CDL CALIFORNIA DIGITAL LIBRARY

eXtensible Text Framework (XTF): Building a Digital Publishing Framework

California Digital Library

Kirk Hastings
Martin Haye

XTF
Topics

- Digital publishing at CDL
- What XTF is (and isn't)
- Design and Features
- Walkthrough: how to publish using XTF
- Wrap-up

Digital Publishing at CDL

- What is CDL?
  - University of California's 11th university library
  - Among world's largest digital libraries
  - We help campuses share resources & holdings
- What do we mean by “Digital Publishing”?
  - Full access to large text collections
  - Quickly navigate large documents
  - Intelligent links within and between docs
  - Search results shown in context

Problem:
Diverse Data

- Our collections: many and diverse
  - eScholarship Editions: Texts and monographs
    - A single text may be 10 megabytes or more
  - OAC: eTexts, finding aids
  - Special collections
    - Free Speech Movement Digital Archive
    - Japanese American Relocation Digital Archives
    - Tobacco Control Archives
    - Hundreds more

Problem:
Proprietary Software

- Digital Publishing Products
  - “Black box” (no control over fixes & features)
  - Often not standards-based
  - Tech companies have short lifespans
  - Support often spotty
  - $$$$$

Problem: Constant Flux

“Hey guys, what if we...”

“That's great, but could we add this too?”

“It should be pretty easy to...”

“You know what would be cool...”

“I just threw in 2500 new files...”

Brian Tingle, a source of constant flux

Problem: Deployment Speed

- Collections growing and multiplying
- Users don't want to wait a year for access
- Programmers are expensive
- Look & feel goes stale quickly
- Barrage of feature requests

Our Solution: XTF

- XTF is:
  - A framework to build search & display services
  - Web and XML centric
  - Focused on textual data
  - Simple to install and configure
  - Good for rapid prototyping and fast deployment

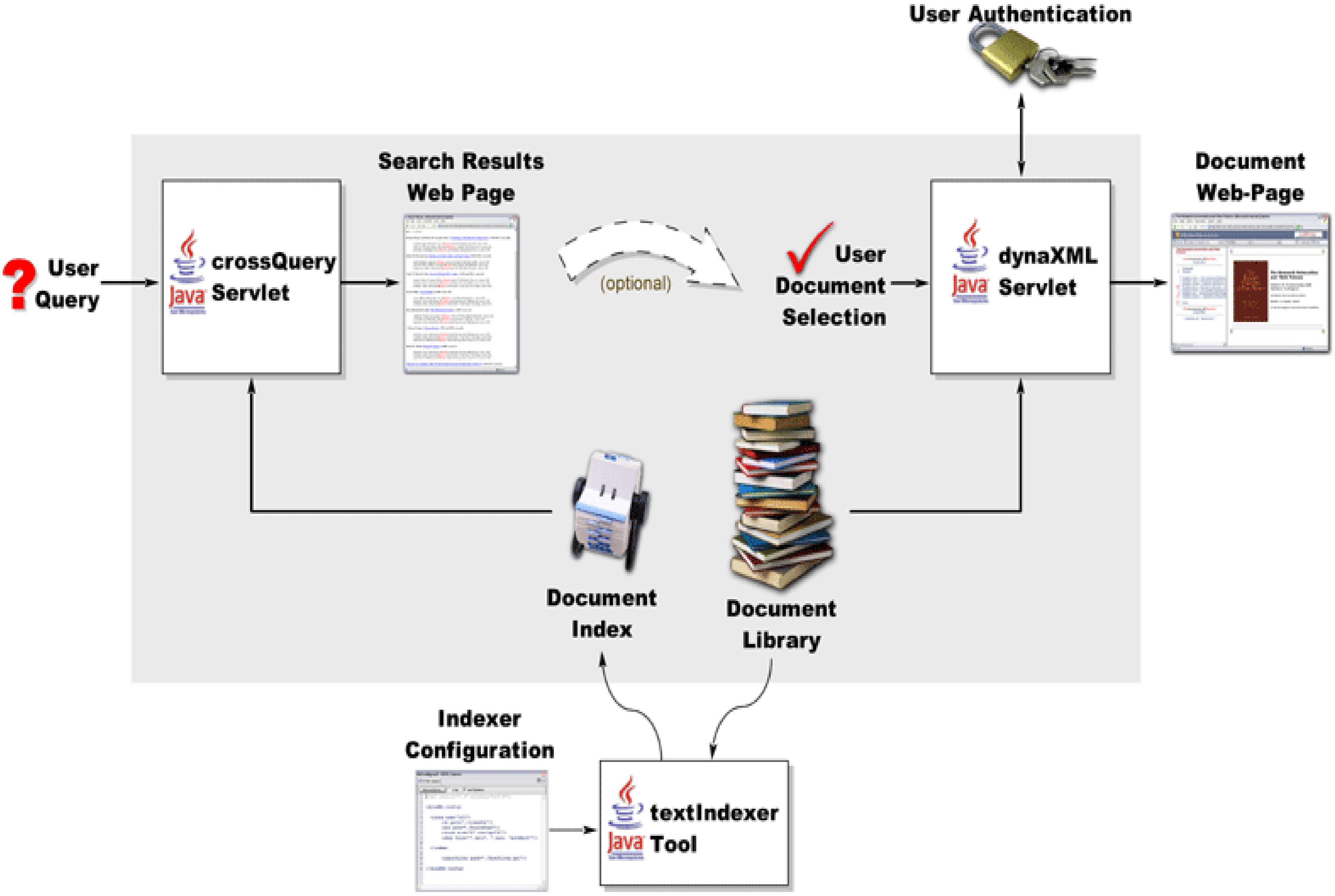
What XTF is Not

- XTF isn't for everyone.
  - It is not a content management system (versioning, archiving, peer review, etc.)
  - Not a content editing environment
  - Not built for remote administration
  - Not a true XML database

XTF is Open Source, Standards Based

- Based on free, open-source tools:
  - Java 2
  - Lucene 1.4 full-text search toolkit
  - Saxon 8.0 XSLT processor
- XTF itself is open-source (BSD license)
- No native code – pure Java and XSLT 2.0
- Runs on SunOS, Linux, Windows
- Drops right in to Tomcat or Resin
- UNICODE support throughout

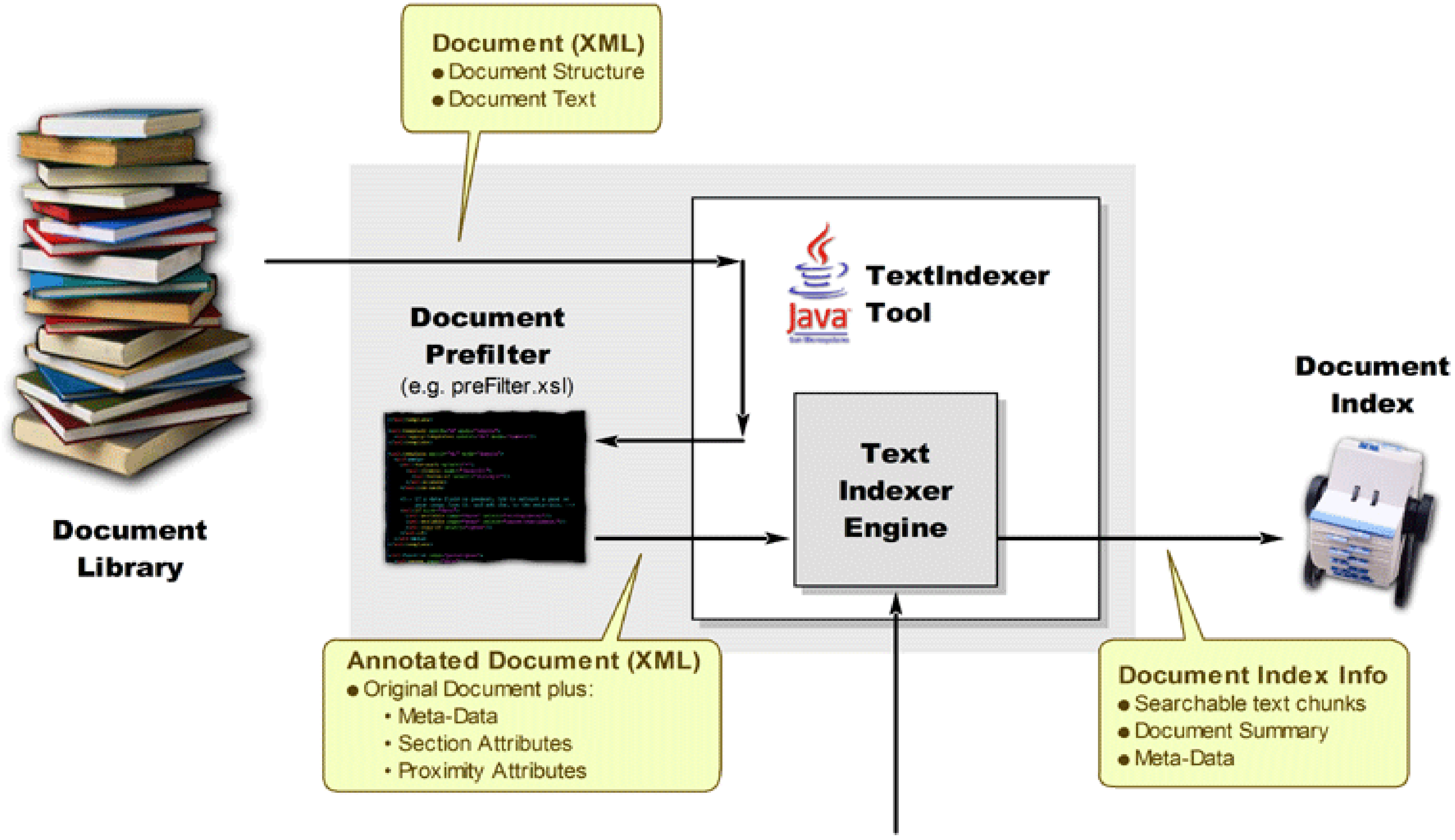
XTF System Overview



Modular

- Use ***crossQuery*** servlet to search, ***dynaXML*** servlet to display and navigate. Mix and match.
- Stylesheets govern flow of data – no Java programming required
- Easy to add features incrementally
- 100% configurable “look and feel”
- Skin & slice: one system can have several interfaces and multiple “brands”

Indexing Process



Document (XML)

- Document Structure
- Document Text

Document Library

Document Prefilter
(e.g. preFilter.xsl)

TextIndexer Tool

Text Indexer Engine

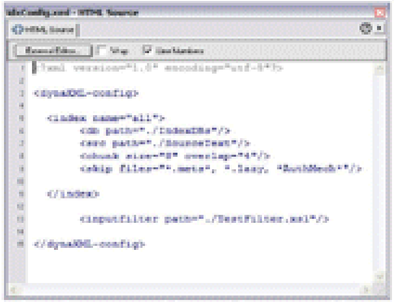
Document Index

Annotated Document (XML)

- Original Document plus:
  - Meta-Data
  - Section Attributes
  - Proximity Attributes

Document Index Info

- Searchable text chunks
- Document Summary
- Meta-Data



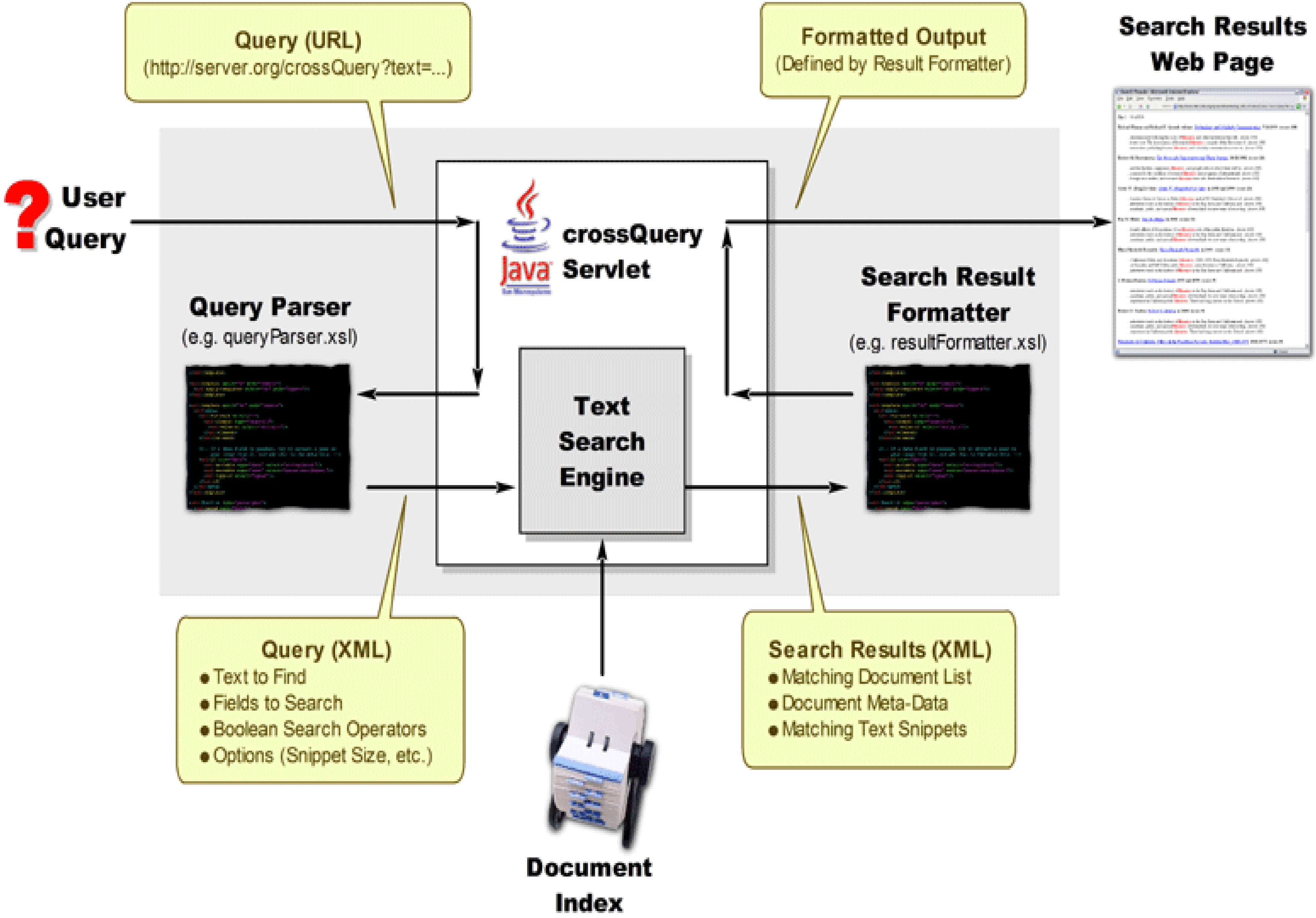
```
<?xml version="1.0" encoding="utf-8"?>

<dynaXML-config>

  <index name="all">
        <db path="./IndexDBs"/>
        <src path="./SourceText"/>
        <chunk size="8" overlap="4"/>
        <skip files="*.meta", "*.lazy, "AuthMech*"/>

  </index>

        <inputfilter path="./TestFilter.xsl"/>

</dynaXML-config>
```

Indexer Configuration
(e.g.textIndexer.conf)

Indexing

- Input filters adapt to many XML doc types
  - Coming soon: Non-XML formats (e.g. PDF)
- XTF is agnostic regarding:
  - Document identifiers
  - Filesystem organization
  - Meta-data storage
- Incremental indexing
  - Simply update filesystem then run indexer.

crossQuery servlet



Query (URL)
(http://server.org/crossQuery?text=...)

Formatted Output
(Defined by Result Formatter)

Search Results
Web Page

? User
Query

crossQuery
Servlet

Query Parser
(e.g. queryParser.xsl)

Search Result
Formatter
(e.g. resultFormatter.xsl)

Text
Search
Engine

Query (XML)
- Text to Find
- Fields to Search
- Boolean Search Operators
- Options (Snippet Size, etc.)

Search Results (XML)
- Matching Document List
- Document Meta-Data
- Matching Text Snippets

Document
Index

Flexible Search/Display

- One query, many collections
  - XTF enables “Virtual collections”
- Output filters for various styles and modes
  - e.g. simple vs. advanced search form, results in brief vs. long format, etc.
- Query parsers for different search interfaces
  - Interface to other query protocols

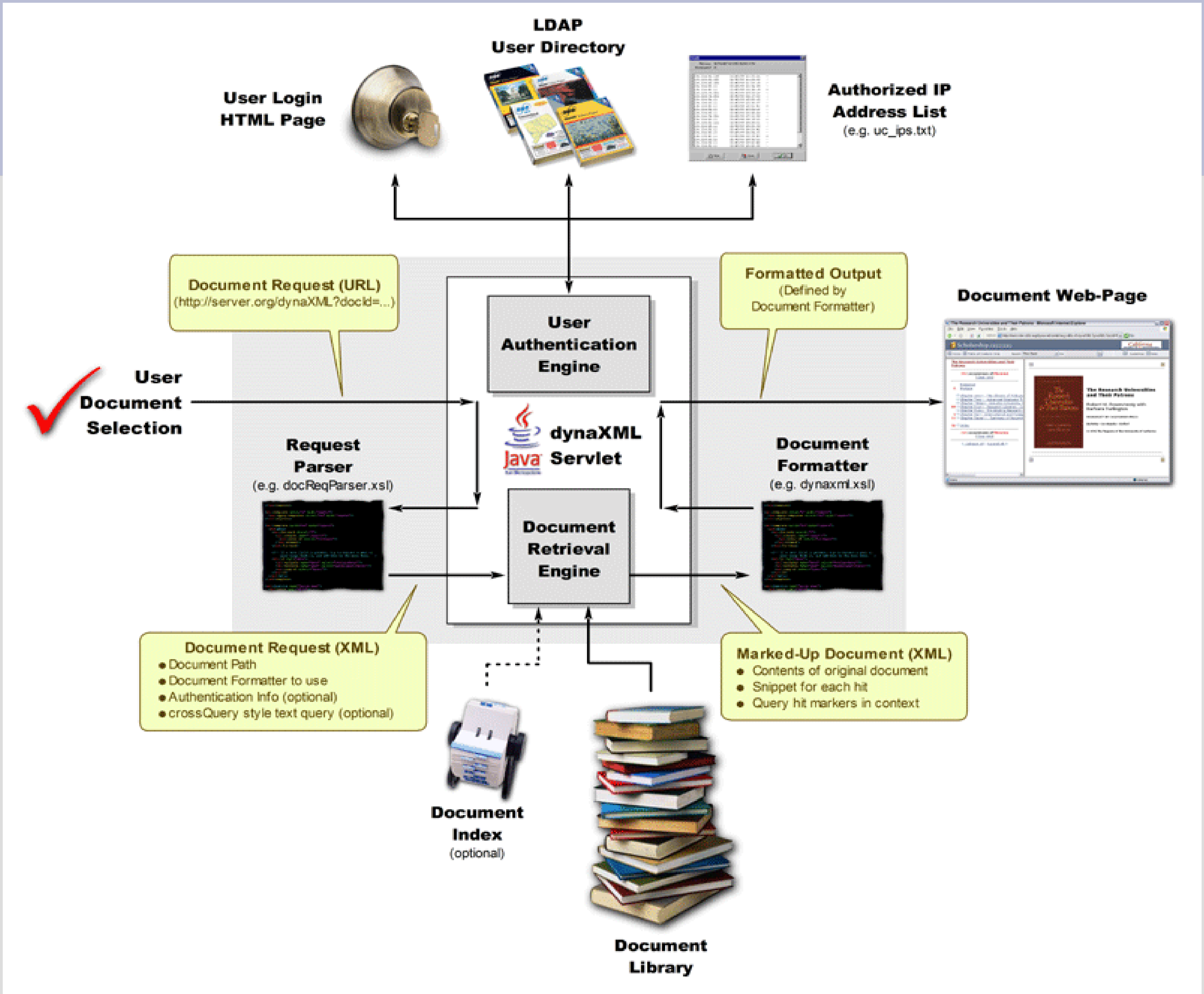
Query Power

- Many operators
  - AND, OR, NEAR, NOT, phrase, range, wildcard
- Arbitrarily complex queries
  - Combine full-text search with meta-data
  - Unusual queries like "dynamic duo" near "red phone"
- Structure-aware searching
  - e.g. search only headings, or only bibliographies
  - But must pre-define which structural elements to search

More Power

- Fixed-length snippets
  - Of course with keywords in context
- Sort by relevance, or any meta-data fields
- No penalty for huge documents
  - XTF “lazily” pulls in only those parts used by a particular request (e.g. show Chapter 1)
- Scalable
  - Lucene proven with extremely large data sets
- Authentication: IP lists, LDAP, or external (e.g. Shibboleth)

dynaXML servlet



Document Request (URL)
(http://server.org/dynaXML?docId=...)

Formatted Output
(Defined by
Document Formatter)

Document Request (XML)
- Document Path
- Document Formatter to use
- Authentication Info (optional)
- crossQuery style text query (optional)

Marked-Up Document (XML)
- Contents of original document
- Snippet for each hit
- Query hit markers in context

Technical Sidebar: How We Adapted Lucene

- XSLT prefilter flattens XML structure
  - Series of text blocks tagged with structual info
- Chunking for efficient searches
  - Lucene performs best with small documents
  - XTF breaks large docs into 200-word chunks
  - Chunks overlap to detect a hit starting in one and ending in the next.

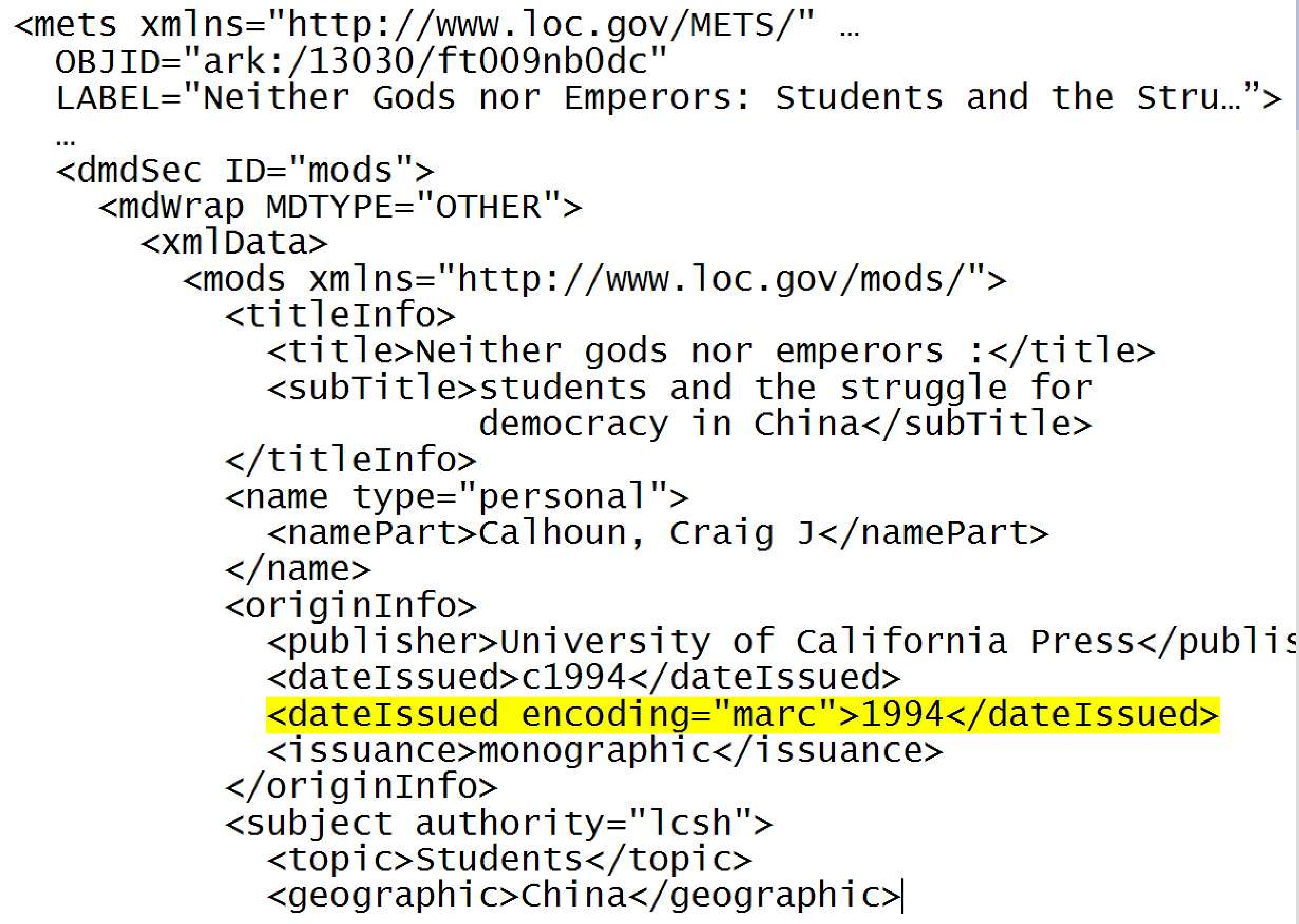
Technical Sidebar: How We Adapted Lucene

- Marking search hits in context
  - Lucene doesn't pinpoint location of hits
  - XTF dynamically annotates original XML doc with ranked hits, then sends to XSLT formatter
- Robust, efficient stop-word handling
  - “the, a, an, it, on...”
  - People do use them, and expect corresponding results (for speed, Lucene ignores stop-words)
  - XTF secretly joins stop-words to adjacent words, forming “*n-grams*”

Walkthrough

- Goal: Modify an existing service
  - Extract new meta-data
  - Add ability to sort by date
  - Customize search form & result display
  - Customize object display

METS Data

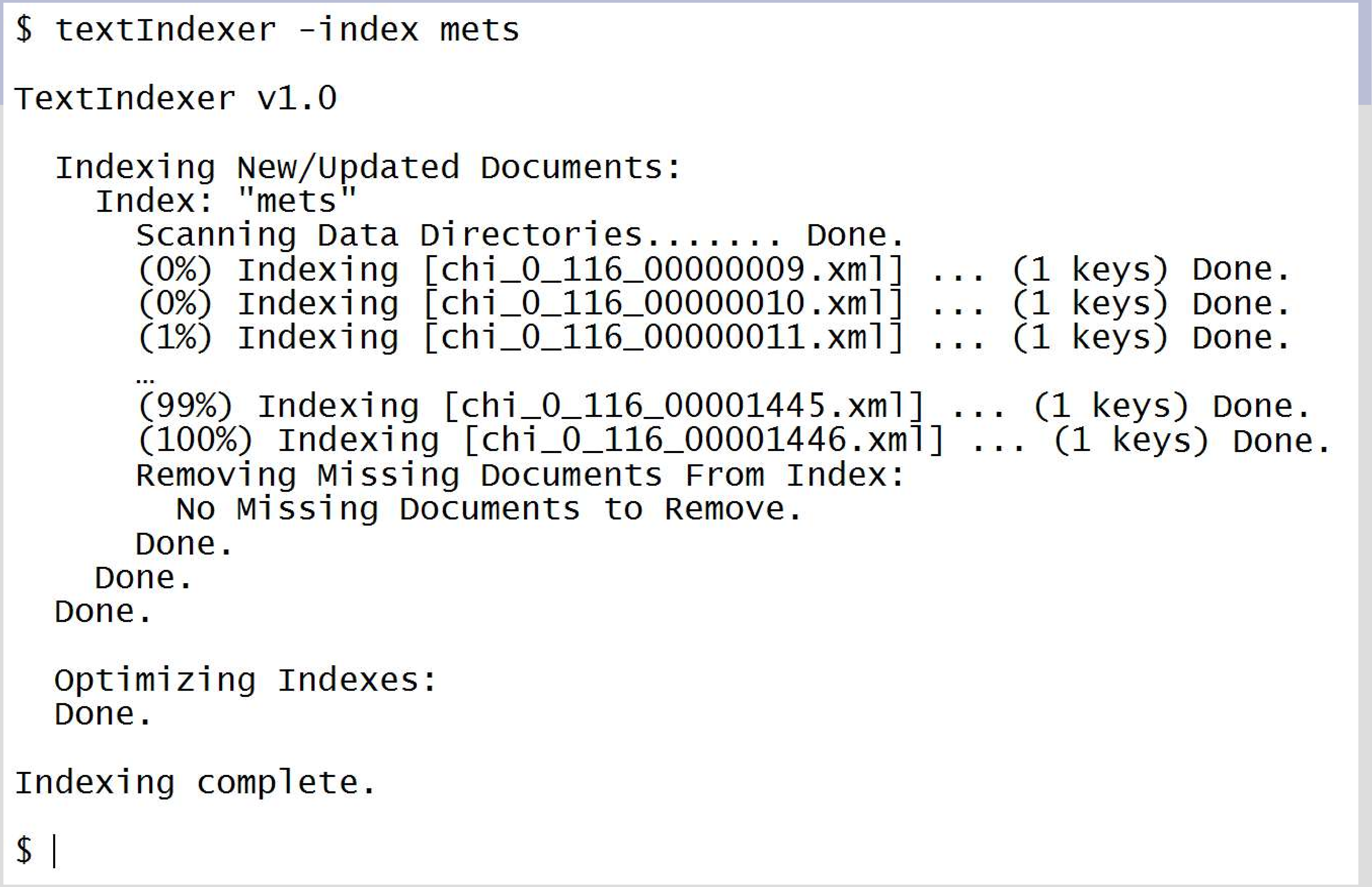


```
<mets xmlns="http://www.loc.gov/METS/" …
  OBJID="ark:/13030/ft009nb0dc"
  LABEL="Neither Gods nor Emperors: Students and the Stru…”>
  …
  <dmdSec ID="mods">
    <mdWrap MDTYPE="OTHER">
      <xmlData>
        <mods xmlns="http://www.loc.gov/mods/">
          <titleInfo>
            <title>Neither gods nor emperors :</title>
            <subTitle>students and the struggle for
                      democracy in China</subTitle>
          </titleInfo>
          <name type="personal">
            <namePart>Calhoun, Craig J</namePart>
          </name>
          <originInfo>
            <publisher>University of California Press</publis
            <dateIssued>c1994</dateIssued>
            <dateIssued encoding="marc">1994</dateIssued>
            <issuance>monographic</issuance>
          </originInfo>
          <subject authority="lcsh">
            <topic>Students</topic>
            <geographic>China</geographic>
```

preFilter.xsl

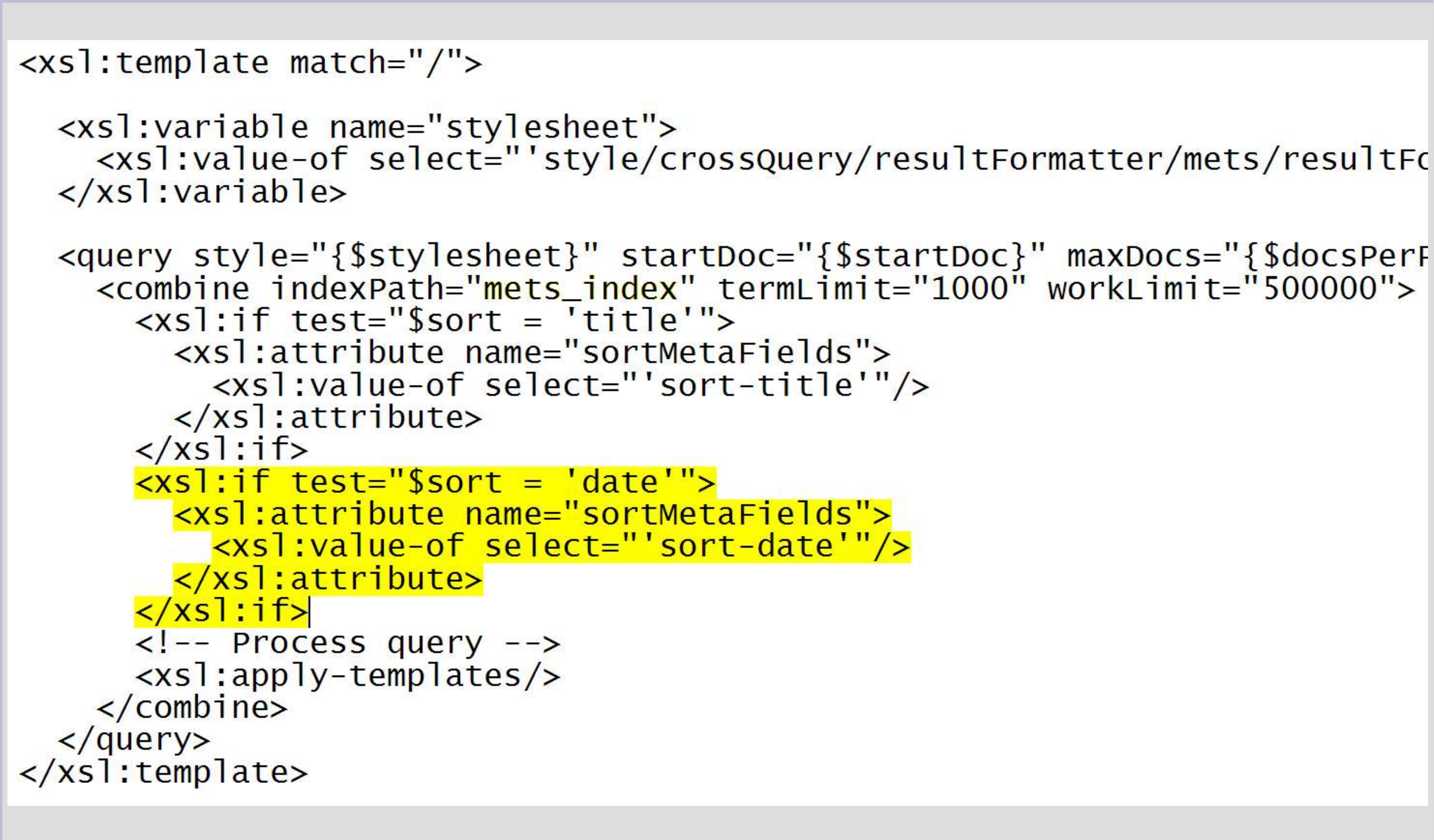
```
<!-- Process MODS -->
<xsl:template match="mods">
  <xsl:variable name="title" select="titleInfo/title[1]"/>
  <title xtf:meta="true">
    <xsl:value-of select="$title"/>
  </title>

  <!-- Process the rest of the MODS -->
  <xsl:apply-templates/>
</xsl:template>

<!-- Process date -->
<xsl:template match="dateIssued[@encoding='marc'][1]">

  <date xtf:meta="true">
    <xsl:value-of select="string()"/>
  </date>

  <sort-date>
    <xsl:attribute name="xtf:meta" select="'true'"/>
    <xsl:attribute name="xtf:tokenize" select="'no'"/>
    <xsl:value-of select="string()"/>
  </sort-date>

</xsl:template>
```

Indexing



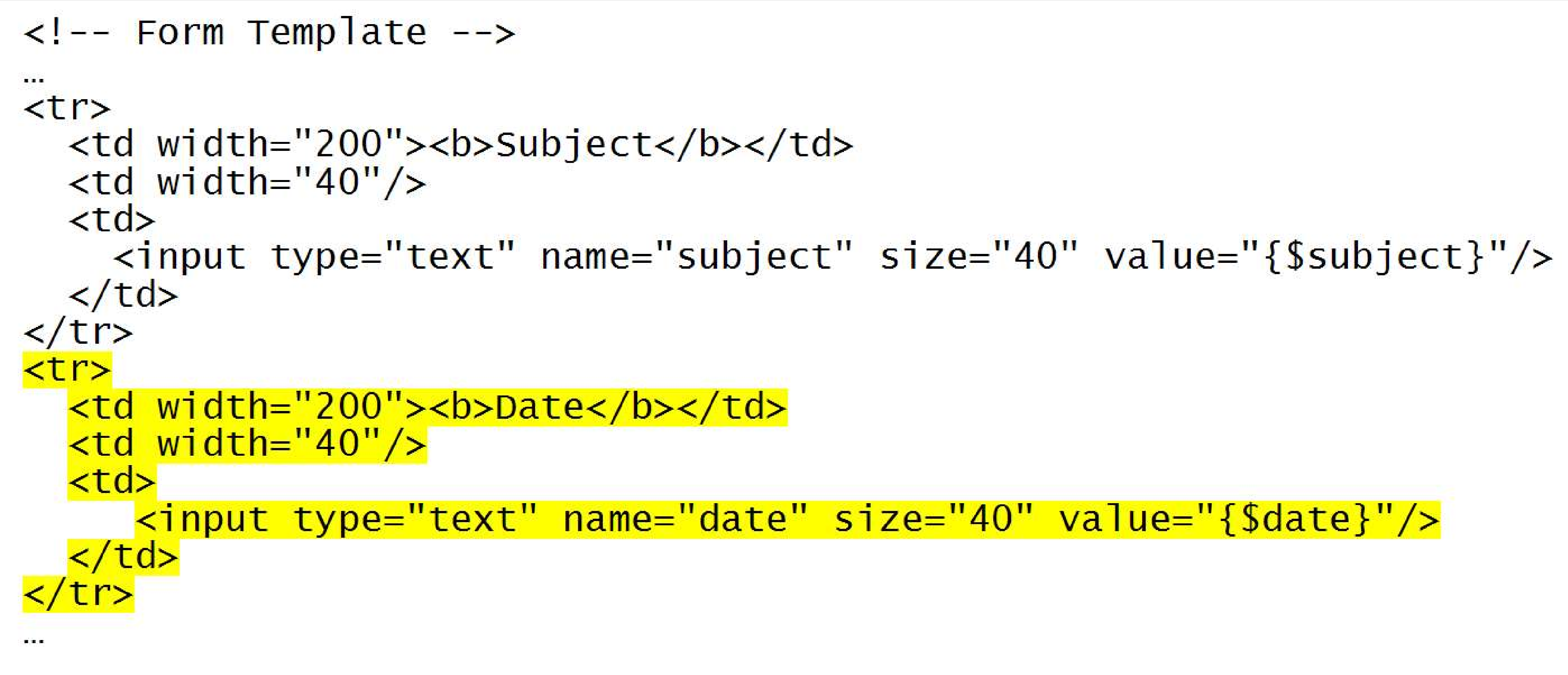
```
$ textIndexer -index mets

TextIndexer v1.0

  Indexing New/Updated Documents:
    Index: "mets"
      Scanning Data Directories....... Done.
      (0%) Indexing [chi_0_116_00000009.xml] ... (1 keys) Done.
      (0%) Indexing [chi_0_116_00000010.xml] ... (1 keys) Done.
      (1%) Indexing [chi_0_116_00000011.xml] ... (1 keys) Done.
      …
      (99%) Indexing [chi_0_116_00001445.xml] ... (1 keys) Done.
      (100%) Indexing [chi_0_116_00001446.xml] ... (1 keys) Done.
      Removing Missing Documents From Index:
        No Missing Documents to Remove.
      Done.
    Done.
  Done.

  Optimizing Indexes:
  Done.

Indexing complete.

$ |
```

queryParser.xsl



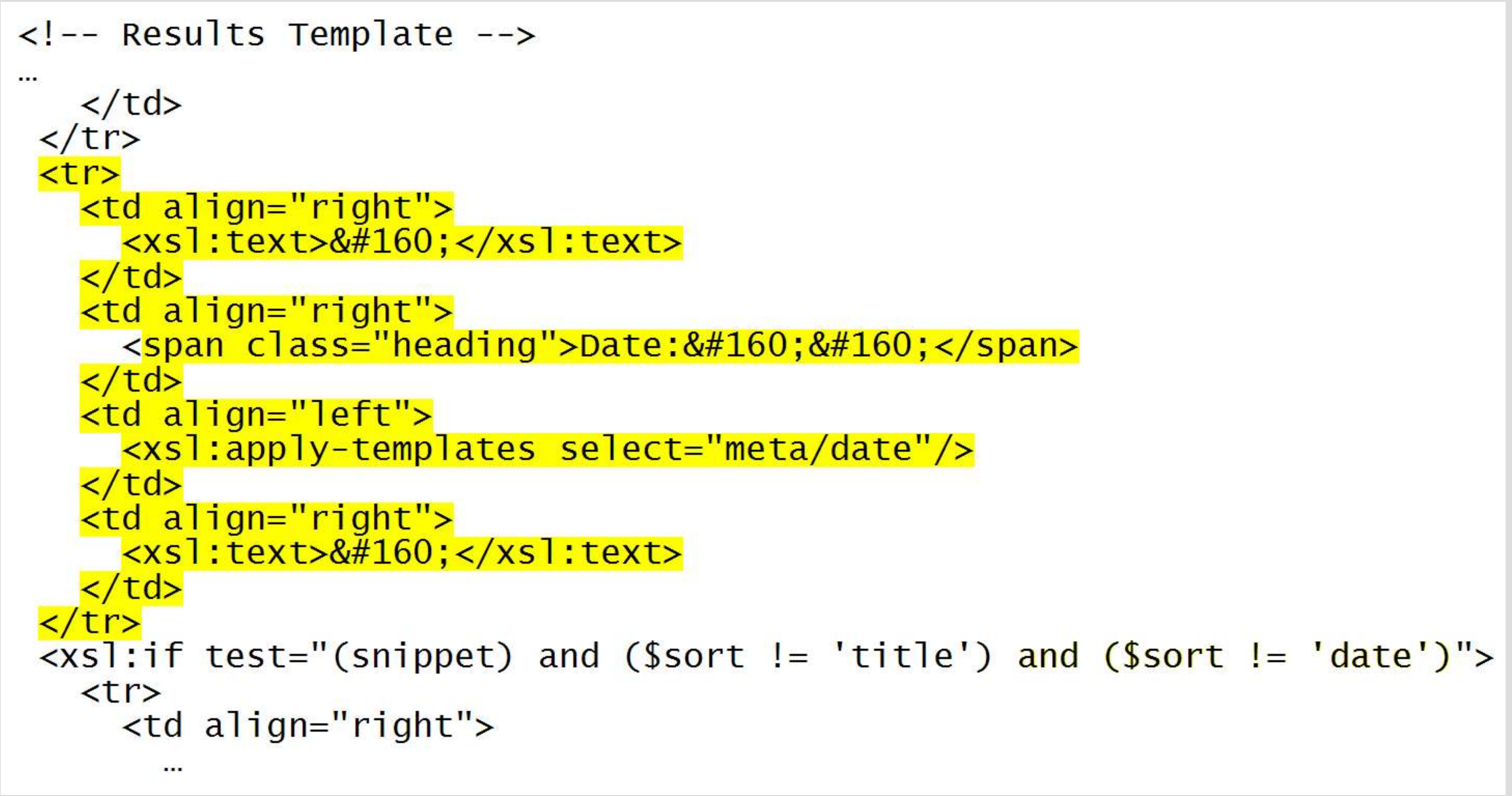
```
<xsl:template match="/">

  <xsl:variable name="stylesheet">
    <xsl:value-of select="'style/crossQuery/resultFormatter/mets/resultFo
  </xsl:variable>

  <query style="{$stylesheet}" startDoc="{$startDoc}" maxDocs="{$docsPerP
    <combine indexPath="mets_index" termLimit="1000" workLimit="500000">
      <xsl:if test="$sort = 'title'">
        <xsl:attribute name="sortMetaFields">
          <xsl:value-of select="'sort-title'"/>
        </xsl:attribute>
      </xsl:if>
      <xsl:if test="$sort = 'date'">
        <xsl:attribute name="sortMetaFields">
          <xsl:value-of select="'sort-date'"/>
        </xsl:attribute>
      </xsl:if>
      <!-- Process query -->
      <xsl:apply-templates/>
    </combine>
  </query>
</xsl:template>
```

resultFormatter.xsl



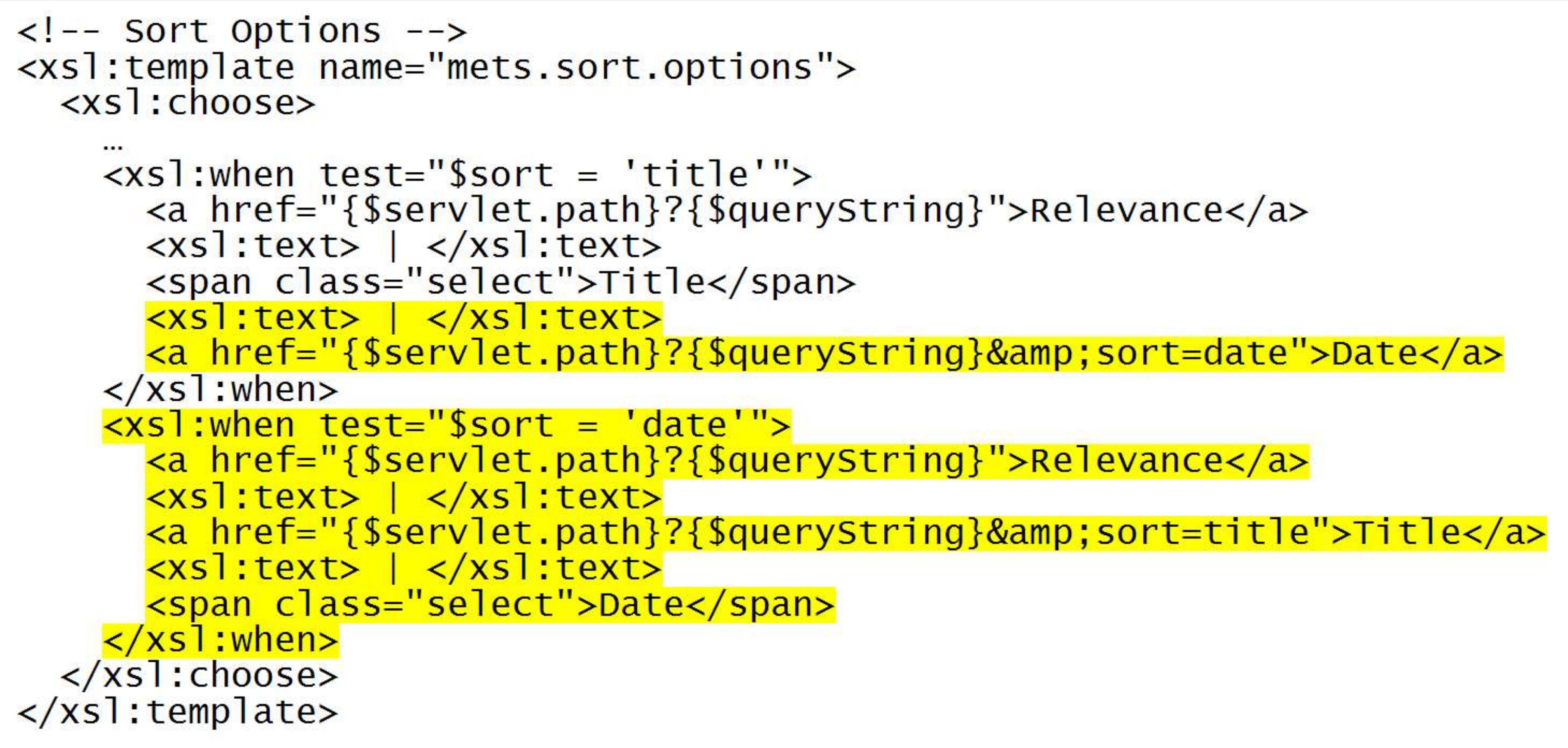
```
<!-- Form Template -->
…
<tr>
  <td width="200"><b>Subject</b></td>
  <td width="40"/>
  <td>
    <input type="text" name="subject" size="40" value="{$subject}"/>
  </td>
</tr>
<tr>
  <td width="200"><b>Date</b></td>
  <td width="40"/>
  <td>
     <input type="text" name="date" size="40" value="{$date}"/>
  </td>
</tr>
…
```

resultFormatter.xsl



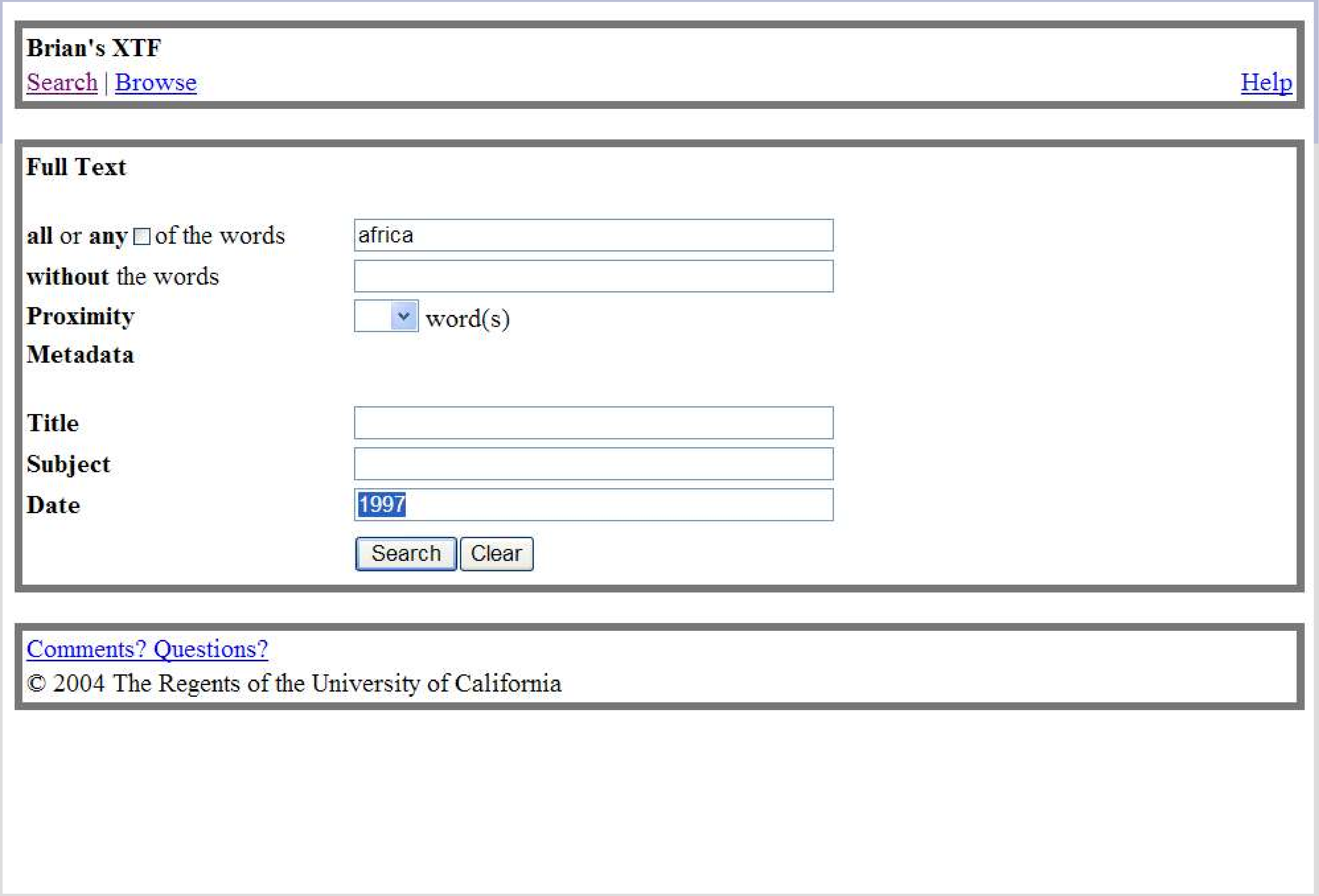
```
<!-- Results Template -->
…
   </td>
 </tr>
 <tr>
   <td align="right">
     <xsl:text>&#160;</xsl:text>
   </td>
   <td align="right">
     <span class="heading">Date:&#160;&#160;</span>
   </td>
   <td align="left">
     <xsl:apply-templates select="meta/date"/>
   </td>
   <td align="right">
     <xsl:text>&#160;</xsl:text>
   </td>
 </tr>
 <xsl:if test="(snippet) and ($sort != 'title') and ($sort != 'date')">
   <tr>
     <td align="right">
       …
```

resultFormatter.xsl



```
<!-- Sort Options -->
<xsl:template name="mets.sort.options">
  <xsl:choose>
    …
    <xsl:when test="$sort = 'title'">
      <a href="{$servlet.path}?{$queryString}">Relevance</a>
      <xsl:text> | </xsl:text>
      <span class="select">Title</span>
      <xsl:text> | </xsl:text>
      <a href="{$servlet.path}?{$queryString}&amp;sort=date">Date</a>
    </xsl:when>
    <xsl:when test="$sort = 'date'">
      <a href="{$servlet.path}?{$queryString}">Relevance</a>
      <xsl:text> | </xsl:text>
      <a href="{$servlet.path}?{$queryString}&amp;sort=title">Title</a>
      <xsl:text> | </xsl:text>
      <span class="select">Date</span>
    </xsl:when>
  </xsl:choose>
</xsl:template>
```

Search Form

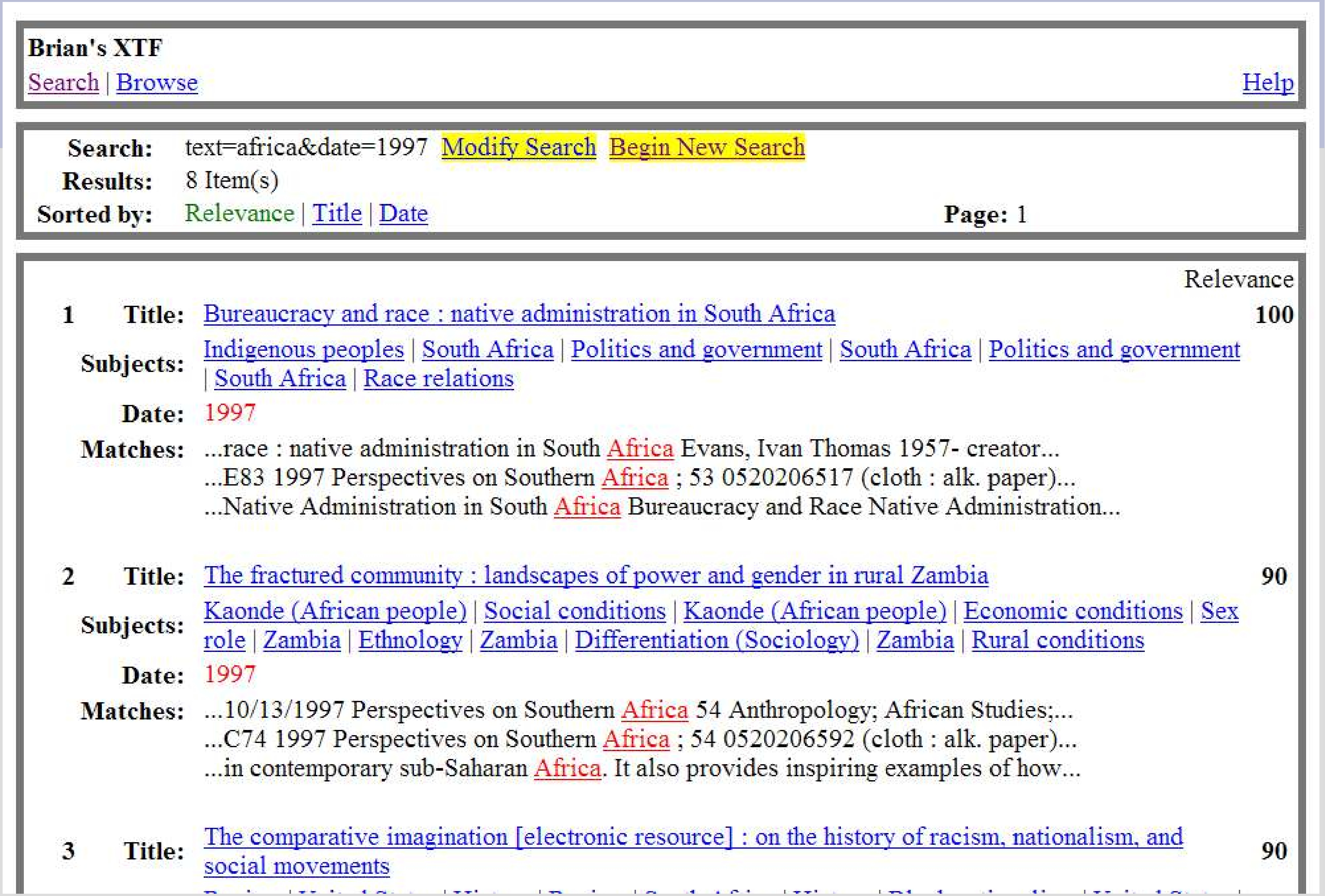


Brian's XTF

Full Text

all or **any** □ of the words	africa
without the words	
Proximity	word(s)
Metadata	
Title	
Subject	
Date	1997

Search Results



Brian's XTF

Search | Browse

Help

Search: text=africa&date=1997 Modify Search Begin New Search

Results: 8 Item(s)

Sorted by: Relevance | Title | Date **Page:** 1

Relevance

1 **Title:** Bureaucracy and race : native administration in South Africa **100**

Subjects: Indigenous peoples | South Africa | Politics and government | South Africa | Politics and government | South Africa | Race relations

Date: 1997

Matches: ...race : native administration in South Africa Evans, Ivan Thomas 1957- creator...
...E83 1997 Perspectives on Southern Africa ; 53 0520206517 (cloth : alk. paper)...
...Native Administration in South Africa Bureaucracy and Race Native Administration...

2 **Title:** The fractured community : landscapes of power and gender in rural Zambia **90**

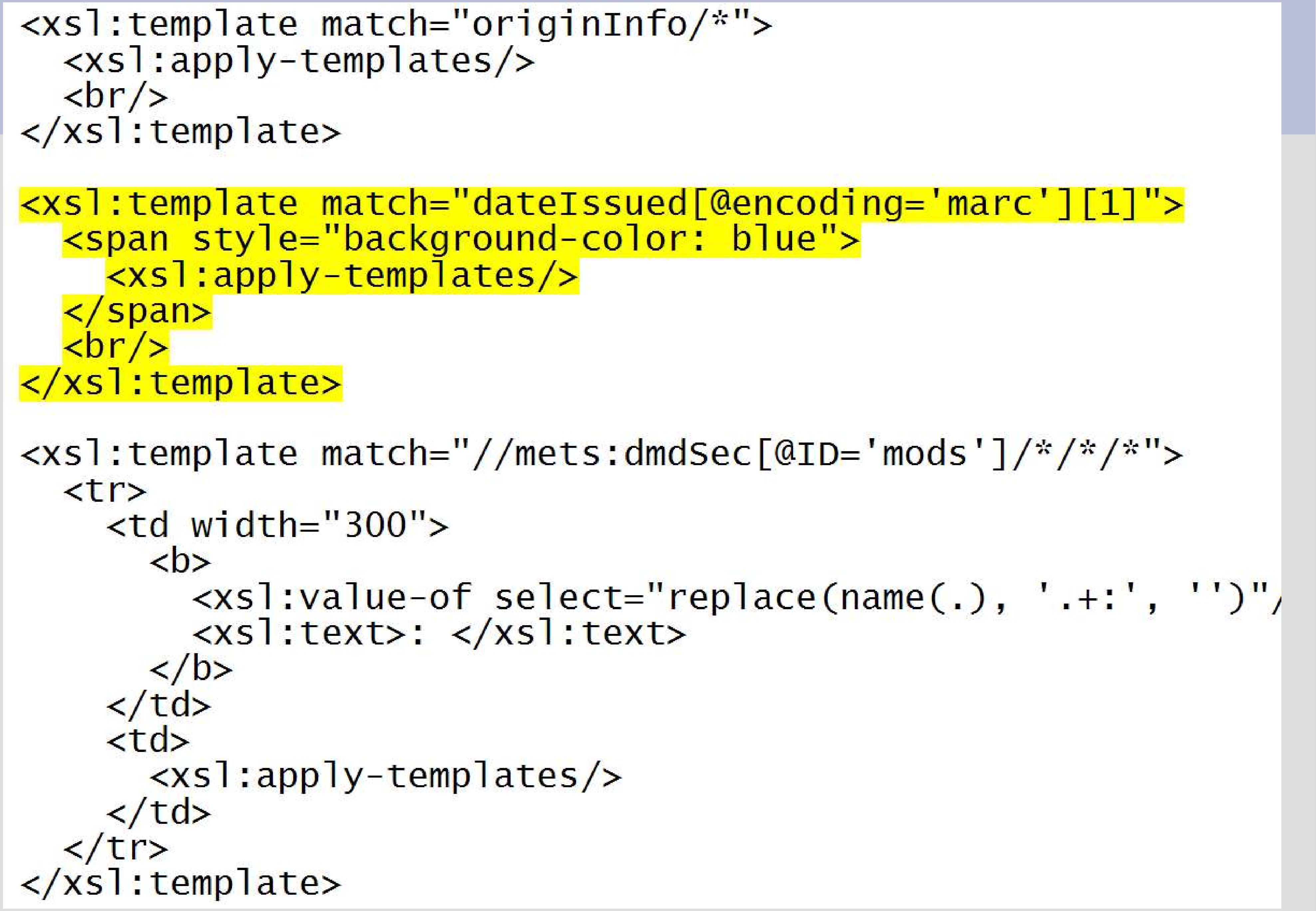
Subjects: Kaonde (African people) | Social conditions | Kaonde (African people) | Economic conditions | Sex role | Zambia | Ethnology | Zambia | Differentiation (Sociology) | Zambia | Rural conditions

Date: 1997

Matches: ...10/13/1997 Perspectives on Southern Africa 54 Anthropology; African Studies;...
...C74 1997 Perspectives on Southern Africa ; 54 0520206592 (cloth : alk. paper)...
...in contemporary sub-Saharan Africa. It also provides inspiring examples of how...

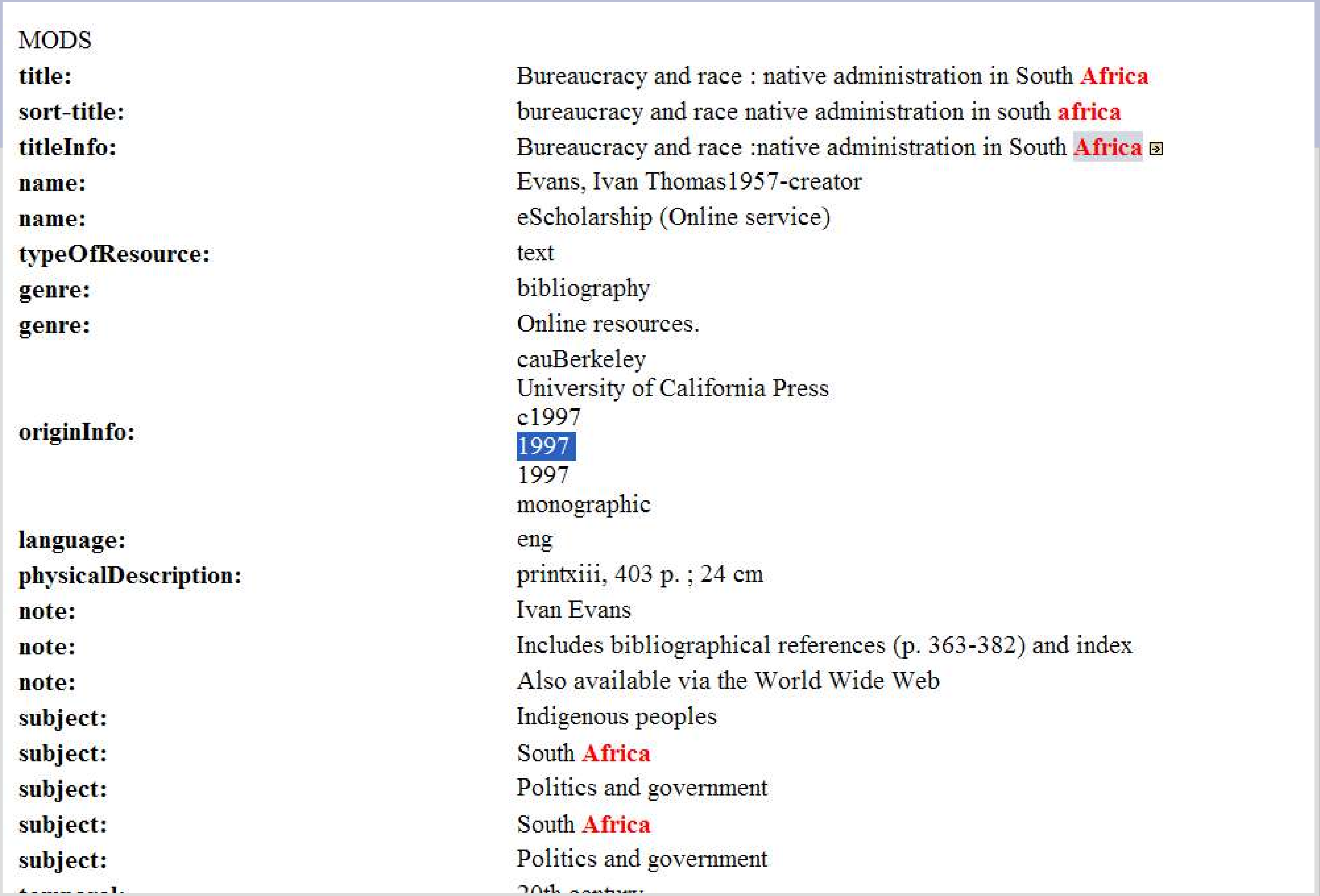
3 **Title:** The comparative imagination [electronic resource] : on the history of racism, nationalism, and social movements **90**

docFormatter.xsl



```
<xsl:template match="originInfo/*">
  <xsl:apply-templates/>
  <br/>
</xsl:template>

<xsl:template match="dateIssued[@encoding='marc'][1]">
  <span style="background-color: blue">
    <xsl:apply-templates/>
  </span>
  <br/>
</xsl:template>

<xsl:template match="//mets:dmdSec[@ID='mods']/*/*/*">
  <tr>
    <td width="300">
      <b>
        <xsl:value-of select="replace(name(.), '.+:', '')"/
        <xsl:text>: </xsl:text>
      </b>
    </td>
    <td>
      <xsl:apply-templates/>
    </td>
  </tr>
</xsl:template>
```

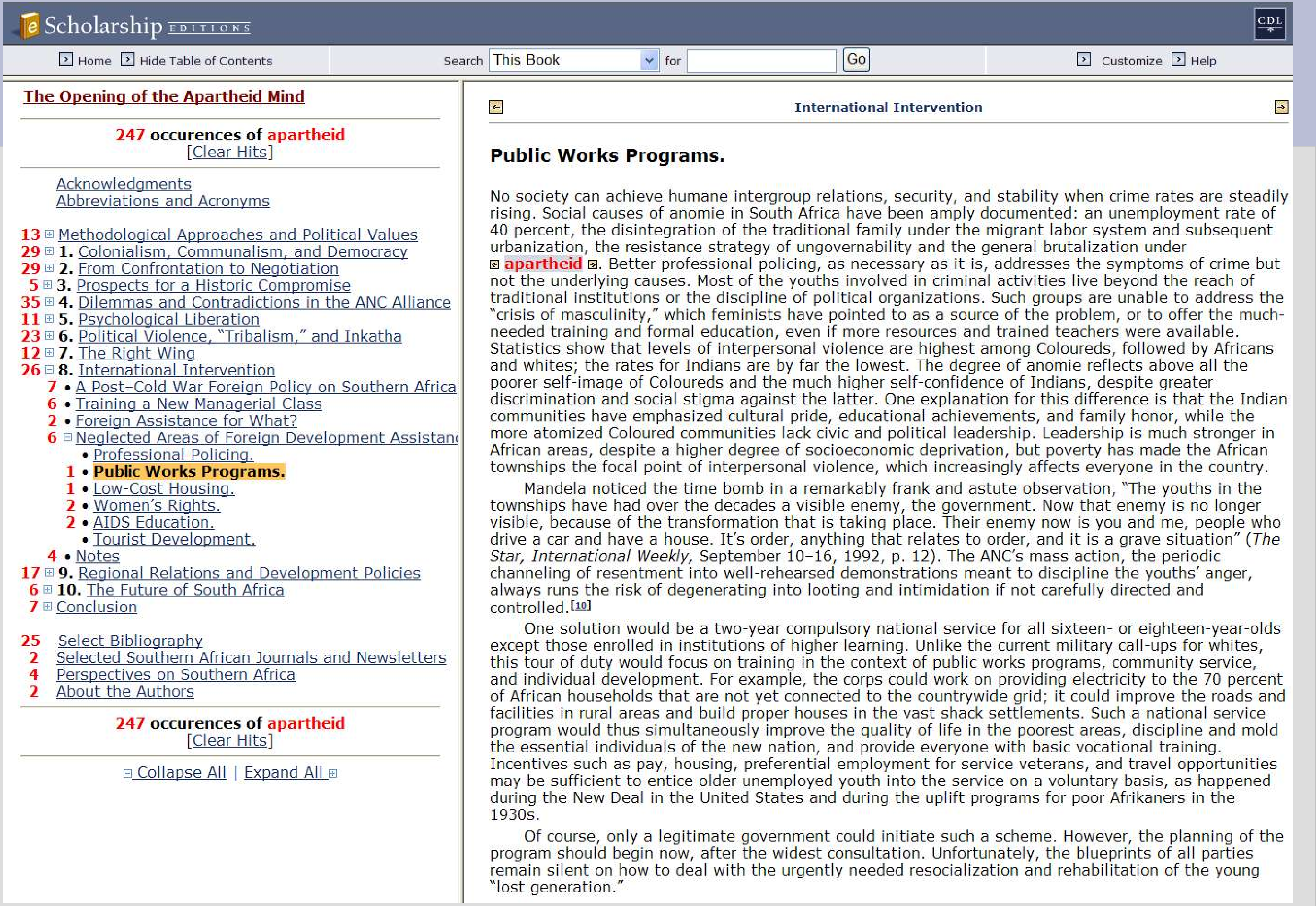
Object View



MODS

title:	Bureaucracy and race : native administration in South **Africa**
sort-title:	bureaucracy and race native administration in south **africa**
titleInfo:	Bureaucracy and race :native administration in South **Africa**
name:	Evans, Ivan Thomas1957-creator
name:	eScholarship (Online service)
typeOfResource:	text
genre:	bibliography
genre:	Online resources.
originInfo:	cauBerkeley University of California Press c1997 1997 1997 monographic
language:	eng
physicalDescription:	printxiii, 403 p. ; 24 cm
note:	Ivan Evans
note:	Includes bibliographical references (p. 363-382) and index
note:	Also available via the World Wide Web
subject:	Indigenous peoples
subject:	South **Africa**
subject:	Politics and government
subject:	South **Africa**
subject:	Politics and government
temporal:	20th century

eScholarship Interface



The Opening of the Apartheid Mind

247 occurences of apartheid
[Clear Hits]

247 occurences of apartheid
[Clear Hits]

⊟ Collapse All | Expand All ⊞

International Intervention

Public Works Programs.

No society can achieve humane intergroup relations, security, and stability when crime rates are steadily rising. Social causes of anomie in South Africa have been amply documented: an unemployment rate of 40 percent, the disintegration of the traditional family under the migrant labor system and subsequent urbanization, the resistance strategy of ungovernability and the general brutalization under apartheid. Better professional policing, as necessary as it is, addresses the symptoms of crime but not the underlying causes. Most of the youths involved in criminal activities live beyond the reach of traditional institutions or the discipline of political organizations. Such groups are unable to address the "crisis of masculinity," which feminists have pointed to as a source of the problem, or to offer the much-needed training and formal education, even if more resources and trained teachers were available. Statistics show that levels of interpersonal violence are highest among Coloureds, followed by Africans and whites; the rates for Indians are by far the lowest. The degree of anomie reflects above all the poorer self-image of Coloureds and the much higher self-confidence of Indians, despite greater discrimination and social stigma against the latter. One explanation for this difference is that the Indian communities have emphasized cultural pride, educational achievements, and family honor, while the more atomized Coloured communities lack civic and political leadership. Leadership is much stronger in African areas, despite a higher degree of socioeconomic deprivation, but poverty has made the African townships the focal point of interpersonal violence, which increasingly affects everyone in the country.

Mandela noticed the time bomb in a remarkably frank and astute observation, "The youths in the townships have had over the decades a visible enemy, the government. Now that enemy is no longer visible, because of the transformation that is taking place. Their enemy now is you and me, people who drive a car and have a house. It's order, anything that relates to order, and it is a grave situation" (*The Star, International Weekly*, September 10–16, 1992, p. 12). The ANC's mass action, the periodic channeling of resentment into well-rehearsed demonstrations meant to discipline the youths' anger, always runs the risk of degenerating into looting and intimidation if not carefully directed and controlled.[10]

One solution would be a two-year compulsory national service for all sixteen- or eighteen-year-olds except those enrolled in institutions of higher learning. Unlike the current military call-ups for whites, this tour of duty would focus on training in the context of public works programs, community service, and individual development. For example, the corps could work on providing electricity to the 70 percent of African households that are not yet connected to the countrywide grid; it could improve the roads and facilities in rural areas and build proper houses in the vast shack settlements. Such a national service program would thus simultaneously improve the quality of life in the poorest areas, discipline and mold the essential individuals of the new nation, and provide everyone with basic vocational training. Incentives such as pay, housing, preferential employment for service veterans, and travel opportunities may be sufficient to entice older unemployed youth into the service on a voluntary basis, as happened during the New Deal in the United States and during the uplift programs for poor Afrikaners in the 1930s.

Of course, only a legitimate government could initiate such a scheme. However, the planning of the program should begin now, after the widest consultation. Unfortunately, the blueprints of all parties remain silent on how to deal with the urgently needed resocialization and rehabilitation of the young "lost generation."

Wrap-up: Possible Future Developments

- Some features on the horizon:
  - Provide automated spelling suggestions
  - Toggle case, diacritic, and plural sensitivity
  - Index non-XML docs (e.g. PDF, Word, HTML)
  - Properly handle non-European languages (Arabic, Chinese, Japanese, Korean, etc.)
  - Store and recall user prefs, searches, results
  - Interface with external thesauri, ontologies, and recommender systems.

Availability

- Pick up a CD after the presentation
- Or go to the source:

  xtf.sourceforge.net

  Documentation, news, source code, feature requests, and more.

Questions?

- We'll take questions now (time permitting), or
- Join our “birds of a feather” session, or
- Catch us at DLF, or
- Send us email:
  - Kirk Hastings: kirk.hastings@ucop.edu
  - Martin Haye: m1@snyder-haye.com