Knowledge Graphs for Digitized Manuscripts in Cultural Heritage Applications

Jan Ignatowicz 12.12.2024

Agenda

- Introduction (including context and motivation)
- Objectives of the project
- Methodology (Computer Vision, AI, Semantic Web Ontologies)
- Key Use Case: Jagiellonian Digital Library
- Current Progress and Challenges
- Future Work and Next Steps

Context and Motivation

Context

- Digitization of cultural heritage is growing rapidly, but metadata remains limited.
- Many digital libraries, like the Jagiellonian Digital Library, provide access to digitized manuscripts, but the metadata often lacks descriptive details.
- Metadata standards vary between institutions, making it difficult to create unified collections.
- Searchability is limited, as existing metadata is often insufficient to support complex queries.

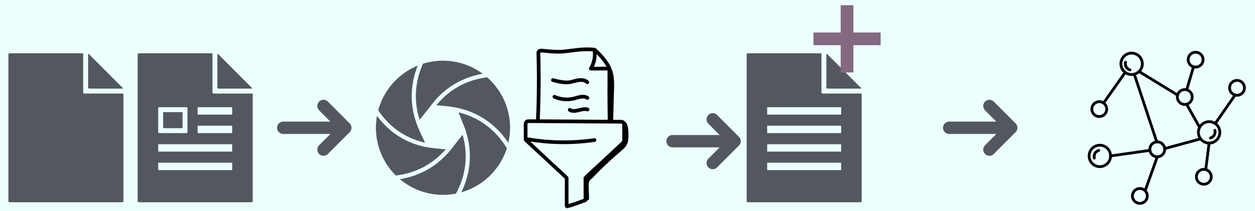
Motivation

- Improve the discoverability of digitized manuscripts.
- Enrich metadata using AI-driven insights.
- Enable cross-collection search by building semantic connections.
- Facilitate novel research methods using enriched metadata and knowledge graphs.

Objectives

- To explore how Knowledge Graphs can enrich metadata for digitized manuscripts.
- To leverage AI and Computer Vision to extract features from historical documents.
- To develop a method that links digital collections using Semantic Web technologies.
- To evaluate the effectiveness of the proposed method on real-world datasets, including the Jagiellonian Digital Library.

Methodology



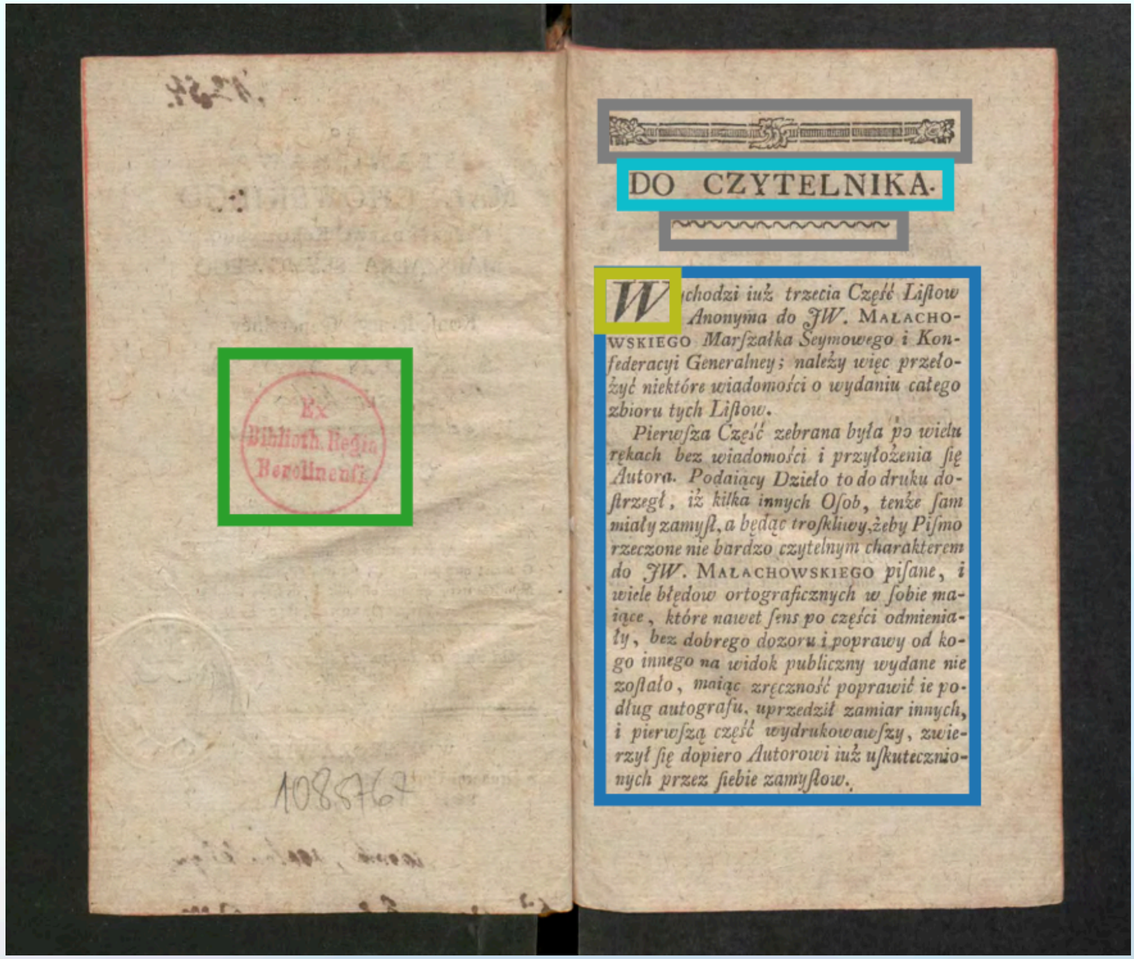
- Data Acquisition: Collecting digitized manuscripts and incunabula from the Jagiellonian Digital Library and other digital archives.
- Feature Extraction (Computer Vision & AI): Using computer vision models (e.g., YOLO, Detectron2, U-Net) to extract visual features from the images (stamps, seals, text regions, ornaments).
- Metadata Enrichment: Incorporating the extracted features into existing metadata to provide richer descriptions for each document.
- Knowledge Graph Construction: Building semantic relationships between enriched metadata using ontologies and linked data principles.
- Evaluation and Refinement: Testing the enriched metadata and knowledge graph on real-world use cases, including the Jagiellonian Digital Library.

Data Acquisition

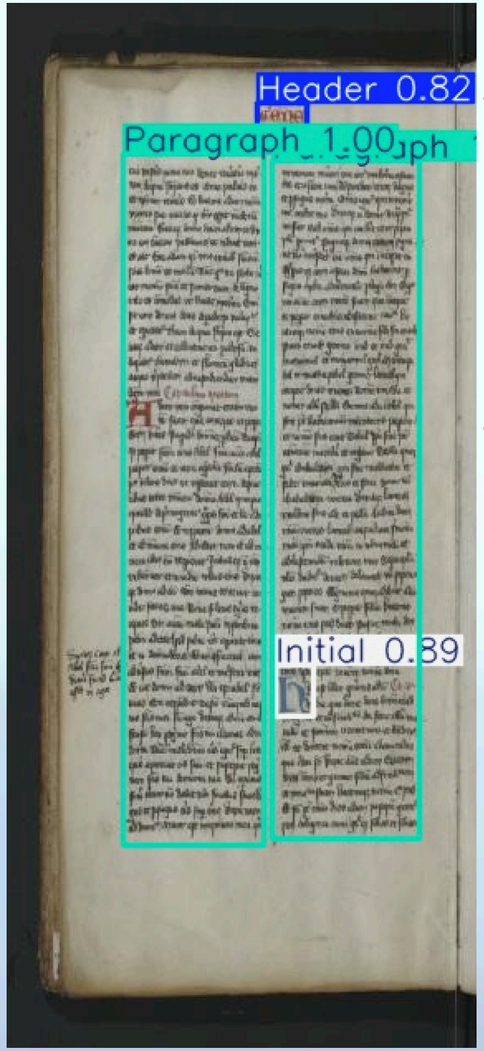
- Main Source:
  - Jagiellonian Digital Library: Primary source for digitized manuscripts and incunabula.
  - Provides access to thousands of historical documents, including unique and rare collections.
- Types of Data:
  - Manuscripts and incunabula images (scanned images).
  - Existing metadata: Title, author, date, and some basic descriptions.
  - Visual elements: Text regions, stamps, seals, signatures, marginalia, illustrations, etc.

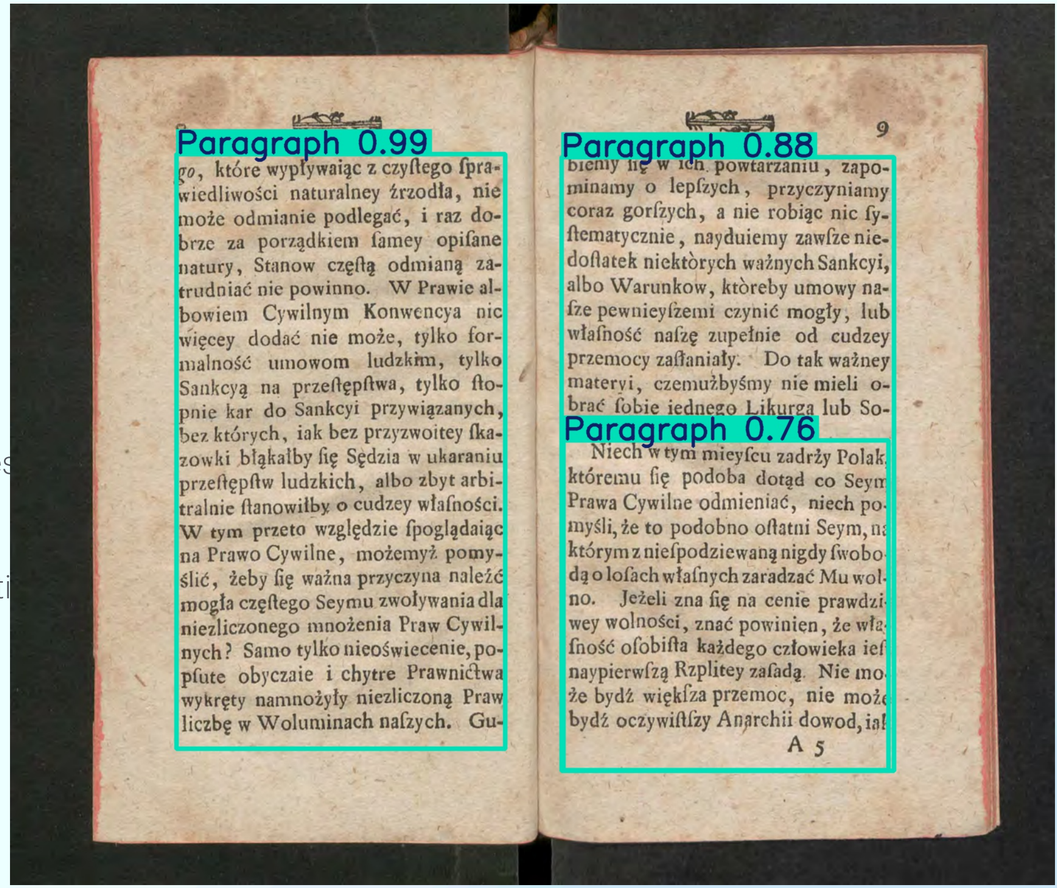
Data Acquisition

- Data Collection Process:
  - OAI-PMH protocol – Used to retrieve datasets in a structured manner.
  - Manual downloads – In cases where APIs are not available, manual extraction is required.
  - Data cleaning and formatting – Images are normalized and annotated for further processing.
- Challenges:
  - Data Heterogeneity: Inconsistent formats, image quality issues, and incomplete metadata.
  - Manual annotations: Requires manual effort to annotate images for training CV models. Creating new datasets of 100 images and retrain retrained models.
  - Limited API support: Not all institutions provide open access APIs, requiring alternative solutions.



Feature Extraction





- Objective of Feature Extraction:
  - Automatically identify stamps, seals, text regions, marginal notes initials, ornaments, and signatures.
  - Provide richer metadata for search, discoverability, and connecti the Knowledge Graph.
- Key Techniques and Tools:
  - Computer Vision Models:
    - YOLO11, Detectron2– Used for object detection
    - DeepLabv3 , U-Net and HRnet – Used for image segmentation to separate key visual features.
  - Manual Annotation:
    - For training, a set of annotated images is required.
    - Annotation process includes marking regions of interest (ROI) for visual elements like text, signatures, and illustrations.

Metadata Enrichment

- Objective of Metadata Enrichment:
  - Link extracted features (stamps, seals, text regions, initials, ornaments, etc.) to existing metadata.
  - Enhance the searchability and discoverability of manuscripts in digital libraries.
  - Facilitate better connections between collections using consistent metadata.
- Key Techniques:
  - Automated metadata augmentation – Detected features are used to automatically populate descriptive metadata fields.
  - Feature-to-metadata mapping – Each detected feature (like a stamp) is linked to a metadata field (like "stamp_location" or "stamp_type").
  - Manual validation – Optionally, enriched metadata can be verified by experts for accuracy.

Metadata Enrichment

- Title: “Example 1”
- Author: "Unknown"
- Description: "Manuscript with Latin text."

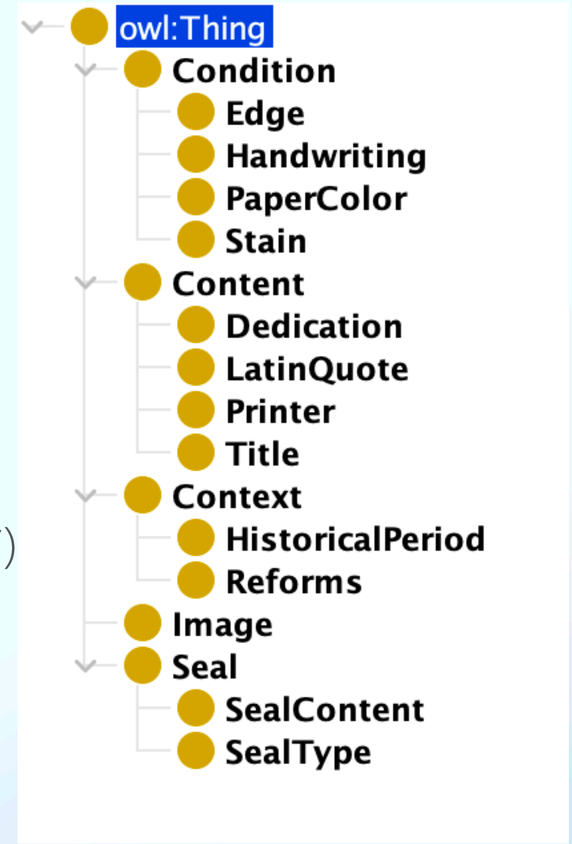
- Title: “Example 1"
- Author: "Unknown"
- Description: "Latin manuscript with handwritten notes, two stamps, and an illuminated initial."

- Detected Stamps: 2 (Top-left, Bottom-right)
- Detected Text Regions: 5 (Paragraphs, Marginal Notes)
- Detected Initials: 1 (Illuminated)

Knowledge Graph Construction

- What is a Knowledge Graph?
  - A graph-based structure where nodes represent objects (manuscripts, features, authors, etc.), and edges represent relationships (e.g., “has stamp”, “was authored by”).
  - Links objects from different collections to discover hidden relationships.
- How the Knowledge Graph is built
  - Nodes: Manuscripts, visual features (stamps, initials), authors, collections, etc.
  - Edges: Represent relationships between nodes, e.g.,
    - "Manuscript A" has "Stamp X"
    - "Author Y" wrote "Manuscript A"
    - "Ornament Z" is found in "Manuscript A"
  - Data sources: Enriched metadata is used to populate the graph.

Knowledge Graph Construction



- Tools and Techniques
  - Ontology design – Defines the types of relationships (e.g., “has stamp”, “has signature”)
  - SPARQL and RDF – Used to create and query the graph.
  - Graph database – RDF triplestore to store the graph.
- Benefits of Knowledge Graphs
  - Improves discoverability – Enables complex queries like "show all manuscripts with similar ornaments."
  - Cross-collection links – Connects collections from multiple libraries.
  - Facilitates exploratory research – Researchers can explore the graph to discover hidden relationships.

Current Progress & Challenges

- Data Collection
  - ✔️ Successfully retrieved manuscripts from the Jagiellonian Digital Library.
  - ✔️ Integrated manual annotation for training purposes.
  - ✔️ Created a custom annotated dataset of 100+ images.
- Feature Extraction
  - ✔️ Fine-tuned YOLO11 and Detectron2 models to detect stamps, text regions, and initials.
  - ✔️ Achieved accuracy of 70-90% for detecting text regions, ornaments, and seals.
- Metadata Enrichment
  - ✔️ Successfully linked extracted features to metadata fields (e.g., number of stamps, stamp location, ornament type).
  - ✔️ Automated enrichment of metadata with visual features.
- Knowledge Graph Construction
  - ✔️ Established initial graph ontology to link manuscripts, authors, and visual features.

- Data-related Challenges
  - ⚠️ Limited API support – Some digital libraries lack API access, requiring manual data collection.
- AI Model Challenges
  - ⚠️ Insufficient annotated data – CV models require larger datasets for higher accuracy.
  - ⚠️ Fine-tuning AI models – Adapting general-purpose models like YOLO11 and Detectron2 to detect historical features (stamps, ornaments) is challenging.
- Knowledge Graph Challenges
  - ⚠️ Ontology design – Defining semantic relationships for new visual features like "has ornament", "has stamp", and "has initial" requires iterative refinement.
  - ⚠️ Data alignment – Linking enriched metadata from different libraries into one graph requires to be done.

Future Work

- Expand the Annotated Dataset
  - Increase the number of annotated images (target: 500+ images).
  - Include more diverse visual elements (seals, ornaments, marginal notes).
- Enhance Feature Detection Models
  - Improve the accuracy of YOLO11 and Detectron2
  - Explore new models for image segmentation, DeepLabv3, U-Net, HRNet
- Refine Metadata Enrichment
  - Add new metadata fields (e.g., "type of ornament", "number of stamps").
  - Test automation of metadata updates with real-time data.
- Expand Knowledge Graph Ontology
  - Add new relationships to the ontology (e.g., "shares ornament style with").
  - Link collections from multiple libraries.

Next Steps

- Scale Up Data Collection
  - Expand beyond the Jagiellonian Digital Library to other digital libraries.
  - Use web scraping to collect data where API access is not available.
- Automate the Annotation Process
  - Develop a semi-automated annotation pipeline to reduce manual effort.
- Pilot Cross-Collection Knowledge Graph
  - Create a pilot version of the Knowledge Graph for a set of 50-100 manuscripts.
  - Demonstrate the cross-collection linking between different collections.
- Research on Industrial Use Case (Digital Twins)
  - Explore the application of Knowledge Graphs for Digital Twins.

Conclusion & Key Takeaways

- Problem:
  - Metadata in digital libraries is often incomplete, inconsistent, and lacks descriptive detail for visual features.
- Solution:
  - My approach enriches metadata using Computer Vision and AI-driven feature extraction, enabling richer, more descriptive metadata.
- Outcome:
  - By linking metadata to Knowledge Graphs, I enable cross-collection connections, improve searchability, and facilitate exploratory research.
- Impact:
  - This work paves the way for a more accessible, connected, and discoverable system for digital cultural heritage.
  - Researchers can now explore hidden relationships between collections.
- Looking Ahead:
  - Ongoing efforts to expand datasets, improve models, and link cross-collection metadata.
  - Potential to apply similar methods in industrial contexts (Digital Twins).