Final Project Introductions

Topic

2. Semantic AI for Enhanced Metadata Creation and Curation: A Comparative Study of Metadata Practices in Digital Archives

By following the steps below, you will have collected, structured, and archived metadata from digital repositories, stored the data in a publicly accessible repository, analyzed metadata trends, and created visual representations of AI-driven metadata practices. This structured approach ensures the project is well-documented and can be easily shared or expanded.

Instructions

1. Understanding Semantic AI and Metadata Curation

- Research how AI is used to enhance metadata creation and curation in digital archives, including:
    - AI-driven metadata generation
    - Natural Language Processing (NLP) for metadata tagging
    - Knowledge graphs and linked data in digital curation
- Identify the role of AI in improving metadata accuracy, discoverability, and interoperability in digital repositories.
- Explore real-world archives and institutions that use AI for metadata enhancement, such as:
    - Europeana
    - Internet Archive
    - Library of Congress Digital Collections
    - Open Library
    - Google Books

---

2. Data Collection and Archiving

Selecting Digital Archives

- Choose at least three digital archives that provide metadata for digital content. Recommended sources include:
  - Europeana (https://www.europeana.eu/en)
  - Internet Archive (https://archive.org/)
  - Open Library (https://openlibrary.org/)
  - Library of Congress Digital Collections (https://www.loc.gov/collections/)
  - Google Books (https://books.google.com/)
- Visit the selected archive websites and explore their metadata structures.

Collecting Metadata

- If the archive provides downloadable metadata, download the metadata file in CSV.
- If metadata is not available for direct download, manually collect relevant metadata fields and enter them into a spreadsheet.
- The key metadata fields to collect include:
  - Title
  - Author/Creator
  - Date of Creation
  - Description
  - Keywords or Subject Tags
  - Source (Digital Archive Name)
  - Metadata Standards Used (e.g., Dublin Core, MODS, METS)
  - AI Enhancement Features (e.g., AI-generated tags, NLP categorization, automatic subject classification)

Organizing the Metadata in a Structured Dataset

- Open Excel and create a table with column headers for each metadata field.
- Enter the metadata collected from the digital archives into the table.
- Ensure data consistency by standardizing date formats and ensuring all required fields are filled as accurately as possible.
- Save the dataset in CSV format for easy sharing and archiving.

---

3. Creating a Repository for Storing the Data

Choosing a Repository Platform

- Select a repository to store and share the collected metadata. Recommended platforms include:
  - GitHub (https://github.com/) – For storing CSV files, documentation, and research findings.
  - Google Drive – For collaborative work on spreadsheets and reports.
  - Zenodo (https://zenodo.org/) – For sharing open-access research datasets.
  - Figshare (https://figshare.com/) – For publishing datasets with permanent links.

Uploading Data to the Repository

- Create a new folder in the chosen repository platform.
- Upload the CSV file containing the collected metadata.
- If using GitHub, create a simple README file describing the dataset, including:
    - The sources of the metadata
    - The fields included in the dataset
    - The purpose of the project

---

4. Analysis and Visualization

Identifying Trends in Metadata Structuring

- Sort and filter the metadata to identify common patterns in how different archives structure their metadata.
- Compare how different digital archives use metadata standards such as Dublin Core, MODS, and METS.
- Identify whether AI-enhanced metadata differs from manually curated metadata.

Creating Visualizations

- Use Excel or Google Sheets (or coding if you are familiar) to create graphs and charts based on the metadata analysis.
- Suggested visualizations include:
    - A bar chart showing the most frequently used metadata fields across archives.
    - A timeline showing how AI-driven metadata generation has evolved over time.
    - A pie chart comparing archives that use AI-enhanced metadata vs. those that rely on manual metadata creation.
    - A network graph illustrating the relationships between metadata elements across different archives.

---

5. Final Report and Presentation

Writing the Report

- The Words limit for each person: no less than 2000 words (If you are working as 4 in the group it should be 8000 words)
- Ensure that information that are used from your articles in your project has in-text citation in APA format
- Structure the report using the following sections:
    1. Abstract
    2. Up to 6 key terms

3. Introduction: Overview of AI in digital archiving and the purpose of the project.
   - Project Question(s): Define the question that you want to answer
4. Related Work: Information from your collected articles
5. Methodology: Explanation of how metadata was collected and organized.
   - Data Collection
   - Data Processing
   - Data Analysis Methods
6. Results and Analysis: Analyze your results
7. Discussion: Discussion of metadata trends observed and findings.
8. Conclusion and Future Work
   - Summary of key insights
   - Contribution of your work
   - Strengths and Limitations
   - Recommendations for improving AI-driven metadata curation in future work.
9. References (Use APA style for references): The three articles for each person in the group is requirement (Based on the year(s) assigned to your group); ==however, you can include more articles. If you like to include extra articles, you do not need to write the summary for those extra articles and they could be chosen from other years as well.== The highlighted part is just for extra articles.

Preparing the Presentation

- Create a 10-minute PowerPoint presentation summarizing the research.
- Include at least two visualizations to illustrate findings.
- Upload the final report and presentation slides to the selected repository for submission and review.

Additional Sources for Metadata

- Kaggle (https://www.kaggle.com/) – Search for "digital archive metadata" datasets.
- Zenodo (https://zenodo.org/) – Open-access research repositories.
- Data.gov (https://www.data.gov/) – Government metadata sources.