# Tools for the Digital Diplomatist

## Open source tools for online publication of charters

**Francesca CAPOCHIANI (Università degli studi di Pisa)**
**Chiara LEONI (Università degli studi di Pisa)**
**Roberto ROSSELLI DEL TURCO (Università degli studi di Torino / Università degli studi di Pisa)**

# Introduction

- searchable charter corpus = effective research tool for historians
- creating a digital corpus: highly desirable, complicated, intimidating
- method proposed: **text encoding** + **visualization software** + **database software**
- workflow accomplished through use of **open source** tools
- (web server/space to publish on the WWW)

# Text Encoding

- to mark-up or not to mark-up
- sticking to simple text: faster path but also a limited one
  - generic search engines → simple searches
  - texts encoded in HTML anyway
- text encoding: slower, more flexible, powerful path
  - XML search engines: complex, powerful searches
  - convert to (X)HTML for web view, PDF for print
  - can be used to perform text analysis
- light encoding schema → enrich it progressively
- wealth of open source XML editors

# XML Schemas

- XML schema = ‘grammar’ of encoded texts
- custom made schemas:
  - as simple as one wants
  - no or difficult interchange with other projects
  - everything (f.i. stylesheets) has to be created *ex novo*
- TEI (http://www.tei-c.org/) schemas:
  - the current "standard"
  - comprehensive set of schema components (modules) → flexible customization
  - well documented (TEI Guidelines)
  - community support (mailing list, wiki, TEI by example)

# XML-based Projects

- non TEI:
  - CDLM http://cdlm.unipv.it/
  - DEEDS http://res.deeds.utoronto.ca:49838/research/
- TEI based:
  - Charters Encoding Initiative http://www.cei.lmu.de/
- TEI projects:
  - Éditions en ligne de l'École des chartes (ELEC) http://elec.enc.sorbonne.fr/
  - The Electronic Sawyer http://www.esawyer.org.uk/
  - Chartae Burgundiae Medii Aevi http://www.artehis-cbma.eu/
  - Monumenta Germaniae Historica http://www.dmgh.de/

# TEI Modules for Charter Encoding

- corpus Metadata for Language Corpora
- msdescription Manuscript Description
- namesdates Names, Dates, People, Places
- textcrit Text Criticism
- transcr Transcription of Primary Sources
- (analysis Analysis and Interpretation)
- (figures Tables, Formulae, Figures)
- (gaiji Character and Glyph Documentation)

- [MENOTA Schemas http://www.menota.org/]

# Future Developments in TEI Land

- absence of a “diplomatic-specific” tag set
- CEI
  - extensions to be integrated with TEI schemas
  - migration through use of XSLT stylesheets
  - see also the Anglo-Saxon Cluster project
- TEI
  - Document / Genetic Criticism module http://www.tei-c.org/SIG/Manuscripts/genetic.html
  - why not a SIG?
- just listen to C. Desenclos and V. Jolivet’s paper :)

# What About Images

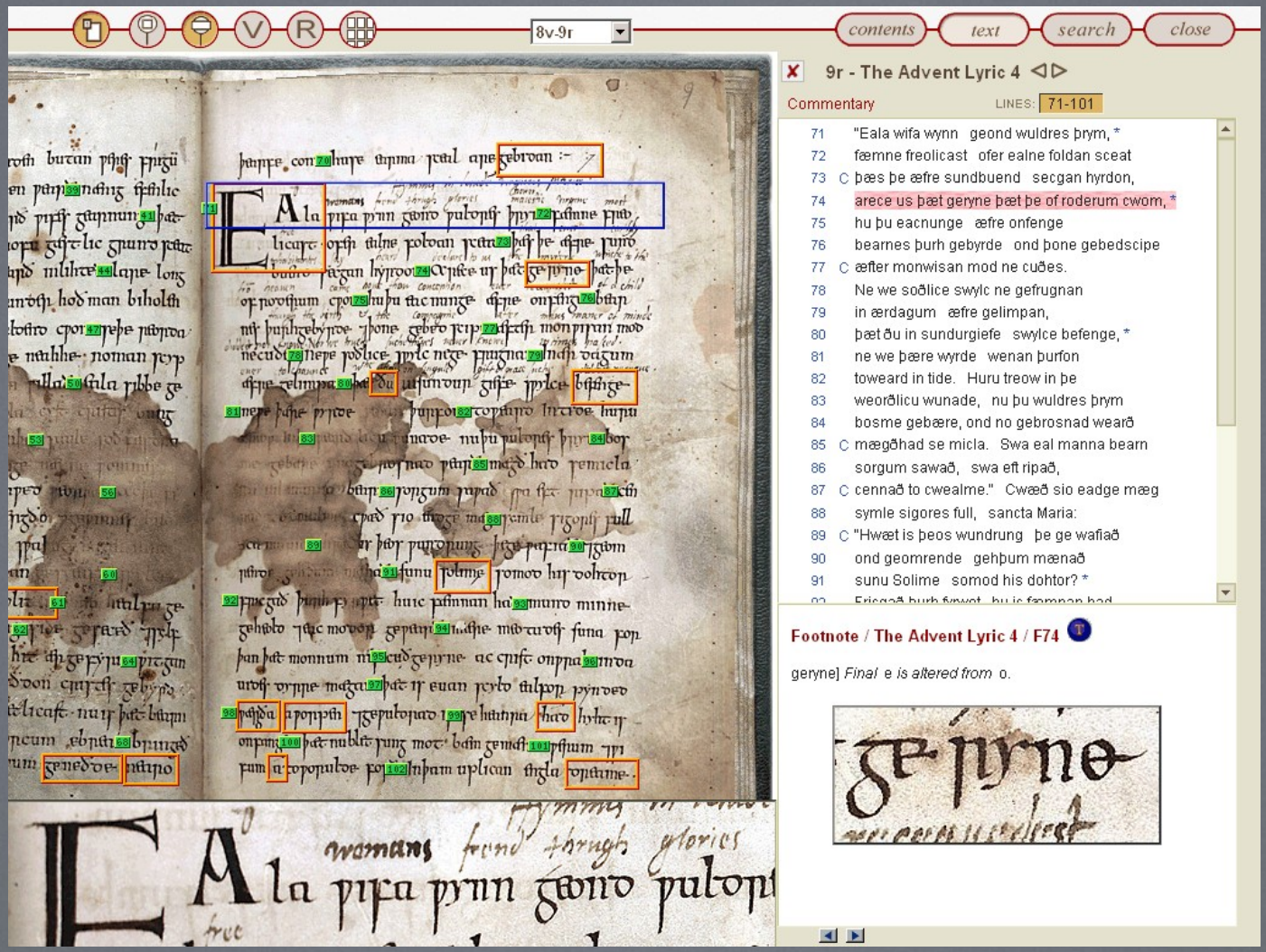
- not a primary goal, but sometimes useful / desirable to include manuscript images
- including the ***transcr*** module in TEI P5 schemas allows to create image-based digital editions or digital facsimiles:
  - global pointer attribute **@facs**
  - element **<facsimile>**
  - element **<surface>**
  - element **<zone>**

# <facsimile> Encoding

- linking text and image
  - <facsimile> is a structural element,  simplest case: collection of <graphic> elements
  - @facs can be used to link text elements to images
  - for more complex interaction between text and images use <surface> and <zone>
  - <surface> defines the written surface as a rectangular zone
  - <zone> defines a rectangular area on the same cartesian space of a surface element
  - @facs and @corresp link text to image and viceversa

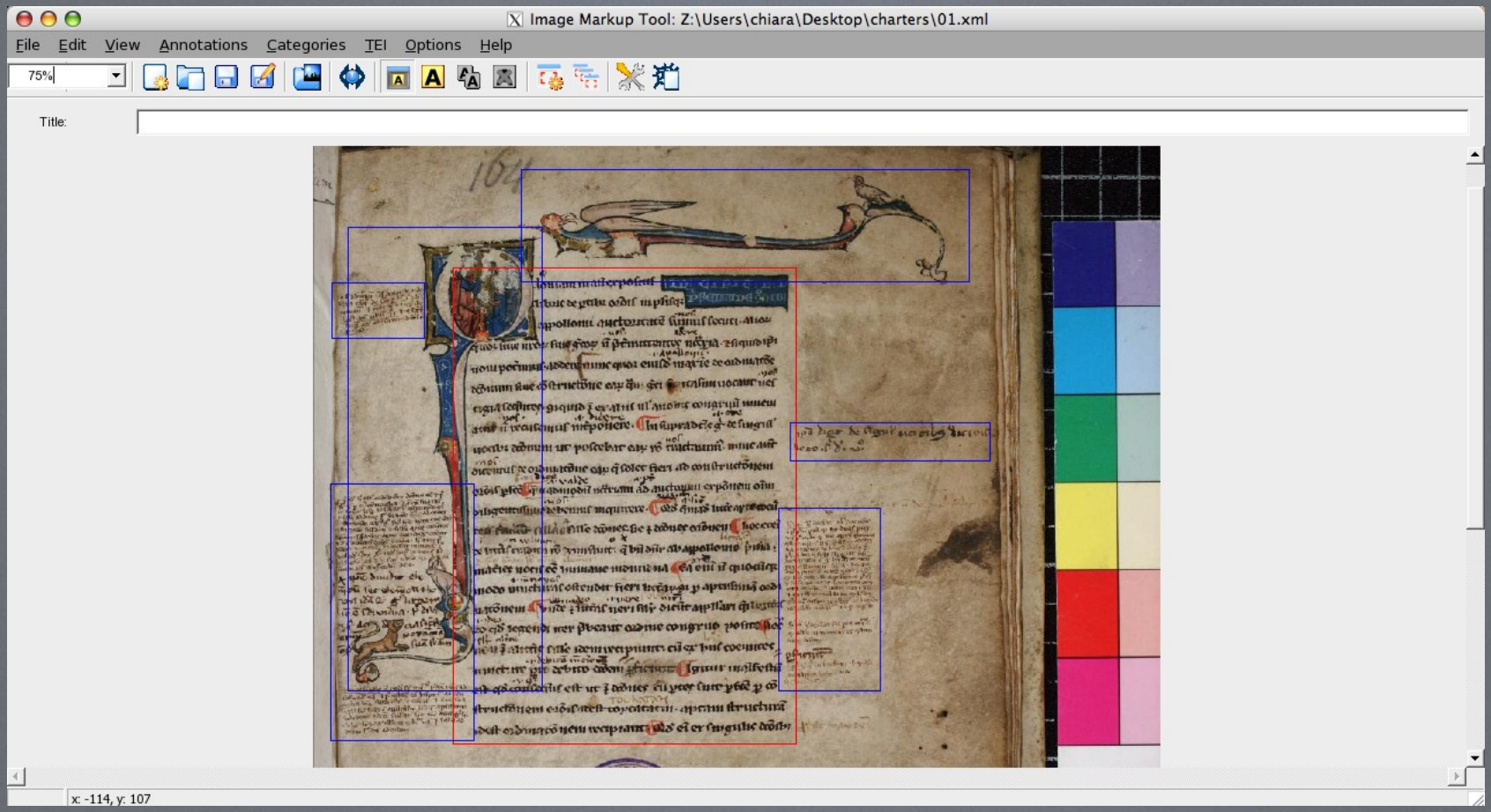
# <surface> and <zone>s

# Text-Image Linking: *Electronic Junius* (by B. Muir)



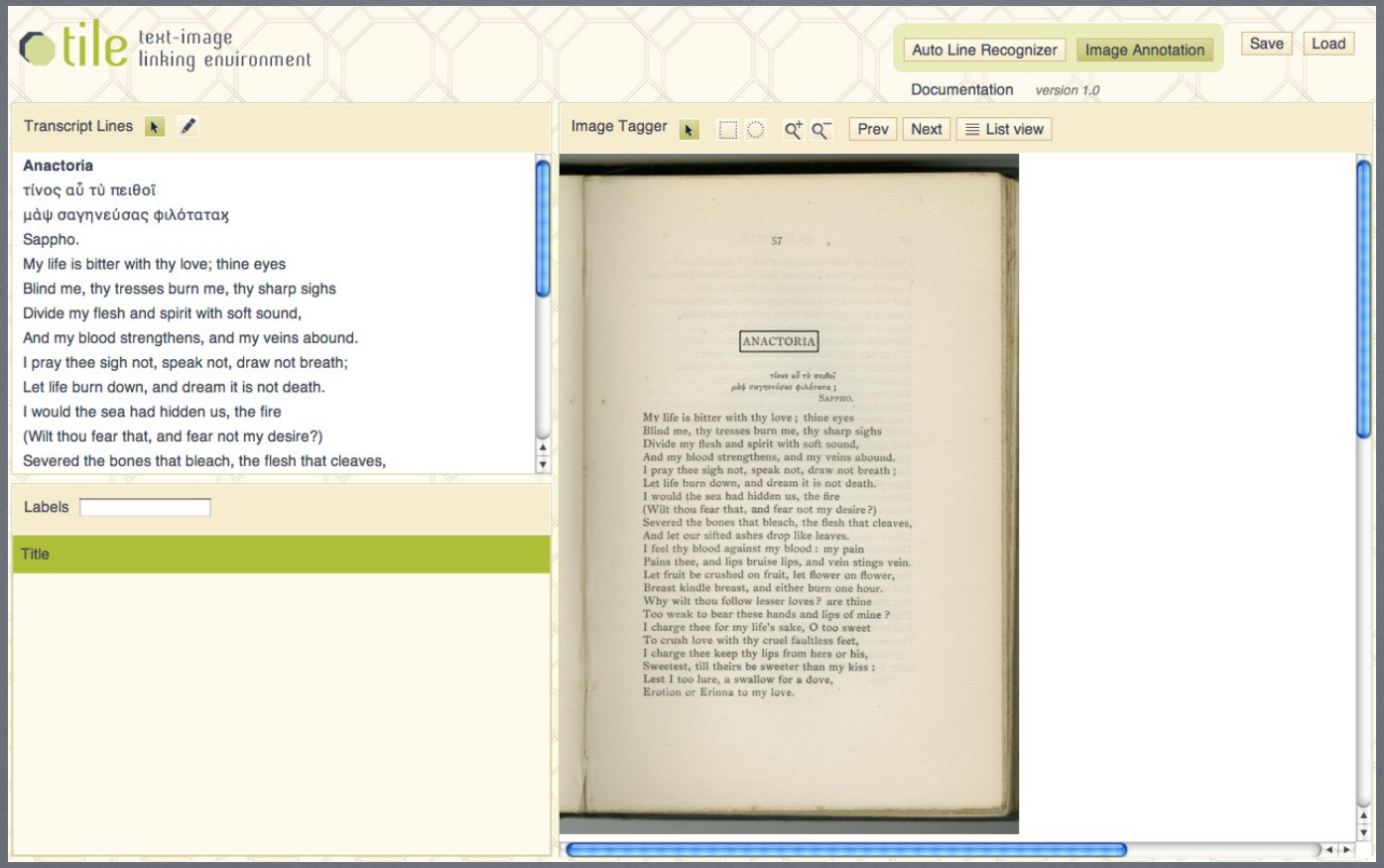
# The UVic Image Markup Tool

http://www.tapor.uvic.ca/~mholmes/image_markup/



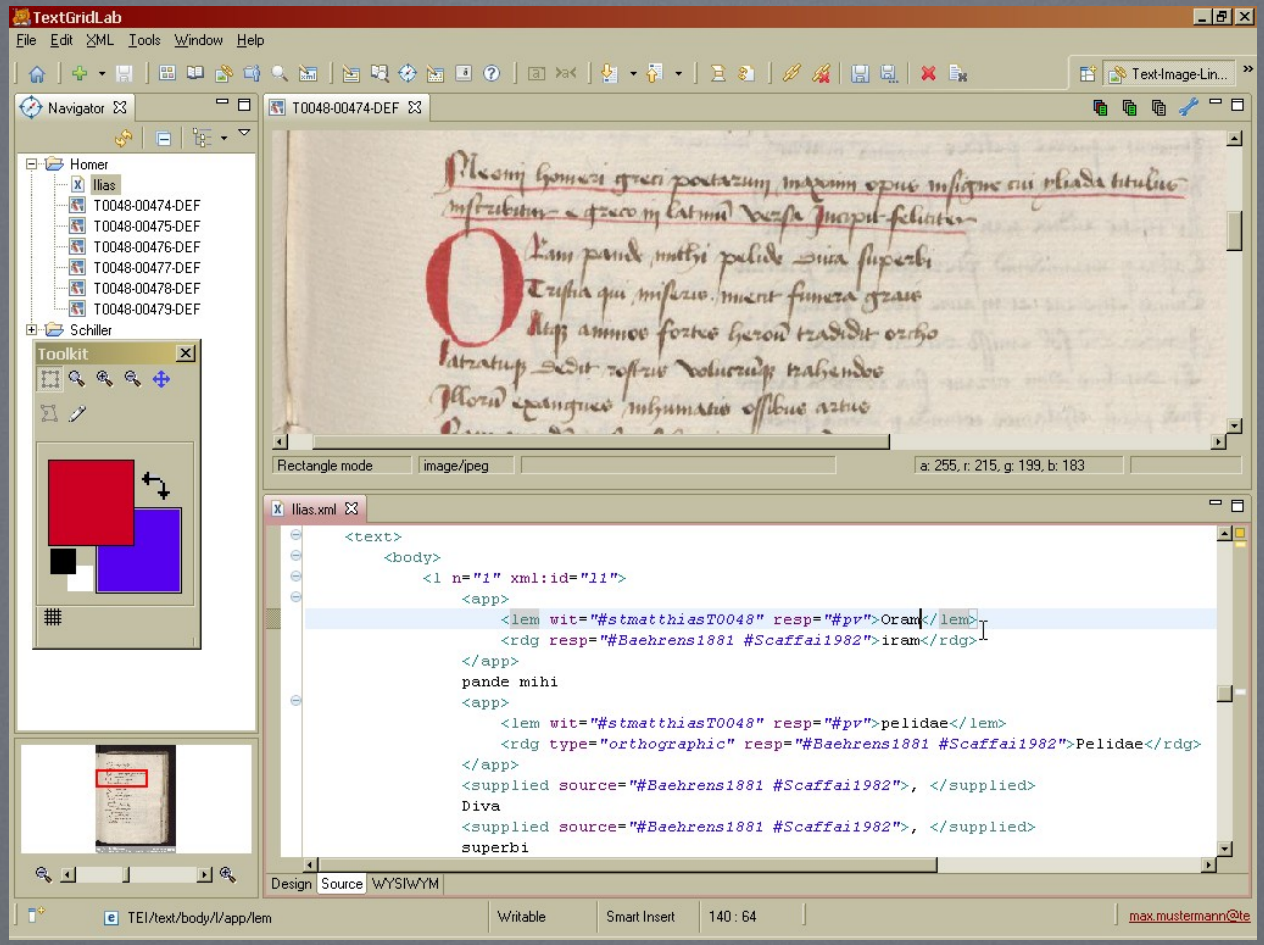
# TILE – Text-Image Linking Environment

http://mith.umd.edu/tile/



# TextGridLab

http://www.textgrid.de/en/1-0.html



# Putting It All Together

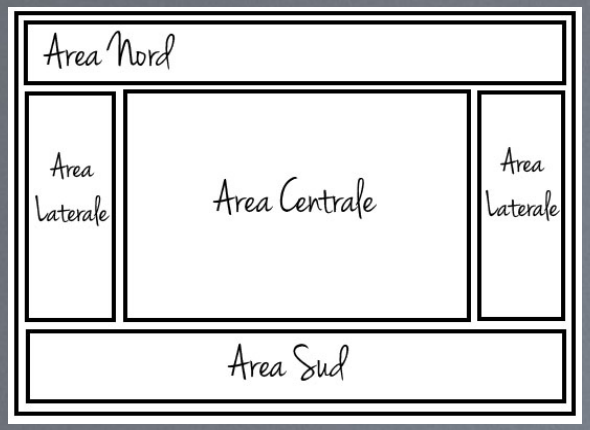
- how to browse text (and possibly images), perform searches of encoded texts?
- solution no. 1: create a web site based on an XML search engine
- solution no. 2: create a web site integrating XML search and other functionalities
- working towards 2 to provide a generic, flexible edition browsing software
- project name: **EVT** – Edition Visualization Technology

# What is EVT?

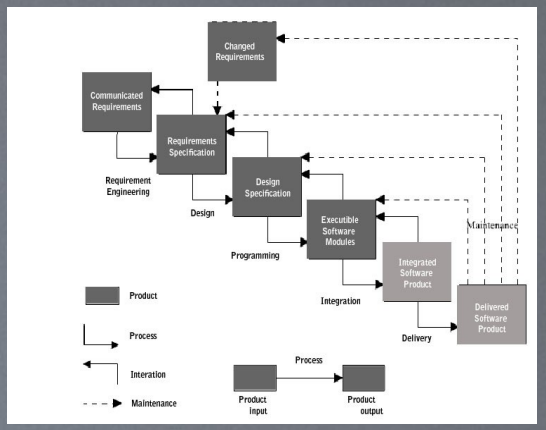
*Edition Visualization Technology*

"...a complete set of flexible and highly customizable editorial tools developed to allow users to view, read, search and compare editions in an electronic environment."

# Building the EVT: Project Planning

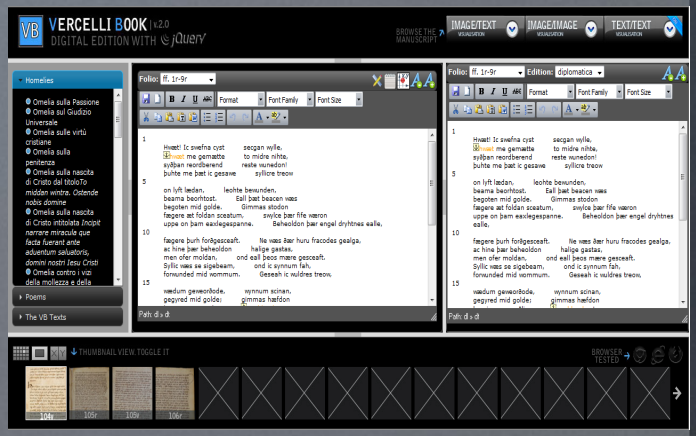



Designing the Gui




Software architecture

Creation of a prototype

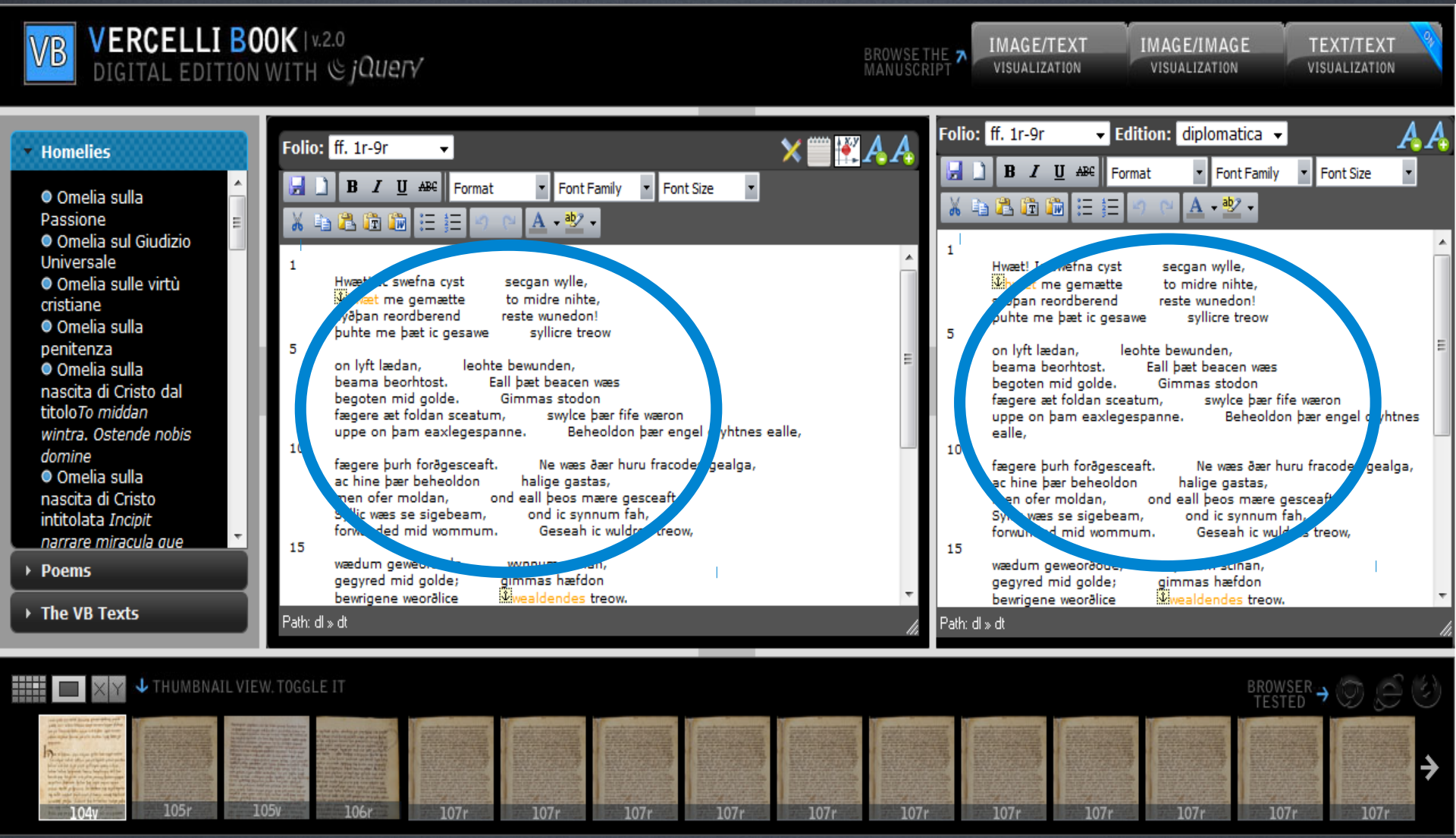


# Building the EVT

| General Criteria |
| --- |
| Human factors (ergonomics) |
| Consistency (link) |
| Control and Navigation |
| Scalability |
| Aestethic Integrity |
| Readability |

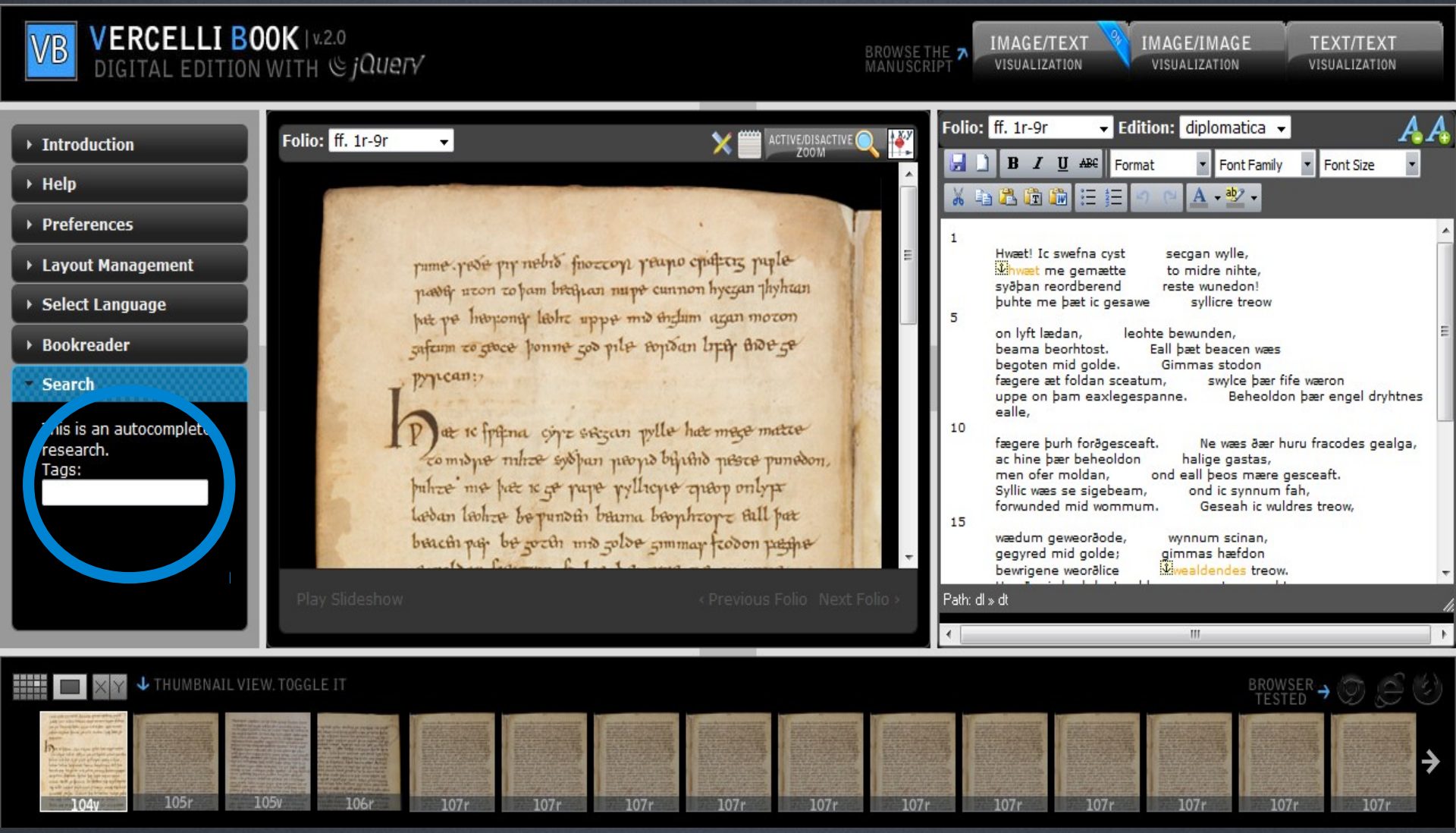
| Specific Requirements |
| --- |
| Good hyper-textual functionality |
| Image manipulation tools |
| Advanced search functionality |

# Benefits of XML with regards to visualization

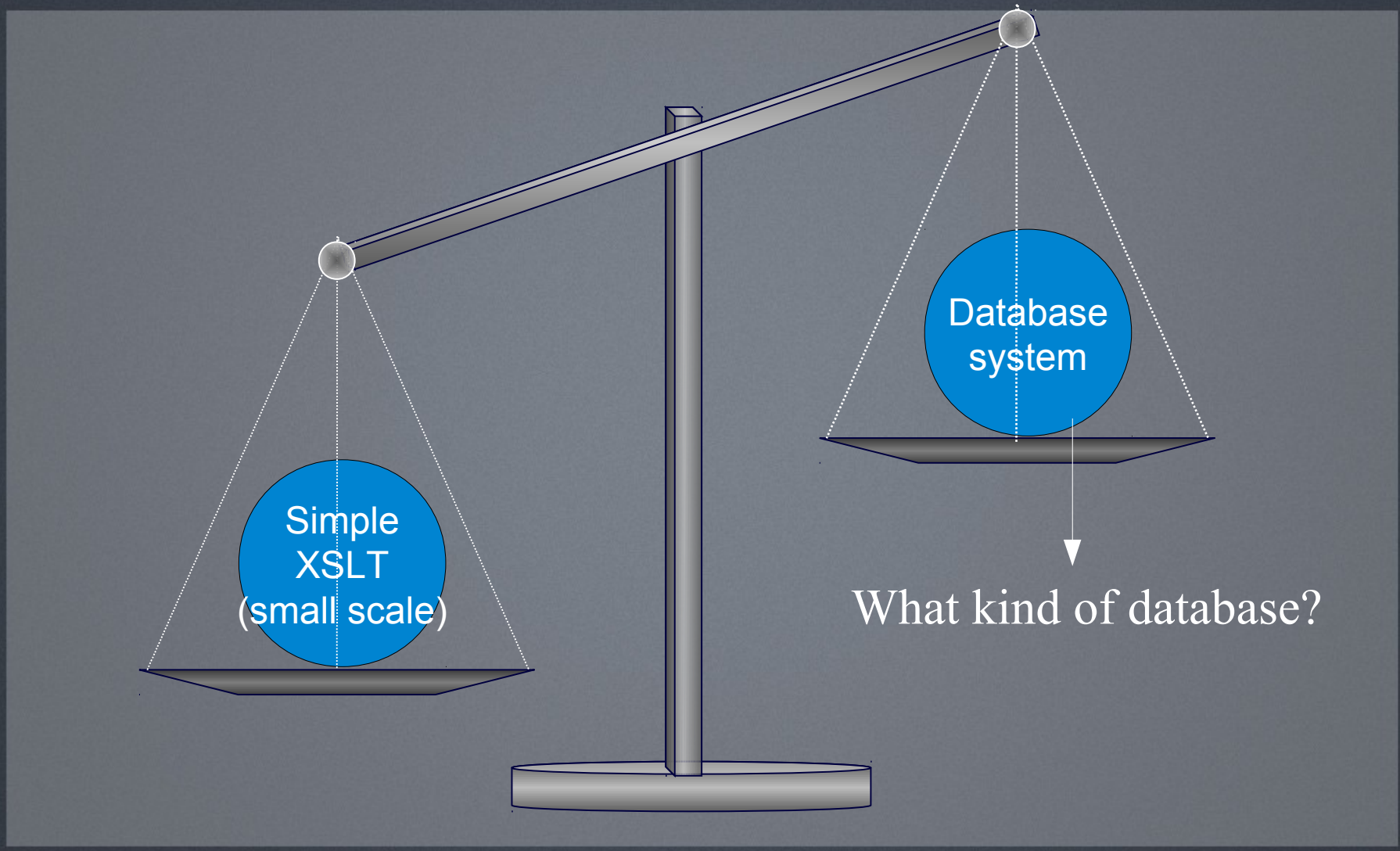


EVT Example Visualisation Software for a Web-based Digital Edition.

# Vercelli Book Digital Edition (using jQuery)

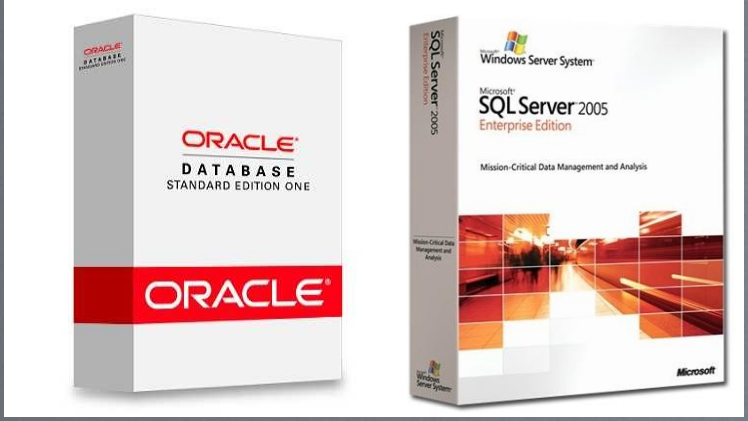


# How to manipulate XML files in EVT?



# XML DATA MANAGEMENT SYSTEM

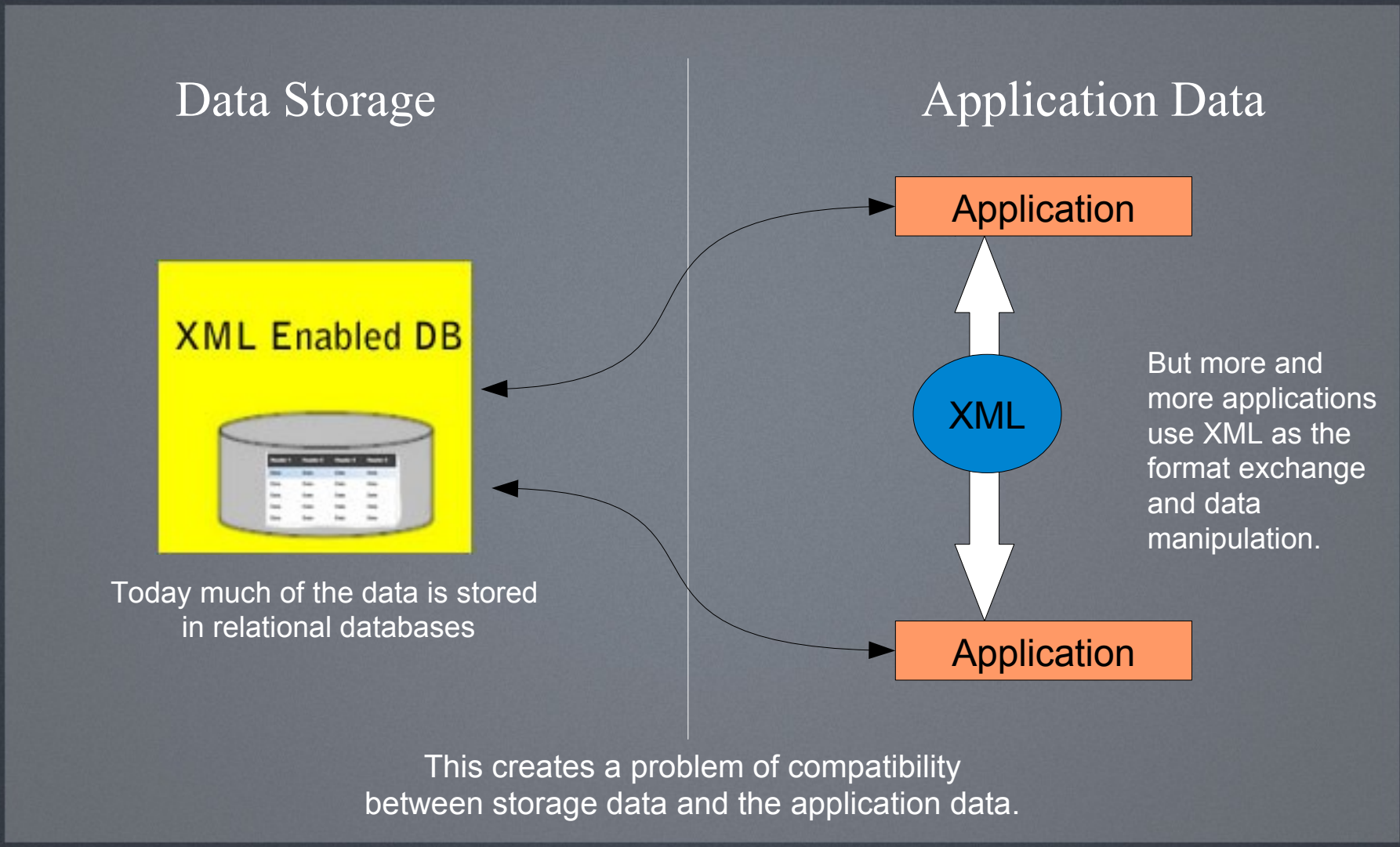
XML-Enabled Database




Native XML Database

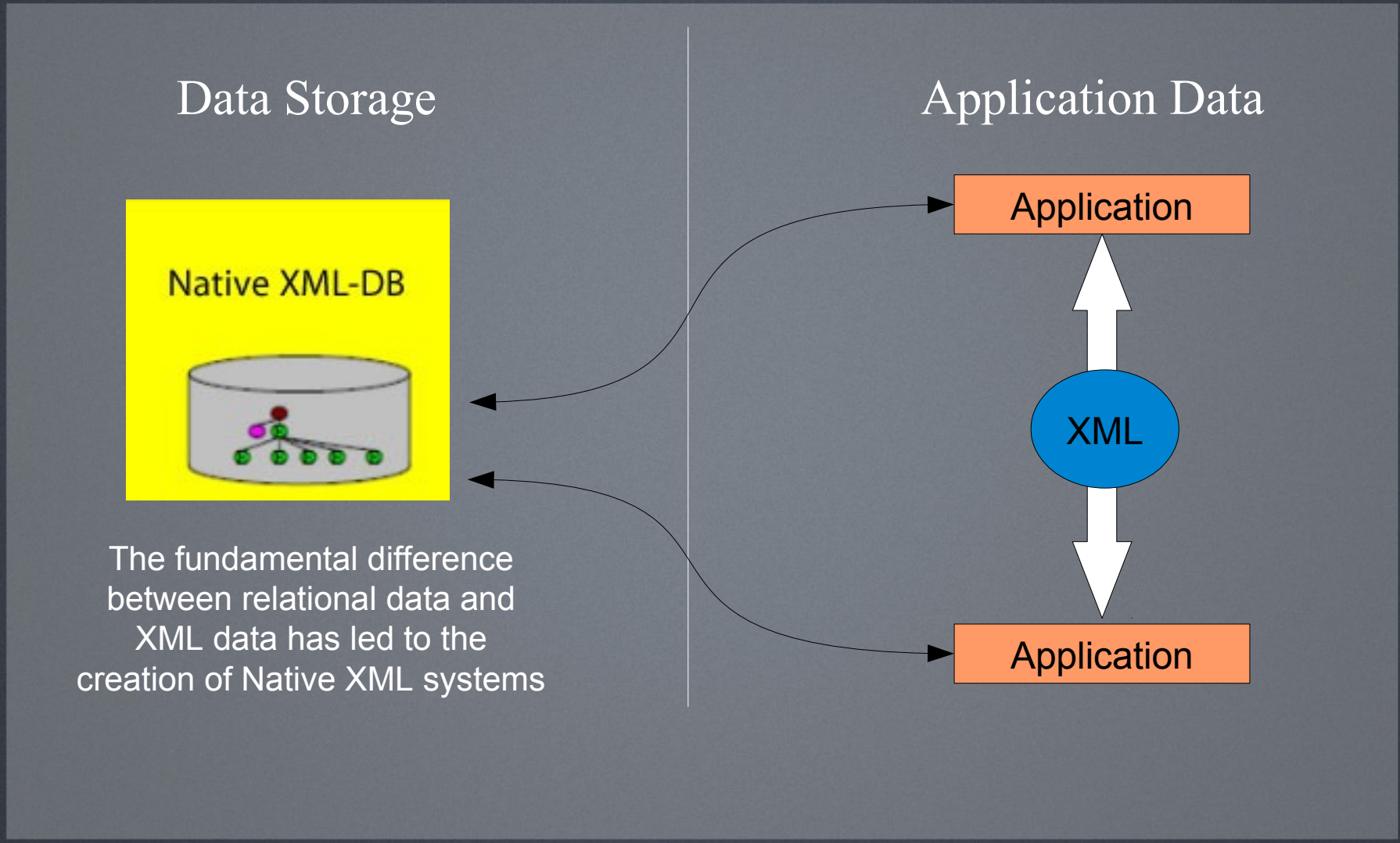
1. What are the differences?
2. What kind of Database for EVT?
3. Is there any open source and TEI-compatible software?

# 1. DATA FORMAT
## (XML-Enabled Database)



# 1. DATA FORMAT

## (XML-Native Database)




The fundamental difference between relational data and XML data has led to the creation of Native XML systems

# XML – Enabled vs Native XML

| | XML- Enabled | Native XML |
|---|---|---|
| Architecture | Record on tables | XML hierarchical struct. (collections, resources, service) |
| DTD or Schema | Necessary | Not necessary |
| Query | SQL | XQuery |
| XML data | Data-centric | Document-centric |
| | | |
| examples | Access2007, DB2 | Exist, Xindice |

# 2. What kind of DB (for EVT)?

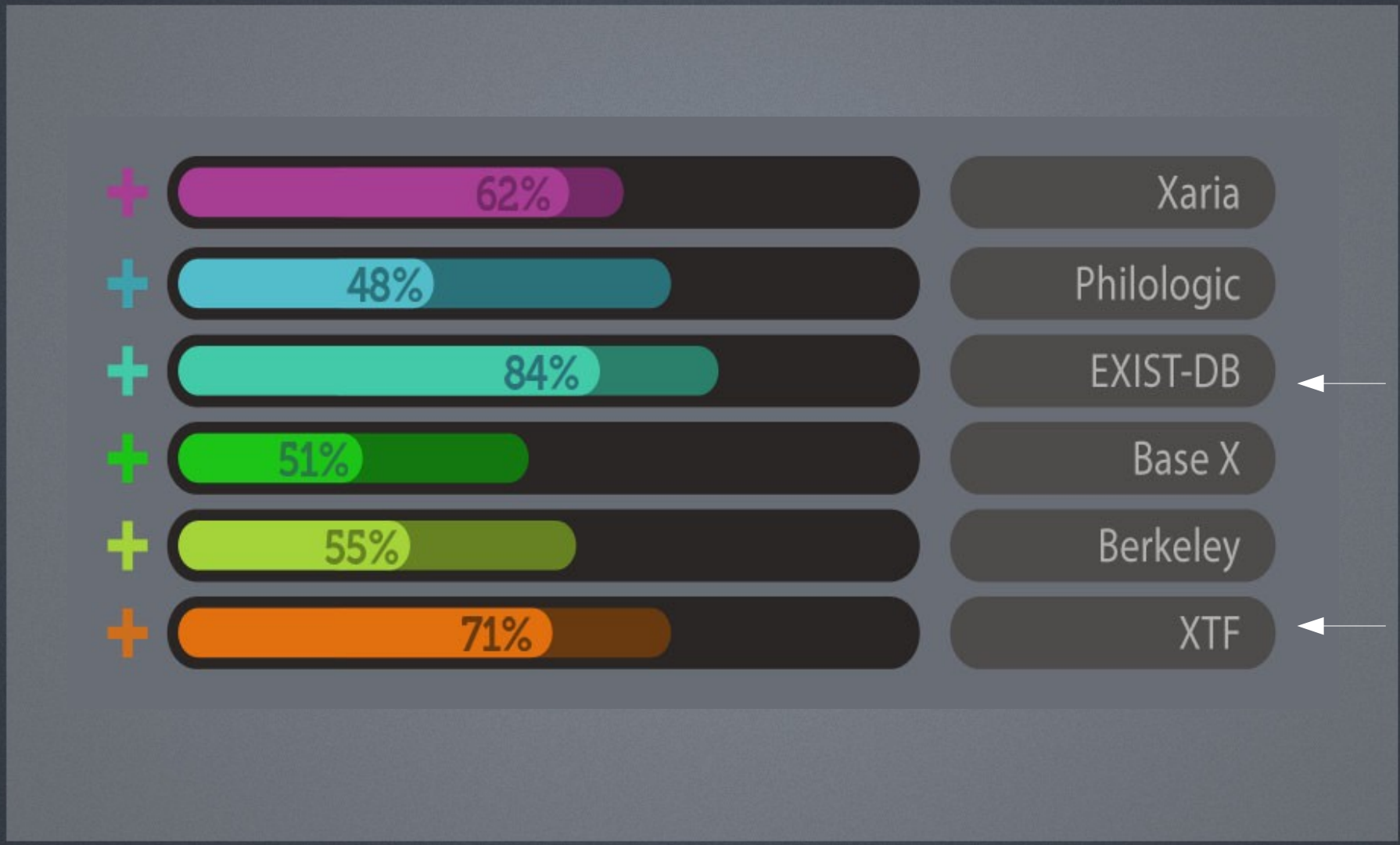
The efficiency of an XML-DB software product can be assessed according to:

- the ability to adapt to documents based on very different schemas.
- the possibility of regenerating XML documents by extracting data from the DB (query).
- the use of query languages (XQuery, XPath).
- the ability to integrate existing data in the XML-DB.

And....

an OPEN SOURCE license

# 3. Categories of Products: XML-DB Opensource



# eXist: Technologies Used

http://exist.sourceforge.net/

*INTERFACE*

–WebDAV

–Javascript+AJAX

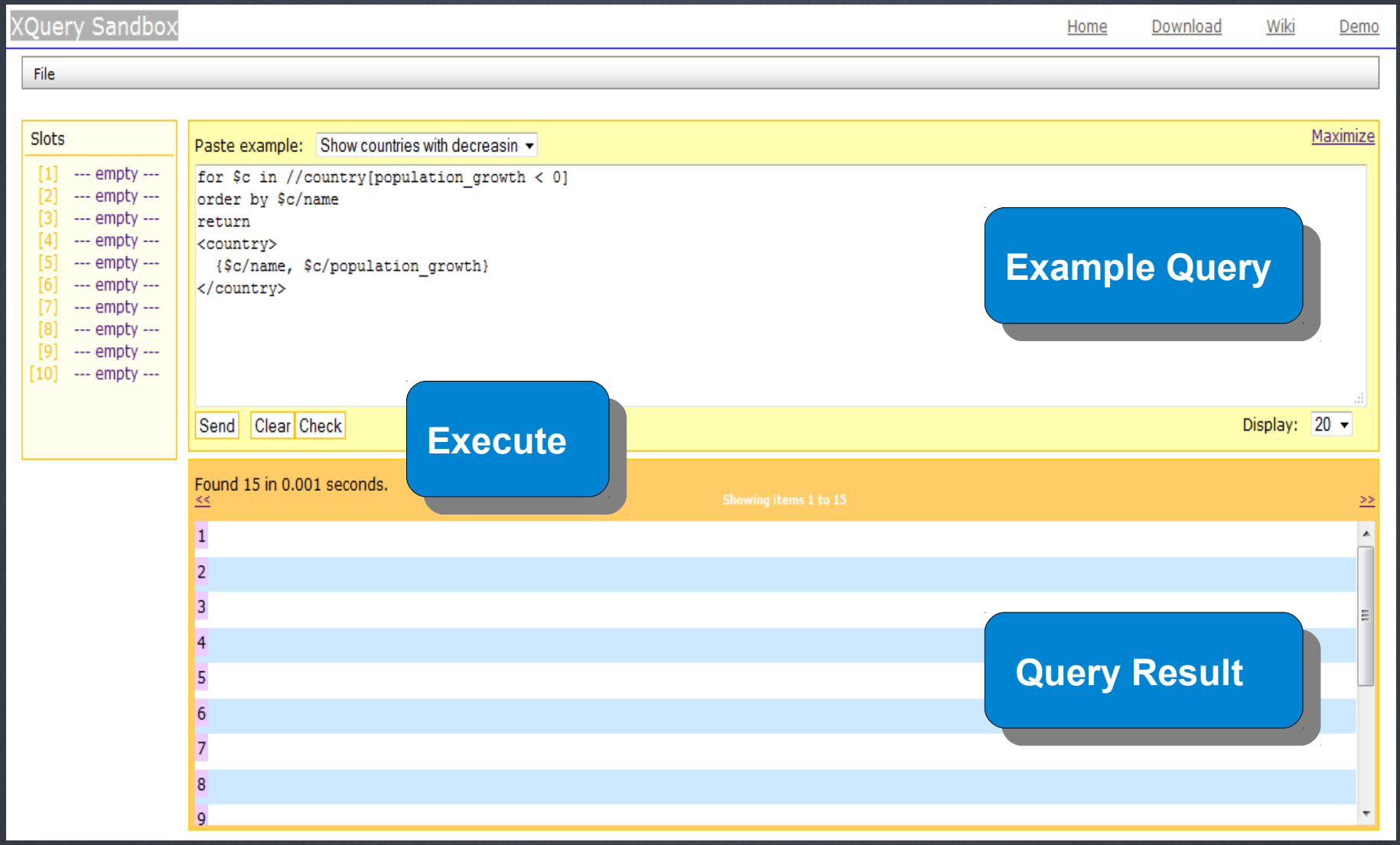
–HTML

*TECHNOLOGIES*

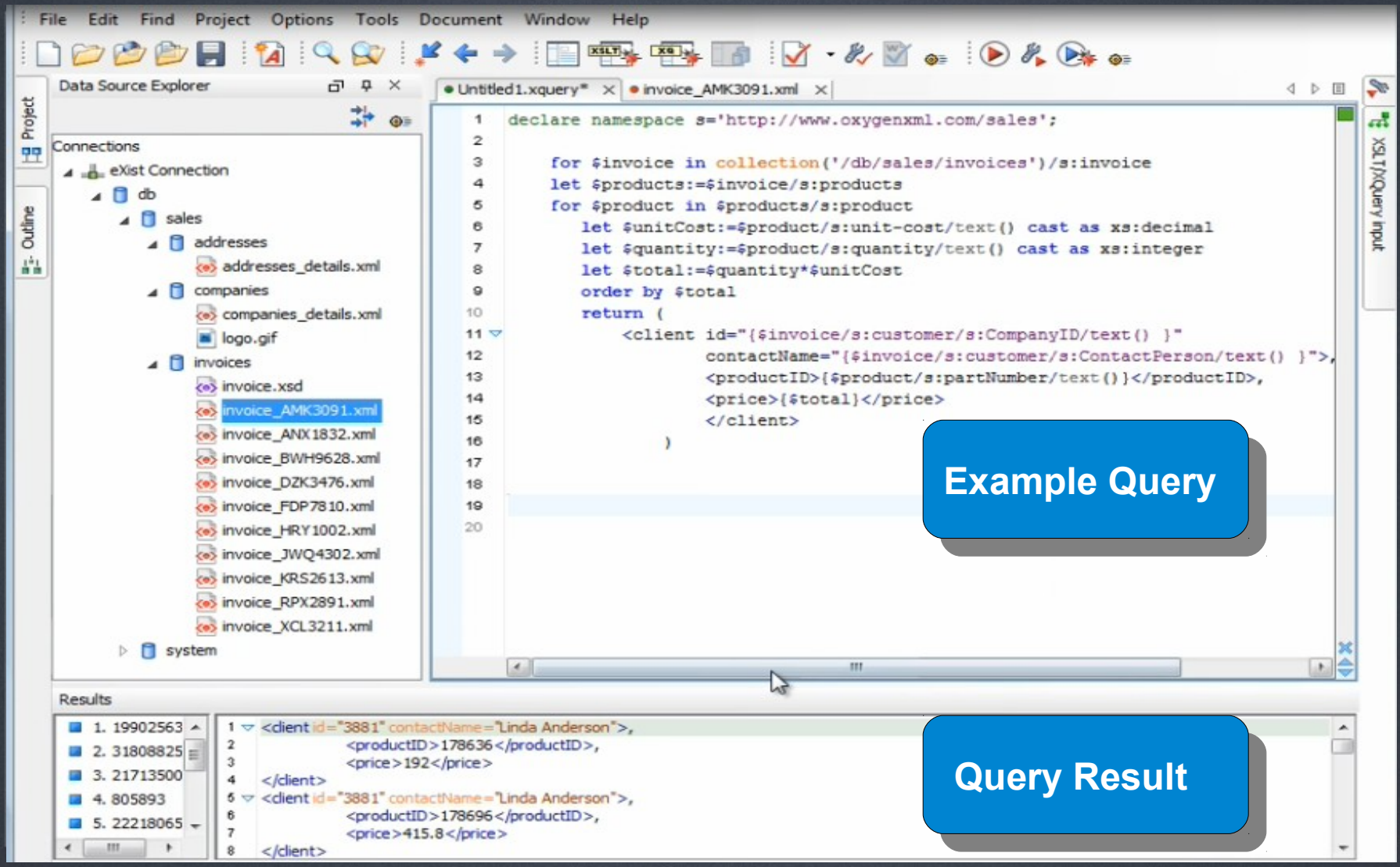
- XQuery
- XPath
- XSLT

# XQuery Sandbox

http://demo.exist-db.org/sandbox/sandbox.xql

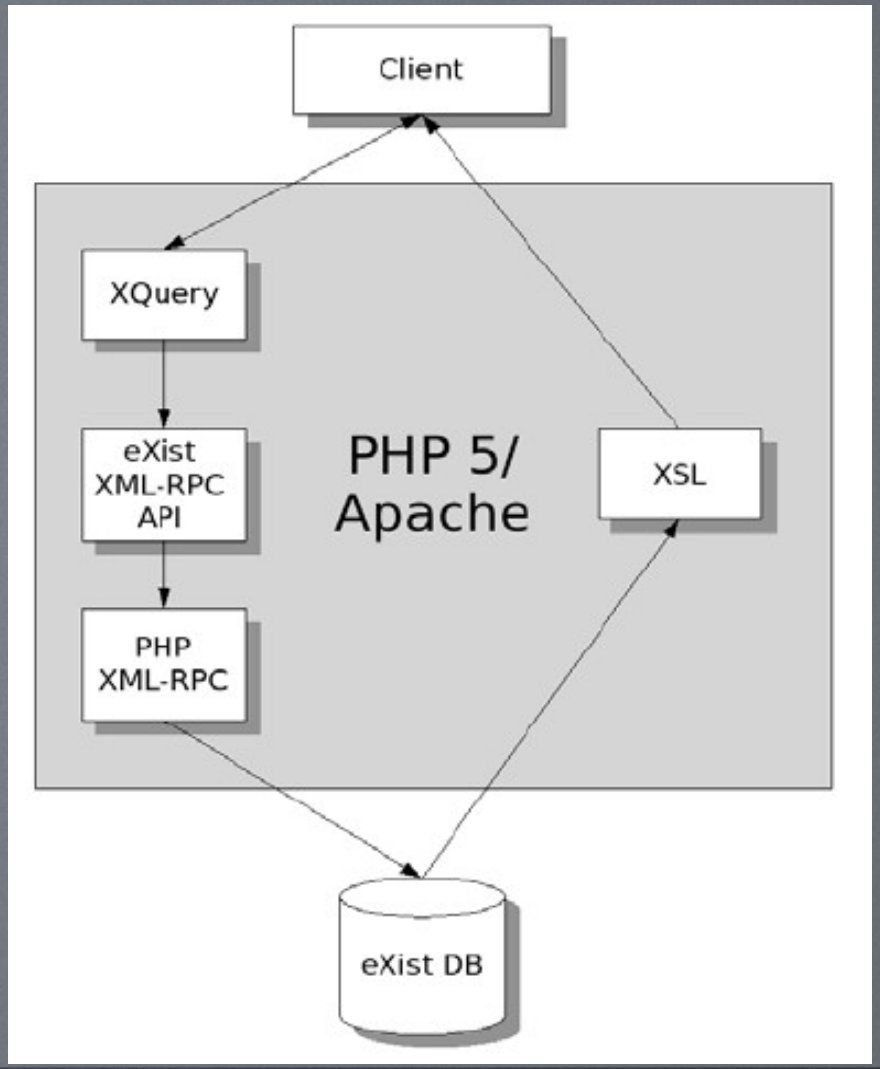


# Configure eXist on Oxygen



# eXist-db Architecture

The website runs on PHP 5/Apache and connects to the eXist database through XML Remote Procedure Call.



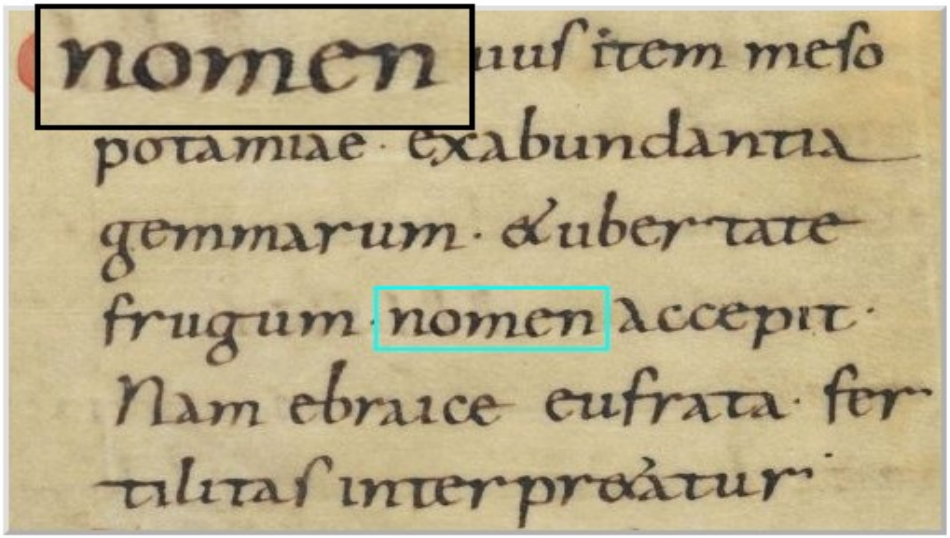
# Example of eXist-Based Project

http://mariage.uvic.ca/

Basic Usage: the query may concern text, author, date...

# Example of eXist-Based Project

Interactive Album of Medieval Palaeography




Advanced Usage:

Select a line
Write on the box
Return from the box

**Transcription**

See the user manual file if you have trouble using this exercise. A double-click on a word in the image displays its transcription.

Ligne 1

Ligne 2

Ligne 3

Ligne 4

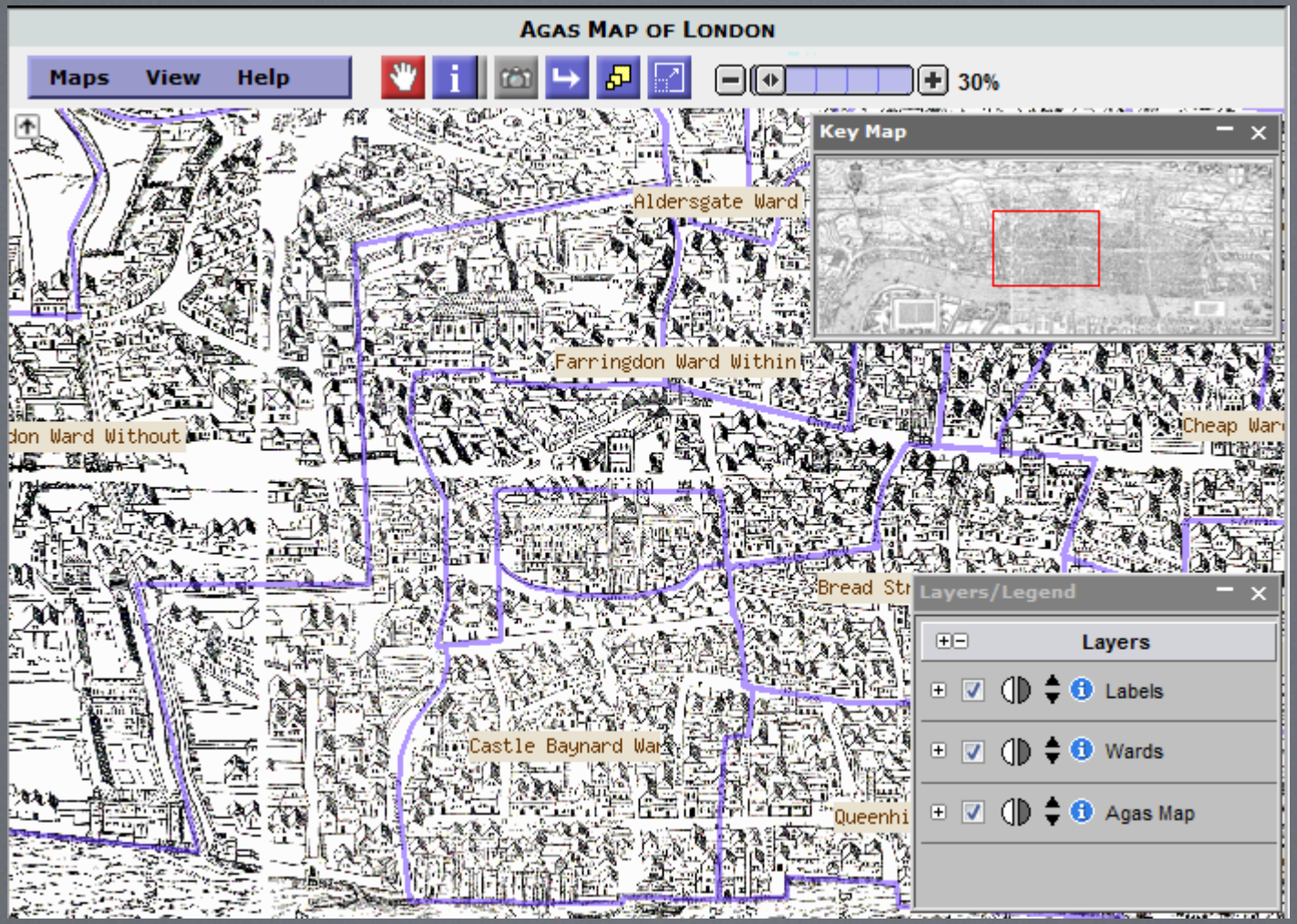
nomen

Ligne 5

http://ciham.ish-lyon.cnrs.fr/paleographie/index.php?l=en

# Example of eXist-Based Project

http://www.mapoflondon.uvic.ca/

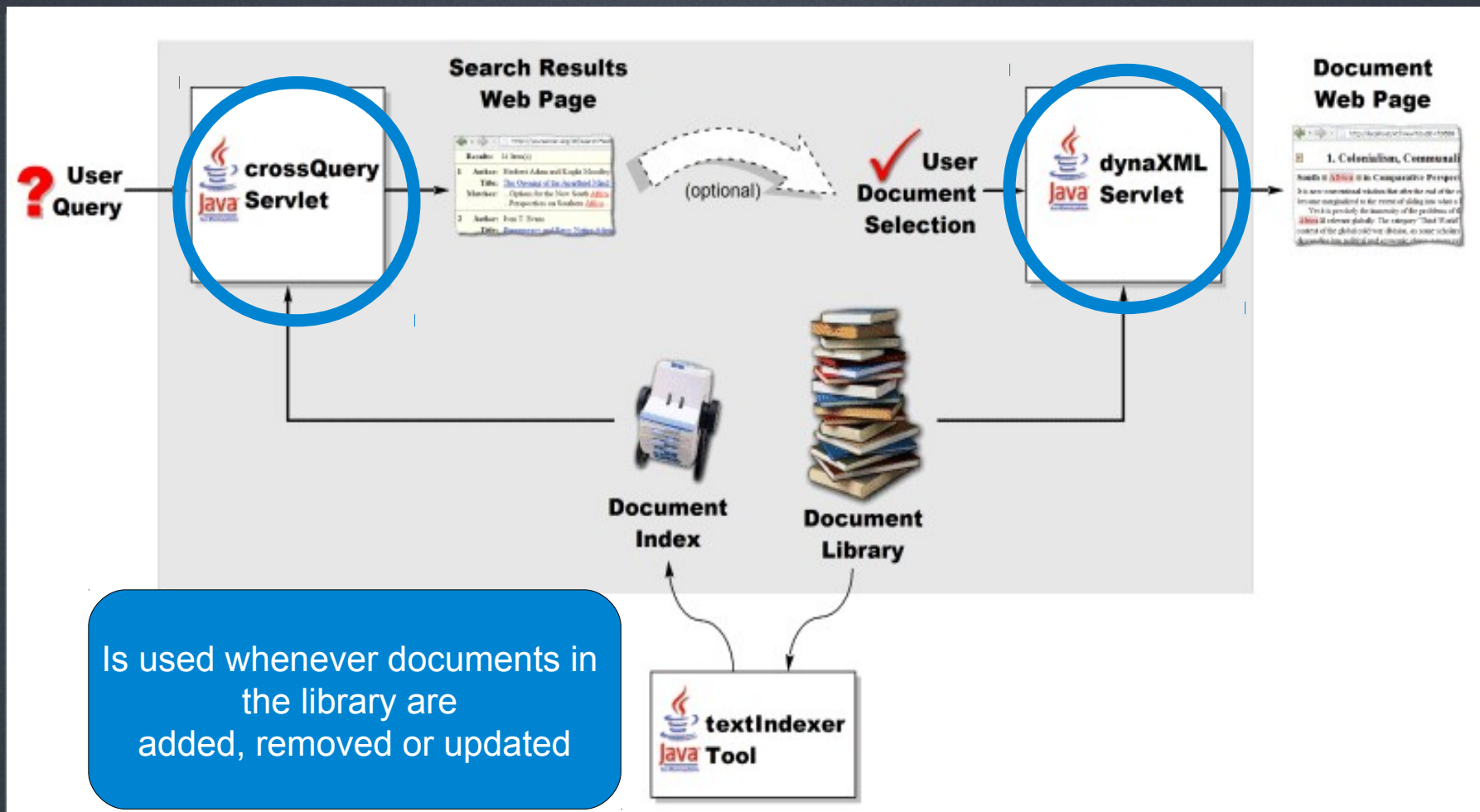


# Sample Implementations

- Diary of Robert Graves:
http://graves.uvic.ca/graves/site/index.xml
- Opera Liber
http://193.204.255.27/operaliber/index.php?page=/operaLiber/home
- The Anglo-Norman Dictionary
http://www.anglo-norman.net/
- Early Americans Digital Archives
http://mith.umd.edu/eada/intro.php
- Corpus Pieter Willems
http://ctb.kantl.be/corpora/CPWNL/
- Sermones.net

# XTF - Extensible Text Framework Architecture

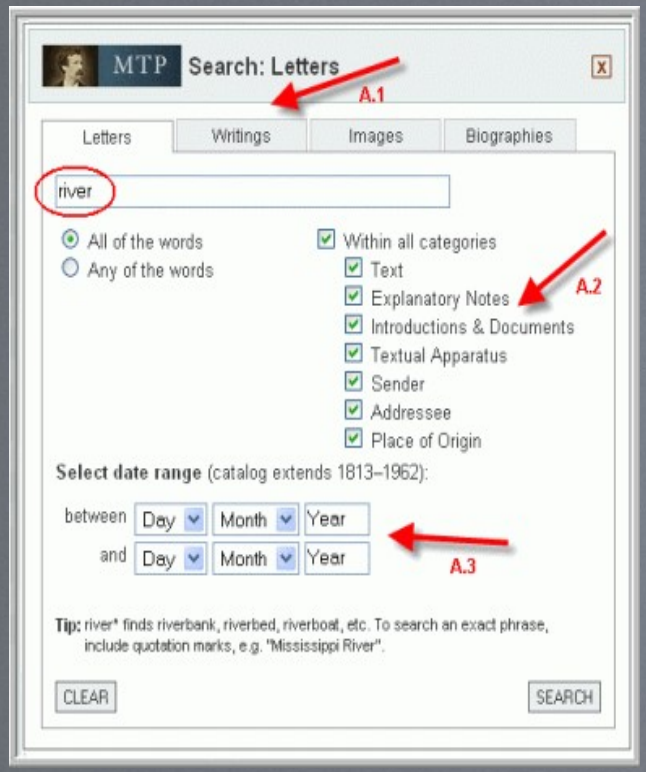
http://xtf.cdlib.org/




The Architecture of XTF. Developed by CDL

# Example of Query on a XTF-based Project

http://www.marktwainproject.org/

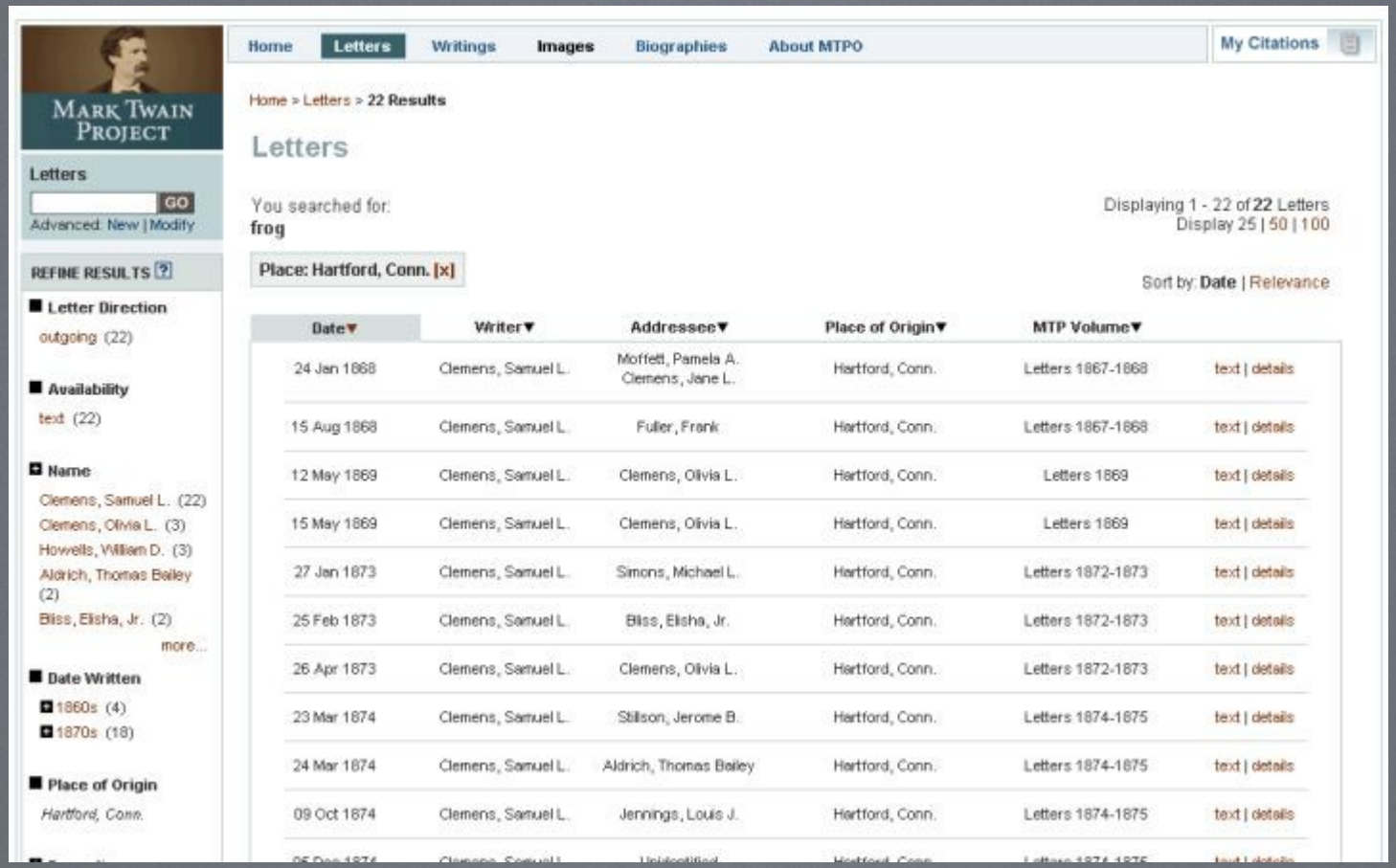



Search individual writings

Search and Browse

# Example of Query on a XTF-based Project

http://www.marktwainproject.org/




**Mark Twain Project - Searching for Objects**

# Example of a Hybrid Project

http://www.labibliotecaitaliana.it

# THANK YOU
# FOR YOUR TIME