Web2XML-Insight

Structured Book Data Extraction and Query System

Team Members

- A Yogi Athish (MSc Computer Science)
- Amogh Mathad (MSc Computer Science)
- Atharva Bapat (MSc Computer Science)

Project Overview

- Analyzing how different genres of books are described by reviewers online.
- Useful for bookstores, authors, and humanities researchers.
- Goal: Extract insights from book data via web scraping, XML transformation, and querying.

Methodology

1. Web Scraping: Using BeautifulSoup to extract data from books.toscrape.com
2. Export: Convert data to JSON format
3. Transformation: Convert JSON to XML using a custom RelaxNG schema
4. Validation: Use RelaxNG or DTD to validate XML
5. Storage: Store XML data in MongoDB
6. Query: Use XPath/XQuery and visualize using XSLT

Tools & Technologies

- Web Scraping: BeautifulSoup, XPath
- XML Schema: RelaxNG
- Database: MongoDB, PyMongo or SQLAlchemy
- Queries: XPath, XQuery
- Visualization: XSLT, HTML

Expected Outputs

- XML File: Structured book metadata (title, author, genre, price, availability)
- Schema: RelaxNG file for validation
- Database: Populated MongoDB with book data
- Queries: Examples like 'books under €20 in the Fiction category'
- HTML Output: XSLT-styled catalog display