results are customizable  ++ (Java)  ++ (XPath, XQuery)  Query results could be sorted by relevance  Query results could be sorted alphabtically by fields (author, publication year)  Datastore for specific languages  Datastore content as file, URL and database cell  Storing of relational data (for specific needs)  Web interfaces  H (JSP)  Web interfaces  H+ (JSP, SOAP with XPath/XQuery, REST, WebDAV, XMLRPC, Atom Publishing Protocol)  Build in WebServer with Servlet-Container for dynamic SQL and XQuery execution (also WebServices)  Maximal size of XML-document  Maximal number of XML-documents  Programmable with XQuery  ++ (XPath, XQuery)  ++ (Java)  +- (Java)  +- (Java)  (Ja				
document (also without index) Xquery  Yanth, Xquery  Yanth, Xquery  Yanth, Xquery  Yanth, Xquery  Xquery  Xquery  Xquery  Yanth, Xquery  Yanth, Xquery  Yanth, Xquery  Xquery  Xquery  Xquery  Yanth, Xquery  Yanth, Xquery  Xquery  Xquery  Xquery  Yanth, Xquery  Yanth, Xquery  Xquery  Yanth, Xquery  Yanth, Xquery  Xquery  Xquery  Yanth, Xquery  Yanth, Xquery  Xquery  Xquery  Yanth, Xquery  Yanth, Xquery  Yanth, Xquery  Xquery  Yanth, Xquery  Yanth, Xquery  Xquery  Yanth, Xquery  Yanth, Xquery  Yanth, Xquery  Xquery  Yanth,		Programmable		
Content encoding in UTF8 (query and content)  Query results are customizable  ++ (Java)  Query results could be sorted by relevance  Query results could be sorted alphabtically by fields (author, publication year)  Datastore for specific languages  Datastore content as file, URL and database cell  Storing of relational data (for specific needs)  Web interfaces  H (JSP)  H+ (JSP, SOAP with XPath/XQuery, REST, WebDAV, XMLRPC, Atom Publishing Protocol)  Build in WebServer with Servlet-Container for dynamic SQL and XQuery yeacution (also WebServices)  Maximal size of XML-document  Maximal number of XML-documents  Navery  H (Java)  H+ (XPath, XQuery)  H+ (SP)  H+ (File)  H+ (Fil				
Content encoding in UTF8 (query and content)  Query results are customizable  ++ (Java)  ++ (XPath, XQuery)  Query results could be sorted by relevance  Query results could be sorted dalphabtically by fields (author, publication year)  Datastore for specific languages  Datastore content as file, URL and database cell  Storing of relational data (for specific needs)  Web interfaces  + (JSP)  Heigh He	document (dies without index)		/ (Quoi y)	/(Quoiy)
and content)  Query results are customizable ++ (Java) ++ (XPath, XQuery) XPath, XQuery)  Query results could be sorted by relevance  Query results could be sorted + + + (?) + + (?) + + (?) + - (?) + - (?) + - (?) + - (?) +	Content encoding in UTF8 (query	•	+	+
Query results are customizable     ++ (Java)     ++ (XPath, XQuery)     ++ (Java, SQL, XPath, XQuery)       Query results could be sorted by relevance     +     -     +       Query results could be sorted alphabtically by fields (author, publication year)     +     + (?)     +       Datastore for specific languages     +     +     +       Datastore content as file, URL and database cell     + (file)     + (file)     ++ (file, url, cell)       Storing of relational data (for specific needs)     + (JSP)     ++ (JSP, SOAP with XPath/XQuery, REST, WebDAV, XMLRPC, Atom Publishing Protocol)     + (JSP, SOAP with XPath/XQuery, REST)     ++ (SST)       Build in WebServer with Servlet-Container for dynamic SQL and XQuery execution (also WebServices)     -     ++ + + + + + + + + + + + + + + + + + +				
Query results could be sorted by relevance Query results could be sorted alphabtically by fields (author, publication year) Datastore for specific languages Datastore content as file, URL and database cell Storing of relational data (for specific needs) Web interfaces  H (JSP)  Build in WebServer with Servlet-Container for dynamic SQL and XQuery execution (also WebServices)  Maximal size of XML-document  Maximal number of XML-documents  A (Path, XQuery)  H (?)  H (File)  H	,	++ (Java)	++ (XPath	++ (Java SQI
Query results could be sorted velevance       +       -       +         Query results could be sorted alphabtically by fields (author, publication year)       +       + (?)       +         Datastore for specific languages       +       +       +         Datastore content as file, URL and database cell       + (file)       + (file)       ++ (file, url, cell)         Storing of relational data (for specific needs)       +       + (JSP)       ++ (JSP, SOAP with XPath/XQuery, REST, WebDAV, XMLRPC, Atom Publishing Protocol)       + (JSP, SOAP with XPath/XQuery, REST)         Build in WebServer with Servlet-Container for dynamic SQL and XQuery execution (also WebServices)       -       ++       +         Maximal size of XML-document       > 2 GB, efficiency is no problem       theoretically as big as the maximum size of a file on the file system; but system hangs if it is too big (> 100 MB)       theoretically as big as the maximum size of a file on the file system;? has to be tested?         Maximal number of XML-documents       > 1.000.000       231       ?	Query results and suctermization	(ouva)	,	1
relevance Query results could be sorted alphabtically by fields (author, publication year) Datastore for specific languages Datastore content as file, URL and database cell Storing of relational data (for specific needs) Web interfaces  + (JSP)  H+ (JSP, SOAP with XPath/XQuery, REST, WebDAV, XMLRPC, Atom Publishing Protocol)  Build in WebServer with Servlet-Container for dynamic SQL and XQuery execution (also WebServices)  Maximal size of XML-document  Maximal number of XML-documents  Aximal number of XML-documents  Maximal number of XML-documents  A + (File)	Query results could be sorted by	+	-	• • • • • • • • • • • • • • • • • • • •
Complete the control of the contro				
alphabtically by fields (author, publication year)  Datastore for specific languages  Datastore content as file, URL and database cell  Storing of relational data (for specific needs)  Web interfaces  + (JSP)  ++ (JSP, SOAP with XPath/XQuery, REST, WebDAV, XMLRPC, Atom Publishing Protocol)  Build in WebServer with Servlet-Container for dynamic SQL and XQuery execution (also WebServices)  Maximal size of XML-document  Page 14  + (file)  ++ (file)  ++ (file)  ++ (JSP, SOAP with XPath/XQuery, REST)  WebDAV, XMLRPC, Atom Publishing Protocol)  ++  ++  ++  ++  ++  ++  ++  theoretically as big as the maximum size of a file on the file system; but system hangs if it is too big (> 100 MB)  Maximal number of XML-documents > 1.000.000  Maximal number of XML-documents > 1.000.000  Maximal number of XML-documents > 1.000.000  A theoretically as big as the maximum size of a file on the file system; as to be tested ?		+	+ (2)	+
Datastore for specific languages Datastore content as file, URL and database cell Storing of relational data (for specific needs) Web interfaces  + (JSP)  H+ (JSP, SOAP with XPath/XQuery, REST)  Build in WebServer with Servlet-Container for dynamic SQL and XQuery execution (also WebServices)  Maximal size of XML-document  Passing as the maximum size of a file on the file system; but system hangs if it is too big (> 100 MB)  Maximal number of XML-documents  Patson + (Hile)  H+ (file)  H+ (SP, SOAP with XPath/XQuery, REST)  WebDAV, XMLRPC, Atom Publishing Protocol)  H+   H+   H+   H+   H+   H+   H+   H+			. (.)	
Datastore for specific languages Datastore content as file, URL and database cell  Storing of relational data (for specific needs)  Web interfaces  + (JSP)  + (JSP, SOAP with XPath/XQuery, REST, WebDAV, XMLRPC, Atom Publishing Protocol)  Build in WebServer with Servlet-Container for dynamic SQL and XQuery execution (also WebServices)  Maximal size of XML-document  Maximal size of XML-document  Maximal number of XML-documents  > 1.000.000  + + + + + + (file)  + (file)  + + (file)  + (file)  + + (file)  + (JSP, SOAP with XPath/XQuery, REST)  WebDAV, XMLRPC, Atom Publishing Protocol)  + + + + + + + + + (complex)				
Datastore content as file, URL and database cell  Storing of relational data (for specific needs)  Web interfaces  + (JSP)  ++ (JSP, SOAP with XPath/XQuery, REST, WebDAV, XMLRPC, Atom Publishing Protocol)  Build in WebServer with Servlet-Container for dynamic SQL and XQuery execution (also WebServices)  Maximal size of XML-document  > 2 GB, efficiency is no problem    Signature		+	+	+
database cell       Storing of relational data (for specific needs)       + (JSP)       + (JSP, SOAP with XPath/XQuery, REST, WebDAV, XMLRPC, Atom Publishing Protocol)       + (JSP, SOAP with XPath/XQuery, REST)         Build in WebServer with Servlet-Container for dynamic SQL and XQuery execution (also WebServices)       -       + + + + + + + + + + + + + + + + + + +				
Storing of relational data (for specific needs)  Web interfaces  + (JSP)  ++ (JSP, SOAP with XPath/XQuery, REST, WebDAV, XMLRPC, Atom Publishing Protocol)  Build in WebServer with Servlet-Container for dynamic SQL and XQuery execution (also WebServices)  Maximal size of XML-document  > 2 GB, efficiency is no problem  > 2 GB, efficiency is no problem    Solution of a file on the file system; but system hangs if it is too big (> 100 MB)    Maximal number of XML-documents   > 1.000.000   2311   ?		(1110)	(1110)	(mo, an, con)
Specific needs   Web interfaces   + (JSP)   + (JSP, SOAP with XPath/XQuery, REST, WebDAV, XMLRPC, Atom Publishing Protocol)   ++   +   +   +   +   +   +   +   +		_	+	++ (complex)
Web interfaces       + (JSP)       ++ (JSP, SOAP with XPath/XQuery, REST, WebDAV, XMLRPC, Atom Publishing Protocol)       + (JSP, SOAP with XPath/XQuery, REST)         Build in WebServer with Servlet-Container for dynamic SQL and XQuery execution (also WebServices)       -       ++       ++         Maximal size of XML-document       > 2 GB, efficiency is no problem       theoretically as big as the maximum size of a file on the file system; but system hangs if it is too big (> 100 MB)       theoretically as big as the maximum size of a file on the file system; ? has to be tested ?         Maximal number of XML-documents       > 1.000.000       231       ?				(complex)
with XPath/XQuery, REST, WebDAV, XMLRPC, Atom Publishing Protocol)  Build in WebServer with Servlet-Container for dynamic SQL and XQuery execution (also WebServices)  Maximal size of XML-document  > 2 GB, efficiency is no problem  > 2 GB, efficiency is no problem    Theoretically as big as the maximum size of a file on the file system; but system hangs if it is too big (> 100 MB)    Maximal number of XML-documents   > 1.000.000   231   ?		+ (.ISP)	++ (JSP SOAP	+ (JSP SOAP
XPath/XQuery, REST, WebDAV, XMLRPC, Atom Publishing Protocol)	VVOD IIItoriaces	(001)	,	1
Build in WebServer with Servlet- Container for dynamic SQL and XQuery execution (also WebServices)  Maximal size of XML-document  Publishing Protocol)  ++  ++  ++  theoretically as big as the maximum size of a file on the file system; but system hangs if it is too big (> 100 MB)  Maximal number of XML-documents  AREST)  REST)  REST)  Hear of Altom Publishing Protocol)  ++  ++  ++  Hear of Altom Publishing Protocol)   100 MB  Postocol  100 MB  231  PREST)  REST)  PHA  Theoretically as big as the maximum size of a file on the file system; 9 has to be tested 9 has to be 100 has 10			=	
Build in WebServer with Servlet- Container for dynamic SQL and XQuery execution (also WebServices)  Maximal size of XML-document  > 2 GB, efficiency is no problem  > 2 GB, efficiency is no problem    Solution				
Build in WebServer with Servlet- Container for dynamic SQL and XQuery execution (also WebServices)  Maximal size of XML-document  > 2 GB, efficiency is no problem    System; but system; but system; hangs if it is too big (> 100 MB)    Maximal number of XML-documents   > 1.000.000   2 <sup>31</sup>			1	112017
Build in WebServer with Servlet- Container for dynamic SQL and XQuery execution (also WebServices)  Maximal size of XML-document  > 2 GB, efficiency is no problem  > 2 GB, efficiency is no problem    Theoretically as big as the maximum size of a file on the file system; but system hangs if it is too big (> 100 MB)    Maximal number of XML-documents   > 1.000.000   2 <sup>31</sup>   ?				
Build in WebServer with Servlet- Container for dynamic SQL and XQuery execution (also WebServices)  Maximal size of XML-document  > 2 GB, efficiency is no problem  > 2 GB, efficiency is no problem    Solution of a file on the file system; but system hangs if it is too big (> 100 MB)    Maximal number of XML-documents   > 1.000.000   2 <sup>31</sup>   ?				
Build in WebServer with Servlet- Container for dynamic SQL and XQuery execution (also WebServices)  Maximal size of XML-document  > 2 GB, efficiency is no problem    Solution				
Container for dynamic SQL and XQuery execution (also WebServices)  Maximal size of XML-document  > 2 GB, efficiency is no problem  > a theoretically as big as the maximum size of a file on the file system; but system hangs if it is too big (> 100 MB)  Maximal number of XML-documents  > 1.000.000  2 31  **Theoretically as big as the maximum size of a file on the file system; ? has to be tested?	Build in WebServer with Servlet-	_	,	++
XQuery execution (also WebServices)  Maximal size of XML-document  > 2 GB, efficiency is no problem    Solution   State   Stat				
WebServices)       > 2 GB, efficiency is no problem       theoretically as big as the maximum size of a file on the file system; but system hangs if it is too big (> 100 MB)       theoretically as big as the maximum size of a file on the file system; has to be tested?         Maximal number of XML-documents       > 1.000.000       231       ?				
Maximal size of XML-document  > 2 GB, efficiency is no problem  > 3 GB, efficiency is no problem    Solution of a file on the file system; but system hangs if it is too big (> 100 MB)    Maximal number of XML-documents   > 1.000.000   231				
efficiency is no problem  big as the maximum size of a file on the file system; but system hangs if it is too big (> 100 MB)  Maximal number of XML-documents > 1.000.000  big as the maximum size of a file on the file system; ? has to be tested ?  100 MB)	/	> 2 GB.	theoretically as	theoretically as
problem maximum size of a file on the file system; but system hangs if it is too big (> 100 MB)  Maximal number of XML-documents > 1.000.000 maximum size of a file on the file system; ? has to be tested?		,		,
of a file on the file system; but system hangs if it is too big (> 100 MB)  Maximal number of XML-documents > 1.000.000 2 <sup>31</sup> ?		•		•
file system; but system hangs if it is too big (> 100 MB)  Maximal number of XML-documents > 1.000.000 2 <sup>31</sup> ?				
system hangs if it is too big (> 100 MB)  Maximal number of XML-documents > 1.000.000 2 <sup>31</sup> ?				
it is too big (> ? 100 MB)  Maximal number of XML-documents > 1.000.000 2 <sup>31</sup> ?			1	1
Maximal number of XML-documents > 1.000.000 2 <sup>31</sup> ?				?
Maximal number of XML-documents > 1.000.000 2 <sup>31</sup> ?			100 MB)	
	Maximal number of XML-documents	> 1.000.000		?
į į				

...

#### Summary:

Lucene could be the system for fulltext querying (with stemming, etc.). eXist could be the system for structural queries (with XQuery).

Oracle could be the system for both (fulltext querying and structurural querying) and also has the web engine.

#### Web based querying and presenting XML-documents

#### Functionality:

- query GUI for searching XML-documents:
- rendering of the document with XSLT
- enrichment of the document with inline images and links to external resources(to e.g. Pollux, geospatial data, etc.)

#### Lucene:

Web engine: Tomcat as JSP-Container, WAR-application files, static XML-files

#### eXist:

Webengine with Servlet-Container already contained in database system (has to be activated), access with XQuery or other techniques

#### Oracle:

Webengine with Servlet-Container already contained in database system (has to be activated), access with XQuery, dynamic SQL or other techniques

#### 3. Hardware

- 1. Developer-Client-Machines: Mac OS X computer or notebook with relevant developer software
- 2. Developer-Server: Linux server (small) for document bases and server software (Indexing and querying software, web engine, backup-software).
- 3. Productional-Server (at the end of the project): Linux server (as big as the user needs) for document bases and server software (Indexing and querying software, web engine).

## 4. Project organisation

All project descriptions (requirements, specifications, project plans, etc.) are available in Trac-Wiki:

https://itgroup.mpiwg-berlin.mpg.de:8080/tracs/mpdl-project-software

Page 5

MPIWG:Users:jwillenborg:Documents:MPDL-Project:requirements:technicalSpec-02.doc Erstellt von J. Willenborg Last modified 18.09.2008 9:34

All produced project software with documentation are maintained by Subversion and could be browsed in Trac-Wiki:						
https://itgroup.mpiwg-berlin.mpg.de:8080/tracs/mpdl-project-software/browser						