Mar 25th, 10:00 AM - 12:20 PM

Open-Source Archives West: Replacing a Proprietary XML Database with BaseX

Tamara Marnell
Orbis Cascade Alliance, tmarnell@orbiscascade.org

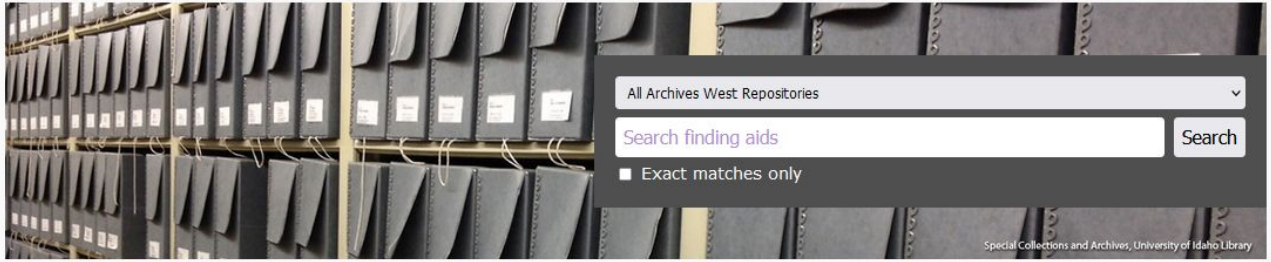
Open-Source Archives West

Replacing a Proprietary XML Database with BaseX

March 25, 2022
Tamara Marnell
Program Manager, Systems
Orbis Cascade Alliance

Background of Archives West

- The website has been operated by Orbis Cascade Alliance since 2007.
- Our repositories contain over 44,000 finding aids from six western states.
    - ID, MT, OR, UT, WA, and WY
- We use a schema based on Encoded Archival Description (EAD) Version 2002.
- The Alliance issues Archival Resource Keys (ARKs) as persistent identifiers.
- Finding aids are released under either CC BY or CC0.



Archives West provides access to descriptions of primary sources in the western United States, including correspondence, diaries or photographs. Digital reproductions of the materials are available in some cases.

By Repository

- Idaho
- Montana
- Oregon
- Utah
- Washington
- Wyoming

By Subject

- Agriculture and Natural Resources
- Arts, Humanities, and Social Sciences
- Business, Industry, Labor, and Commerce
- Education
- Environment and Conservation
- Immigration and American Expansion
- Places
- People, Ethnicity, and Culture
- Politics, Government, and Law
- Science, Technology and Health
- Social Life and Customs
- Material Type

All Archives West Repositories

native american tribes in oregon Search

Exact matches only

The following terms were excluded from the search: ***in****.*

Refine Search

Sort by Relevance

Repository

Confederated Tribes of the Siletz, Tribal Cultural Collections (82)

University of Washington Libraries, Special Collections (81)

Oregon Historical Society Research Library (33)

Oregon State University Libraries, Special Collections and Archives Research Center (33)

University of Oregon Libraries, Special Collections and University Archives (30)

Show more

Place

Siletz Indian Reservation (Or.) (21)

Newport (Or.) (18)

Siletz (Or.) (12)

Portland (Or.) (12)

465 finding aids

Results per Page 10

Page 1 of 47

Next page »

Oregon Native American Language Sound Recordings, 1962-1964

Repository: Oregon State University Libraries, Special Collections and Archives Research Center

Abstract: The Oregon Native American Language Sound Recordings were made by Joe E. Pierce during his study of Native American languages in Oregon and include interviews with members of the Coquille and Siletz tribes.

Native American Newspaper Collection, 1975-1993

Repository: Confederated Tribes of the Siletz, Tribal Cultural Collections

Abstract: The collection contains newspapers and clippings regarding issues and events affecting Native American tribes throughout the United States from 1975-1993.

Native American Maps Collection, 1875-1972

Repository: Oregon State University Libraries, Special Collections and Archives Research Center

Abstract: The Native American Maps Collection consists of maps of the Warm Springs Indian Reservation in north central Oregon; a map of the tribes of Alaska; and maps of reservation lands in the United States.

Reasons for Going “Open”

- The Alliance Council wrote a Commitment to Open Principles & Practices.
- Our proprietary XML database software was expensive.
- The servers required to run the proprietary software were expensive.
- Over 14 years, different developers made additions without renovations.

Rebuilding with BaseX

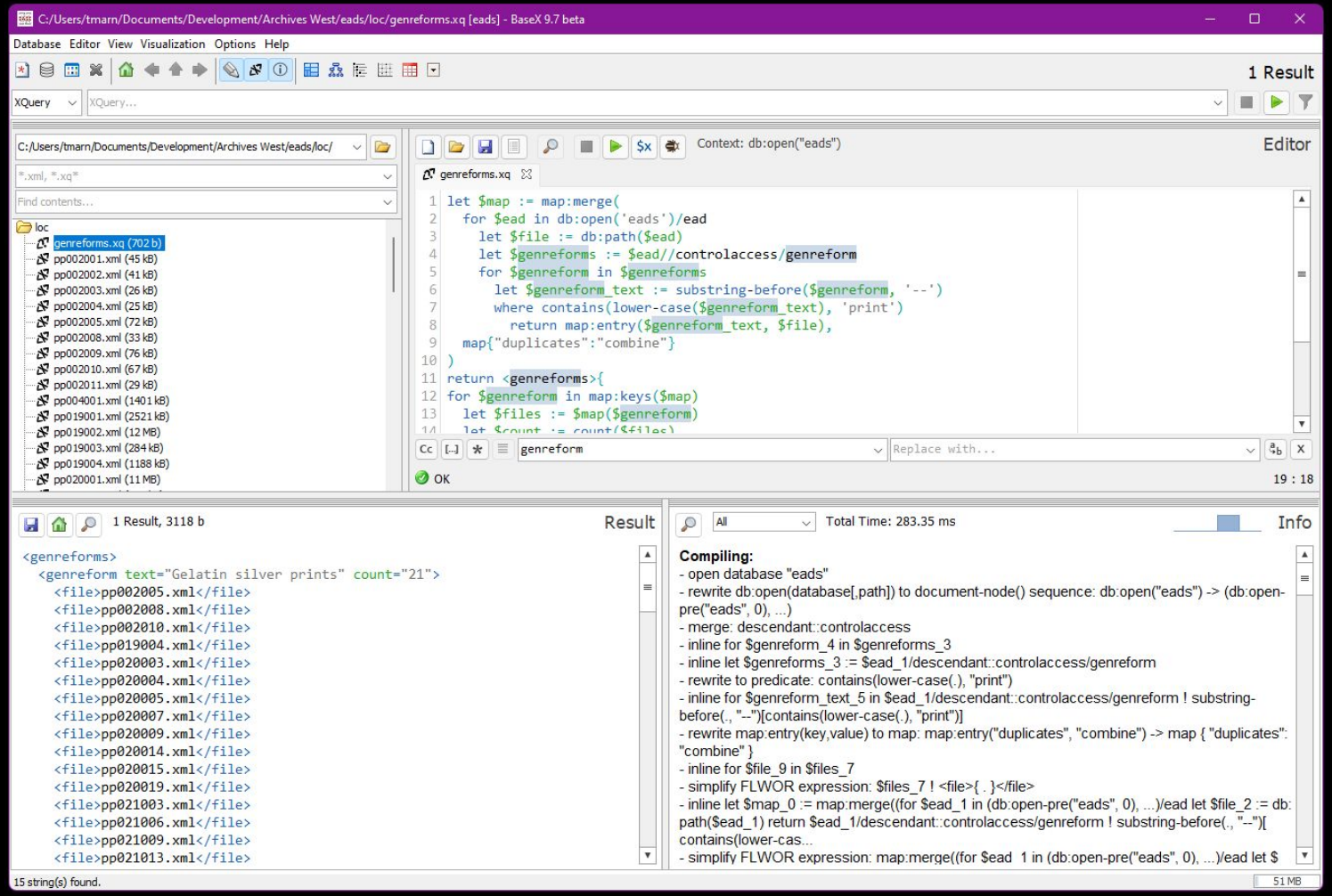
What Is BaseX?

“BaseX is a light-weight, high-performance and scalable XML Database and an XQuery 3.1 Processor with full support for the W3C Update and Full-Text extensions.”

- It’s a Java program that builds searchable databases out of XML files.
- XQuery is the programming language used to read, update, or full-text search the databases.
- XPath expressions select nodes within a document.

Client/Server Architecture

- The structure of BaseX is similar to MySQL.
  - A system-wide service stores the databases.
  - Clients send commands to build databases, query them, update or drop them.
- Multiple client types are available.
  - Command line
  - Graphical User Interface
  - Programming languages


C:/Users/tmarn/Documents/Development/Archives West/eads/loc/genreforms.xq [eads] - BaseX 9.7 beta
Database Editor View Visualization Options Help
1 Result
XQuery
XQuery...
C:/Users/tmarn/Documents/Development/Archives West/eads/loc/
*.xml, *.xq*
Find contents...
loc
genreforms.xq (702 b)
pp002001.xml (45 kB)
pp002002.xml (41 kB)
pp002003.xml (26 kB)
pp002004.xml (25 kB)
pp002005.xml (72 kB)
pp002008.xml (33 kB)
pp002009.xml (76 kB)
pp002010.xml (67 kB)
pp002011.xml (29 kB)
pp004001.xml (1401 kB)
pp019001.xml (2521 kB)
pp019002.xml (12 MB)
pp019003.xml (284 kB)
pp019004.xml (1188 kB)
pp020001.xml (11 MB)
Context: db:open("eads")
Editor
genreforms.xq
let $map := map:merge(
for $ead in db:open('eads')/ead
let $file := db:path($ead)
let $genreforms := $ead//controlaccess/genreform
for $genreform in $genreforms
let $genreform_text := substring-before($genreform, '--')
where contains(lower-case($genreform_text), 'print')
return map:entry($genreform_text, $file),
map{"duplicates":"combine"}
)
return <genreforms>{
for $genreform in map:keys($map)
let $files := $map($genreform)
genreform
Replace with...
OK
19 : 18
1 Result, 3118 b
Result
<genreforms>
<genreform text="Gelatin silver prints" count="21">
<file>pp002005.xml</file>
<file>pp002008.xml</file>
<file>pp002010.xml</file>
<file>pp019004.xml</file>
<file>pp020003.xml</file>
<file>pp020004.xml</file>
<file>pp020005.xml</file>
<file>pp020007.xml</file>
<file>pp020009.xml</file>
<file>pp020014.xml</file>
<file>pp020015.xml</file>
<file>pp020019.xml</file>
<file>pp021003.xml</file>
<file>pp021006.xml</file>
<file>pp021009.xml</file>
<file>pp021013.xml</file>
All
Total Time: 283.35 ms
Info
Compiling:
- open database "eads"
- rewrite db:open(database[,path]) to document-node() sequence: db:open("eads") -> (db:open-pre("eads", 0), ...)
- merge: descendant::controlaccess
- inline for $genreform_4 in $genreforms_3
- inline let $genreforms_3 := $ead_1/descendant::controlaccess/genreform
- rewrite to predicate: contains(lower-case(.), "print")
- inline for $genreform_text_5 in $ead_1/descendant::controlaccess/genreform ! substring-before(., "--")[contains(lower-case(.), "print")]
- rewrite map:entry(key,value) to map: map:entry("duplicates", "combine") -> map { "duplicates": "combine" }
- inline for $file_9 in $files_7
- simplify FLWOR expression: $files_7 ! <file>{ . }</file>
- inline let $map_0 := map:merge((for $ead_1 in (db:open-pre("eads", 0), ...)/ead let $file_2 := db:path($ead_1) return $ead_1/descendant::controlaccess/genreform ! substring-before(., "--")[contains(lower-cas...
- simplify FLWOR expression: map:merge((for $ead 1 in (db:open-pre("eads", 0), ...)/ead let $
15 string(s) found.
51 MB

```
// Build a database
function build_db($repo_id) {
  $repo = new AW_Repo($repo_id);
  $this->session->execute('CREATE DB eads' . $repo_id . ' ' . AW_REPOS . '/' . $repo->get_folder() . '/eads');
}

// Drop a database
function drop_db($repo_id) {
  $this->session->execute('DROP DB eads' . $repo_id);
}

// Add a finding aid to a database
function add_document($repo_id, $file) {
  $repo = new AW_Repo($repo_id);
  $this->session->execute('OPEN eads' . $repo_id);
  $this->session->execute('ADD ' . AW_REPOS . '/' . $repo->get_folder() . '/eads/' . $file);
  $this->session->execute('OPTIMIZE');
  $this->session->execute('CLOSE');
}

// Populate text index for a database
function index_text($repo_id) {
  try {
    $query = $this->get_query('index-text-db.xq');
    $query->bind("d", $repo_id);
    $query->bind("s", $this->get_stopwords());
    $query->execute();
    $query->close();
  }
  catch (BaseXException $e) {
    print $e->getMessage();
  }
}
```

The Importance of Indexes

- The performance of BaseX depends on the efficiency of the project's index structures and XQuery.
- BaseX offers built-in indexes like full-text.
- Use XQuery to build custom indexes as databases.
- In production, users should search index databases, not document databases.
    - Searching and navigating original XML documents is slow and CPU intensive.
    - BaseX will process one read or update request at a time.

```
<subjects>
  <subject term="Accidents--1920-1970.">
    <id>loc.pnp/eadpnp.pp021032</id>
    <id>loc.pnp/eadpnp.pp021033</id>
    <id>loc.pnp/eadpnp.pp021034</id>
  </subject>
  <subject term="Actors--1920-1940.">
    <id>loc.pnp/eadpnp.pp002004</id>
  </subject>
  <subject term="Actors--1950-1960">
    <id>loc.pnp/eadpnp.pp002009</id>
  </subject>
  <subject term="Actors--United States--1950-1990.">
    <id>loc.pnp/eadpnp.pp019004</id>
  </subject>
  <subject term="Adultery--1900-1910.">
    <id>loc.pnp/eadpnp.pp021004</id>
  </subject>
  <subject term="Aerial photographs--1940-1960">
    <id>loc.pnp/eadpnp.pp002009</id>
  </subject>
  <subject term="Aeronautics--1900-1950.">
    <id>loc.pnp/eadpnp.pp020008</id>
  </subject>
  [...]
</subjects>
```

Lessons Learned

“Free” Software Has a Cost

- The Archives West rebuild required half a year of intensive development and continued for two months after go-live.
  - Tweaks to full-text search matching and scoring.
  - Addition of dedicated production text indexes.
  - Batch jobs for optimizing indexes.
- Implementers need to develop skills as well as code.
  - BaseX for Newbies (PDF)
- An organization can pay for software, or they can pay an employee.

Questions?