Applying XQuery and OWL to The World Factbook, Wikipedia and Project Gutenberg

Kenneth B. Sall, Science Applications International Corporation (SAIC) and Ronald P. Reck, RRecktek LLC

December 7, 2006

XML 2006, Boston, MA
Agenda

• Background
• Problem Statement
• Data Sources:
  – Project Gutenberg, Wikipedia, The World Factbook
• Modifications to Data Sources
• eXist XML Database
• Queries We Used
• Future Work Potentials
Background

- Metadata Cards for Describing Project Gutenberg Texts – Ron Reck
- Project Gutenberg e-texts.
- Metadata represented by a combination of RDF, Dublin Core, and FOAF markup for 15,000 texts.
- Harvested Wikipedia HTML pages for 1,700+ authors, extracting each page’s text which contains author’s country/nationality, year of birth, year of death, etc.
- The World Factbook (produced by the CIA) is a source of detailed information about geography, people, government, economy, communications, and transportation for every country in the world. [Details in paper appendix.]
Problem Statement

- Find all the Project Gutenberg books written by male European [Scandinavian] authors in the 19th century.

Find all PG books & all authors. [PG]

Determine gender of authors. [W]

Define European countries. [WFB]

Find all European authors. [W]

Determine authors who wrote in 1800’s. [W]
Data Sources: Project Gutenberg

<table>
<thead>
<tr>
<th>Bibliographic Record</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Creator</strong></td>
</tr>
<tr>
<td><strong>Title</strong></td>
</tr>
<tr>
<td><strong>Language</strong></td>
</tr>
<tr>
<td><strong>LoC Class</strong></td>
</tr>
<tr>
<td><strong>Subject</strong></td>
</tr>
<tr>
<td><strong>Subject</strong></td>
</tr>
<tr>
<td><strong>EText-No.</strong></td>
</tr>
<tr>
<td><strong>Release Date</strong></td>
</tr>
<tr>
<td><strong>Copyright Status</strong></td>
</tr>
</tbody>
</table>

Download this ebook for free

<table>
<thead>
<tr>
<th>Formats Available For Download</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Edition</strong></td>
</tr>
<tr>
<td>11</td>
</tr>
</tbody>
</table>
Data Sources: Wikipedia

wget
http://en.wikipedia.org/wiki/Andrew_Lang
Andrew_Lang.html

- Found 1,733 unique Wikipedia authors pages of 4,702 PG authors.
Data Sources: The World Factbook [CIA]

- Country Profiles
  - Introduction (Background)
  - Geography
  - People
  - Government
  - Economy
  - Communications
  - Transportation
  - Military
  - Transnational Issues

- Reference Maps
- Rank Order Pages
- Flags of the World
- Notes and Definitions (explanation of fields)
- Appendices
Two centuries of Viking raids into Europe tapered off following the adoption of Christianity by King Olaf TRYGGVASON in 994. Conversion of the Norwegian kingdom occurred over the next several decades. In 1397, Norway was absorbed into a union with Denmark that lasted more than four centuries. In 1814, Norwegians resisted the union of their country to Sweden and adopted a new constitution. Sweden then invaded Norway but agreed to let Norway keep its constitution in return for accepting the union under a Swedish king. Rising nationalism throughout the 19th century led to a 1905 referendum granting Norway independence. Although Norway remained neutral in World War I, it suffered heavy losses to its shipping. Norway proclaimed its neutrality at the outset of World War II, but was nonetheless occupied for five years by Nazi Germany (1940-45). In 1949, neutrality was abandoned and Norway became a member of NATO. Discovery of oil and gas in adjacent waters in the late 1960s boosted Norway’s economic fortunes. The current focus is on containing spending on the extensive welfare system and planning for the time when petroleum reserves are depleted. In referendums held in 1972 and 1994, Norway rejected joining the EU.
Modifications to Sources: Project Gutenberg → RDF

- *dc:title* – title of an ebook
- *dc:language* – language an ebook was presented in
- *dc:creator* – author of an ebook
- *dc:available* – date an ebook became available in Project Gutenberg in the format YYYY-MM
- *foaf:name* – name of the PG text editor/translator/producer
- *foaf:mbox* – used if a producer’s email address was provided and if determinable
- *foaf:sha1* – adopted to express the checksum of the archive file
- Plus 11 other metadata properties described in our paper
<rdf:RDF
   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
   xmlns:dc="http://purl.org/dc/elements/1.1/"
   xmlns:dcterms="http://purl.org/dc/terms/"
   xmlns:pg="http://iama.recktek.com/daml/ont/pg#"
   xmlns:foaf="http://xmlns.com/foaf/0.1/" >

      <dc:title>A Horse's Tale</dc:title>
      <dcterms:available rdf:datatype="http://www.w3c.org/2000/10/XMLSchema#date">1997-10</dcterms:available>
      <pg:etext>1086</pg:etext>
      <pg:producer>David Price</pg:producer>
      <foaf:person rdf:parseType="Resource">
         <foaf:name>David Price</foaf:name>
         <foaf:mbox rdf:resource="mailto:ccx074@coventry.ac.uk"/>
      </foaf:person>
      <pg:linecount rdf:datatype="http://www.w3c.org/2000/10/XMLSchema#int">2365</pg:linecount>
      <pg:wordcount rdf:datatype="http://www.w3c.org/2000/10/XMLSchema#int">19257</pg:wordcount>
      <pg:charactercount rdf:datatype="http://www.w3c.org/2000/10/XMLSchema#int">107174</pg:charactercount>
      <foaf:sha1>386126b01230dd062894742701cb208c526471db</foaf:sha1>
      <pg:fcount>ASCII English text, with CRLF line terminators</pg:fcount>
      <pg:csnum rdf:datatype="http://www.w3c.org/2000/10/XMLSchema#int">44497</pg:csnum>
      <pg:ucsize rdf:datatype="http://www.w3c.org/2000/10/XMLSchema#int">109539</pg:ucsize>
      <pg:cratio rdf:datatype="http://www.w3c.org/2000/10/XMLSchema#int">59.4</pg:cratio>
   </book:Book>
</rdf:RDF>

- For each of the 1733 unique author pages downloaded from Wikipedia, determine value of four attributes for each author:
  - Gender, Nationality, Year of Birth, Year of Death

“Born in Selkirk, Scotland (March 31, 1844 – July 20, 1912) was a prolific Scots man of letters. He was a poet, novelist, and literary critic, and contributor to anthropology.”

“Agatha Mary Clarissa, Lady Mallowan, DBE (15 September 1890 – 12 January 1976), also known as Dame Agatha Christie, was an English crime fiction writer.”

“A Alexis-Charles-Henri Clérel de Tocqueville (July 29, 1805–April 16, 1859) was a French political thinker and historian.”

HTML not XML! “title” for instance rather than class info.

```
<a href="/wiki/Scotland" title="Scotland">Scotland</a>, 
<a href="/wiki/March_31" title="March 31">March 31</a>, 
<a href="/wiki/1844" title="1844">1844</a>
```
## Nouns to Determine Nationality

<table>
<thead>
<tr>
<th>Actor</th>
<th>Actress</th>
<th>Anthropologist</th>
<th>Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biographer</td>
<td>Biologist</td>
<td>Businessman</td>
<td>Correspondent</td>
</tr>
<tr>
<td>Diplomat</td>
<td>Divine</td>
<td>Educator</td>
<td>Engineer</td>
</tr>
<tr>
<td>Entomologist</td>
<td>Essayist</td>
<td>Explorer</td>
<td>Explorer</td>
</tr>
<tr>
<td>Feminist</td>
<td>Geographer</td>
<td>Governess</td>
<td>Grammarian</td>
</tr>
<tr>
<td>Historian</td>
<td>Journalist</td>
<td>Jurist</td>
<td>Lawyer</td>
</tr>
<tr>
<td>Naturalist</td>
<td>Novelist</td>
<td>Officer</td>
<td>Philanthropist</td>
</tr>
<tr>
<td>Philosopher</td>
<td>Poet</td>
<td>Politician</td>
<td>Psychologist</td>
</tr>
<tr>
<td>Publicist</td>
<td>Publisher</td>
<td>Statesman</td>
<td>Statesman</td>
</tr>
<tr>
<td>Suffragist</td>
<td>Thinker</td>
<td>Writer</td>
<td></td>
</tr>
</tbody>
</table>
DAML: CIA World Factbook in OWL: Norway

Separate files per country, so nodeIDs not unique across set.
Our Node Modifications to DAML OWL

<?xml version='1.0'?>
<rdf:Description rdf:nodeID="no-A0">
  <NS1:note>including 109 km of expressways</NS1:note>
</rdf:Description>

<rdf:Description rdf:nodeID="no-A3">
    <NS1:latitude>62.0</NS1:latitude>
    <NS1:longitude>10.0</NS1:longitude>
  </rdf:type>
</rdf:Description>

<rdf:Description rdf:nodeID="no-A85">
  <NS1:note>includes mainland 3,419 km, large islands 2,413 km, long fjords, numerous small
  islands, and minor indentations 16,093 km</NS1:note>
</rdf:Description>

<rdf:Description rdf:about ="http://www.daml.org/2001/09/countries/fips#NO">
  <NS1:location>Northern Europe, bordering the North Sea and the North Atlantic Ocean, west of Sweden</NS1:location>
  <NS1:geographicCoordinates rdf:nodeID ="no-A3"/>
  <NS1:mapReferences>Europe</NS1:mapReferences>
  <NS1:totalArea>324220</NS1:totalArea>
  <NS1:landArea>307860</NS1:landArea>
  <NS1:waterArea>16360</NS1:waterArea>
  <NS1:comparativeArea>slightly larger than New Mexico</NS1:comparativeArea>
  <NS1:landBoundaries>2544</NS1:landBoundaries>
  <NS1:coastline rdf:ID ="no-A85">21925</NS1:coastline>
</rdf:Description>

Added FIPS 10-4 country code to nodeIDs for uniqueness.
Sample Entry from Our author-wiki.rdf

<dc:creator
    rdf:about="http://skosaurus.rrecktek.com/ont/author#andrew_lang">
    <au:Gender>male</au:Gender>
    <au:Nationality>Scottish</au:Nationality>
    <au:FromCountry>Scotland</au:FromCountry>
    <au:BirthYear>1844</au:BirthYear>
    <au:DeathYear>1912</au:DeathYear>
</dc:creator>
Excerpt (1) from author-wiki.rdf

```xml
1    <rdf:RDF
2        xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
3            xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
4              xmlns:dc="http://purl.org/dc/elements/1.1/"
5              xmlns:owl="http://www.w3.org/2002/07/owl#"
6                  xmlns:au="http://iana.rrecktek.com/daml/ont/au">
7                  <owl:Ontology rdf:about="">
8                      <owl:versionInfo rdf:datatype="http://www.w3.org/2001/XMLSchema#string">
9                          Author ontology version 1.0</owl:versionInfo>
10                         </owl:Ontology>
11                            <owl:Class rdf:ID="BirthYear">
12                                <rdfs:label xml:lang="en">Author's Year of Birth</rdfs:label>
13                              </owl:Class>
14                            <owl:Class rdf:ID="DeathYear">
15                                <rdfs:label xml:lang="en">Author's Year of Death</rdfs:label>
16                              </owl:Class>
17                            <owl:Class rdf:ID="Nationality">
18                                <rdfs:label xml:lang="en">Author's Birth Nationality</rdfs:label>
19                              </owl:Class>
20                            <owl:Class rdf:ID="Gender">
21                                <rdfs:label xml:lang="en">Author's Gender</rdfs:label>
22                              </owl:Class>
23                            <owl:ObjectProperty rdf:ID="hasBirthYear">
24                                <rdf:type rdf:resource="FunctionalProperty" />
25                                      <owl:Restriction owl:minCardinality="1" owl:maxCardinality="1" />
26                                      <rdfs:domain rdf:resource="#Birth" />
27                                      <rdfs:range rdf:resource="#BirthYear" />
28                              </owl:ObjectProperty>
```

Excerpt (2) from author-wiki.rdf

```xml
<owl:ObjectProperty rdf:ID="hasDeathYear">
  <rdf:type rdf:resource="FunctionalProperty"/>
  <owl:Restriction owl:minCardinality="1" owl:maxCardinality="1"/>
  <rdfs:domain rdf:resource="#Death"/> <rdfs:range rdf:resource="#DeathYear"/>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="hasGender">
  <rdf:type rdf:resource="FunctionalProperty"/>
  <owl:Restriction owl:minCardinality="1" owl:maxCardinality="1"/>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="hasNationality">
  <rdf:type rdf:resource="FunctionalProperty"/>
  <owl:Restriction owl:minCardinality="1" owl:maxCardinality="1"/>
</owl:ObjectProperty>

<dc:creator rdf:about="http://skosaurus.rrecktek.com/ont/author#a_a_milne">
</dc:creator>

<dc:creator rdf:about="http://skosaurus.rrecktek.com/ont/author#andrew_lang">
</dc:creator>
```
Open Source native XML database with efficient, index-based XQuery processing, automatic indexing, extensions for full-text search.

- Adding RDF files to the XML database
- Using the XQuery Sandbox feature to develop and test XQueries
- Debugging our XQueries via the as-you-type XQuery interpreter
Queries Used in Our Solution - Namespaces

declare namespace dc="http://purl.org/dc/elements/1.1/";
declare namespace owl="http://www.w3.org/2002/07/owl#";
declare namespace rdfs="http://www.w3.org/2000/01/rdf-schema#";
declare namespace foaf="http://xmlns.com/foaf/0.1/";
declare namespace rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#";
declare namespace NS0="http://www.w3.org/2002/07/owl#";
declare namespace NS1="http://www.daml.org/2003/09/factbook/factbook-ont#";
declare namespace au="http://iama.rrecktek.com/daml/ont/au";
Queries Used – Find Continents

for $x$ in
doc("/db/metacards/factbook.rdf")//rdf:
    Description/di distinct-values(NS1:mapReferences)

order by $x$ ascending

return <Continent>{$x}</Continent>
Queries Used – Find Continent of Country

for $x$ in
doc("/db/metacards/factbook.rdf")//rdf:Desc ription[@rdf:about[ends-with(., '#NO')]])
return $x/NS1:mapReferences

(where “NO” is FIPS 10-4 code for Norway) returns
<NS1:mapReferences>Europe</NS1:mapReferences>
Queries Used – Get Country Profile

for $x in
doc("/db/metacards/factbook.rdf")//rdf:Description[@rdf:about[ends-with(., 'NO')]]
return $x

Wrapper element (around hundreds of lines) returned:

<rdf:Description rdf:about="http://www.daml.org/2001/09/countries/fips#NO">
Queries Used – Find 45 Countries in Europe

for $country in
doc("/db/metacards/factbook.rdf")//rdf:Description

where starts-with
  ($country/NS1:mapReferences,"Europe")

return
  string($country/NS1:conventionalShortCountryName)
Queries Used – Continent-to-Countries Function

(: Function to return all countries of given continent. :) 

declare function local:Continent-to-Countries($Continent as xs:string)

{for $country in 
   doc("/db/metacards/factbook.rdf")//rdf:Description
where starts-with ($country/NS1:mapReferences,$Continent)
return
string($country/NS1:conventionalShortCountryName)
}

(: call function :) 

for $country in (local:Continent-to-Countries("Europe"))
return $country
Queries Used – Male Authors from 3 Countries

for $country in ("England", "France", "Germany")

for $author in doc("/db/metacards/author-wiki.rdf")//dc:creator

where starts-with($author/au:BirthYear,"18") and
starts-with($author/au:Gender,"male") and
contains($author/au:FromCountry, $country)

return substring-after(string($author/@rdf:about),'#')
Queries – Male European Authors in 19th Century

for $country in (local:Continent-to-Countries("Europe")) for $author in doc("/db/metacards/author-wiki.rdf")//dc:creator where
starts-with($author/au:BirthYear,"18") and
starts-with($author/au:DeathYear,"18") and
starts-with($author/au:Gender,"male") and
contains($author/au:FromCountry, $country)
return substring-after(string($author/@rdf:about),'#')
Final Query in eXist Sandbox

XQuery Sandbox

Imports

declare namespace dc="http://purl.org/dc/elements/1.1";
declare namespace owl="http://www.w3.org/2002/07/owl#";
declare namespace rdfs="http://www.w3.org/2000/01/rdf-schema#";
declare namespace xsd="http://www.w3.org/2001/XMLSchema#";
declare namespace rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#";
declare namespace ns1="http://www.daml.org/2002/03/daml+oOnt#";
declare namespace au="http://www.rtektek.com/daml/ont/au#";

declare function local:Continent-to-Countries([Country as xs:string])

for $country in doc("/db/metacode/Facebook.rdf")/rdf:Description
where starts-with($country/NS1:conventions/Continent, "Africa")
return string($country/NS1:conventions/ShortCountryName)

for $country in (local:Continent-to-Countries("Europe")
for $author in doc("/db/metacode/author-wiki.rdf")/decreator where
starts-with($author/xs:birthYear, "19") and
starts-with($author/xs:deathYear, "19") and
starts-with($author/xs:gender, "Male") and
starts-with($author/xs:country, "Country")
return substring-after(string($author/xs:about), "," )</Author>

Send  Clear  Check  Display:  20  Show Options

Found 64 in 5.703 seconds.

Showing items 1 to 20

1. <Author>adriatic_coffee</Author>
2. <Author>charlie_le_poster</Author>
3. <Author>hendrik_conscience</Author>
4. <Author>julius_kabig</Author>
5. <Author>pieter_bloem</Author>
6. <Author>manic__mulligan</Author>
7. <Author>christian_sanderson</Author>
8. <Author>herman_bang</Author>
9. <Author>adriatic_coffee</Author>
eXist Java Client – Admin and Cmd Line

type help or ? for help.
exist:/db> cd "metacards"
exist:/db/metacards>
eXist Java Client – Query Dialog
Future Work Potentials

• Conduct performance comparisons between eXist and several key commercial XML database products such as:
  - DataDirect XQuery
  - MarkLogic Server
  - TigerLogic XML Data Management Server by RainingData

• Obtain The World Factbook database from the CIA, convert to XML and then to OWL

• Consider methods that could draw inferences about which countries are considered Scandinavian (rather than hard-coding a short list).
Future Work Potentials (continued)

- Expand Wikipedia page analysis (e.g., category data) to change link “title” attribute to something actually useful, possibly by referencing a fourth data source, WordNet.

- Investigate Wikipedia3, a conversion of the English Wikipedia into RDF, a dataset containing approximately 47 million triples, updated monthly. Wikipedia3 uses a custom ontology with elements from SKOS, Dublin Core, OWL & RDF. For example, Wikipedia3 could serve as a quality check or possibly replacement for our birth year, death year, and nationality triples.

- Investigate other technology this is relevant to our purposes including:
  - SPARQL (SPARQL Query Language for RDF)
  - SWRL (Semantic Web Rule Language Combining OWL and RuleML)
Future Work Potentials (continued)

- Could improve the analysis of the Wikipedia pages 2 ways.
  1. Leverage the information presented by a page’s category data as demonstrated in Danny Ayers Wikipedia3 effort.
  2. Each page’s hyperlinks can be typed in a manner similar to the method described in a paper entitled Wikipedia and the Semantic Web: The Missing Links by Krotzsch et al.

<table>
<thead>
<tr>
<th>Resource</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>eXist Open Source XML Database</td>
<td><a href="http://www.exist-db.org/">http://www.exist-db.org/</a></td>
</tr>
<tr>
<td>Gutenberg Texts</td>
<td></td>
</tr>
<tr>
<td>Mondial Database in XML</td>
<td><a href="http://www.dbis.informatik.uni-goettingen.de/Mondial/#XML">http://www.dbis.informatik.uni-goettingen.de/Mondial/#XML</a></td>
</tr>
<tr>
<td>Project Gutenberg</td>
<td><a href="http://www.gutenberg.org/wiki/Main_Page">http://www.gutenberg.org/wiki/Main_Page</a></td>
</tr>
<tr>
<td>Resource Description Framework (RDF)</td>
<td><a href="http://www.w3.org/RDF/">http://www.w3.org/RDF/</a></td>
</tr>
<tr>
<td>Web Ontology Language (OWL)</td>
<td><a href="http://www.w3.org/2004/OWL/">http://www.w3.org/2004/OWL/</a></td>
</tr>
<tr>
<td>Wikipedia3 [RDF]</td>
<td><a href="http://labs.systemone.at/wikipedia3">http://labs.systemone.at/wikipedia3</a></td>
</tr>
<tr>
<td>XML Query (XQuery)</td>
<td><a href="http://www.w3.org/XML/Query/">http://www.w3.org/XML/Query/</a></td>
</tr>
</tbody>
</table>