INF 551
Project Guidelines
Theme: A Cloud-based Metadata Extraction & Search App

Develop an application that allows users to search over a collection of (non-text) documents using metadata extracted from the documents.

Requirements:

- Documents: Documents should be in **non-text** formats. Acceptable formats include PDF, Microsoft Office documents (word, excel, ppt, etc.), music files (e.g., mp3), videos, etc. Note that they cannot be plain-text and text-based (e.g., XML or JSON).
- Size of collection: you should have at least **100** documents in your collection.
- Metadata: Your app should include a function that extracts metadata from the chosen documents. Example metadata for a PDF document include: author, title, number of pages, creation/modification date, frequent keywords/phrases in the content of the document, etc. Example metadata for a MP3 file includes its ID3 tags: artist name, song title, year, and genre.
- Metadata extraction: you may develop the extraction function as a utility program that takes a collection of documents as the input and output the metadata in the **JSON** format. You may use existing libraries for your development. For example, Apache PDFBox (https://pdfbox.apache.org/) for working with PDF documents.
- Data upload: you also need to develop a program that automatically uploads the metadata to a cloud database.
- Cloud database: needs to be a NoSQL database. Examples: Firebase, MongoDB, DynanoDB, etc.
- Faceted search: your search app should present an interface that allows users to search documents by facets. Each facet may correspond to an attribute of your metadata. For example, you may allow users to search a collection of mp3 files via facets like artist, year, and genre. For each facet, a list of top/frequent values in the facet should be displayed, along with the number of documents that takes the value. See example screenshot below.
- Keyword search: Your app should also present a keyword search interface that allows users to search over the values of metadata. For example, searching "Maria Kelly" will return all her songs in the collection. The search should be case-**insensitive**.
- Index: you should use an **index** to speed up the search process.
- Search result: for each document in the search result, you should list all its metadata and provide a link to download the document (URL is fine; you may store the documents in a cloud storage service, e.g., Firebase cloud storage, Amazon S3, etc.).
- Use case: Your app should be interesting and useful.
- User interface: it can be either web browser-based or mobile app.

Phases:

The project consists of 3 phases: proposal, midterm report, final report & demo. The total point of the project is 100, broken down as follows.

- Proposal: 10 points
- Midterm report: 10 points
- Final report: 20 points
- Demo: 10 points
- Project implementation: 50 points

Proposal (1-2 pages):

Your proposal should include the following content. Please also prepare a few slides for a short presentation (2 minutes) of your project idea.

- A description of your project idea.
    - What does your app do?
    - Where do you plan to obtain the documents?
    - What metadata do you plan to extract?
    - Where (i.e., which cloud database) do you plan to store the metadata?
    - How do you plan to implement programs for extraction and uploading of extracted metadata to the cloud database?
    - What kind of user interface do you plan to implement (web browser, Android, iOS)?
    - Which programming languages and software libraries will you use?
- Group formation: who are in your group? What is each person's responsibility? Is your group equipped to implement the application by the end of the semester?
- Milestones: a project timeline with milestones.

Midterm progress report (1-2 pages):

- Provide a checklist showing the items in your timeline and the status on each time (complete, on-going, etc.).
- Provide a screenshot of the components (e.g., working of extraction programs) you have completed.
- Are you on track to achieve your milestones?
- Any challenges you have encountered? Any helps that you will need?
- Any other things you think should be reported in the midterm?

Final report (5-10 pages):

It should be a comprehensive report. You may include the contents from your proposal and midterm report, with changes to reflect the final implementation of your project. The final report should have the following parts.

- Project idea.
- Screenshot for each working component with a description.

- Description of documents and metadata extracted. Show examples.
- Implementation details.
- Responsibility and work of each group member.

Final Demo:

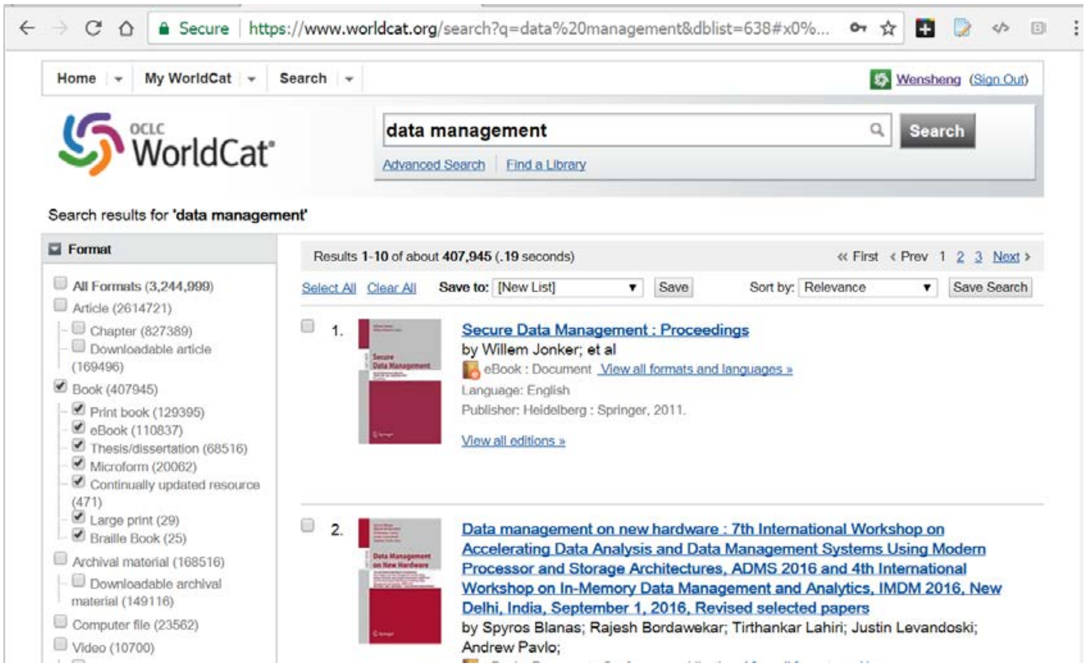
- Demo of your app (10-minute) will be in the last week.
- Show the working of each component: extraction, loading, and search. Show examples of documents and extracted metadata.
- All group members should be present at the presentations.

Deliverables:

Your phase & final reports and project codes.

Example of faceted & keyword search:

Please visit worldcat.org for an example of (faceted & keyword) search application.



Secure | https://www.worldcat.org/search?q=data%20management&dblist=638#x0%...

data management

Search results for 'data management'

Format

☐ All Formats (3,244,999)
☐ Article (2614721)
☐ Chapter (827389)
☐ Downloadable article (169496)
☑ Book (407945)
☑ Print book (129395)
☑ eBook (110837)
☑ Thesis/dissertation (68516)
☑ Microform (20062)
☑ Continually updated resource (471)
☑ Large print (29)
☑ Braille Book (25)
☐ Archival material (168516)
☐ Downloadable archival material (149116)
☐ Computer file (23562)
☐ Video (10700)

Results **1-10** of about **407,945** (.**19** seconds)

Save to: [New List] ▼

Sort by: Relevance ▼

☐ 1. Secure Data Management : Proceedings
by Willem Jonker; et al
eBook : Document View all formats and languages »
Language: English
Publisher: Heidelberg : Springer, 2011.
View all editions »

☐ 2. Data management on new hardware : 7th International Workshop on Accelerating Data Analysis and Data Management Systems Using Modern Processor and Storage Architectures, ADMS 2016 and 4th International Workshop on In-Memory Data Management and Analytics, IMDM 2016, New Delhi, India, September 1, 2016, Revised selected papers
by Spyros Blanas; Rajesh Bordawekar; Tirthankar Lahiri; Justin Levandoski; Andrew Pavlo;