DIGITAL HERITAGE (2025)
S. Campana, D. Ferdani, H. Graf, G. Guidi, Z. Hegarty, S. Pescarin, and F. Remondino (Editors)

From Scanned Pages to Semantic Graphs: Scalable Methods for Extracting Historical and Cultural Knowledge Across Heterogeneous Texts

P. Malak[1], A. Letowska[1] and J. Wodzinski
[1]University of Wroclaw, Poland
: https://orcid.org/0000-0002-8701-1831

Abstract

We present a multilayered methodology for processing digitized historical texts, enabling cross-relational analysis across time periods, languages, and subject domains. Drawing from multiple DH platforms (Tsadikim, Two Enlightenments, Corporeality), we demonstrate an integrated pipeline combining adaptive OCR, noise-tolerant keyword extraction, and NER. Custom preprocessing and fuzzy matching techniques allow for meaningful text recovery from degraded scans in Polish, German, and Yiddish. Data are enriched with spatial and temporal metadata, indexed by topic and linked across projects. The resulting datasets support trend analysis, social network modeling, and discourse mapping. Our approach enables researchers to trace linguistic shifts and intellectual networks over centuries without manual review of source pages. This workflow facilitates interoperable exploration of cultural data and demonstrates how machine learning can assist in recovering semantic relationships from fragmented historical records. The methodology was tested on Enlightenment-era and early 20th-century journals, revealing both technical challenges and insights into evolving ideological, medical, and theological vocabularies.

CCS Concepts

*· **Information systems** → Digital libraries and archives; · **Computing methodologies** → Natural language processing; Machine learning; ·**Applied computing** → Arts and humanities; Digital humanities; ·**Human-centered computing** → Visualization; ·**Theory of computation** → Ontologies;*

1.Introduction

The integration of digital tools in the humanities has revolutionized the way how historical texts are accessed and analyzed. Digitized historical periodicals, particularly those digitized from 19th- and early 20th-century archives, present an invaluable yet complex source of information. Researchers frequently encounter significant obstacles in processing these materials, including poor scan quality, inconsistent layouts, multilingual and archaic language, and the lack of structured metadata.

This article introduces a comprehensive methodology that leverages OCR (optical character recognition), NLP (natural language processing) and ML (machine learning) to extract, index, and analyze content from digitized historical sources.

Using data from the *Tsadikim* [TSAD], *Two Enlightenments* [TENL] and *Corporeality* [CRRL] projects, we demonstrate how combining technical solutions with humanistic inquiry can uncover rich intellectual networks, evolving linguistic patterns, and socio-cultural transformations across a time.

The Jewish culture related multi-hub is an idea of prof. dr hab. Marcin Wodzinski – professor of Jewish history and literature, former Head of Taube's Department of Jewish Studies of University of Wroclaw. In current paper we present tools and methods we deployed in order to implement his concept.

2. OCR and Machine Learning techniques for processing historical texts

Advancements in OCR in combination with machine learning have opened new possibilities for transforming vast archives of historical texts into machine-readable and analysis-ready formats. However, the successful application of these technologies requires careful adaptation to the unique characteristics of distinct archival material. Before outlining the methodological framework, it is

DOI: 10.2312/dh.20253133

essential to examine the specific challenges that digitized historical sources present.

2.1. Challenges in digitized archival materials

Historical periodicals are often preserved in physical formats that have degraded over time, resulting in digitized images that are:

- Blurred, stained, or poorly contrasted,
- Structured in multi-column or irregular layouts,
- Written in old orthographies or multiple languages (e.g., Polish, German, Hebrew, and Yiddish),
- Incomplete or partially damaged.

Figure 1: *Excerpt of the front page of Ewa. Pismo Tygodniowe, no. 1, 19 February 1928.*

Figure 1. presents the visible physical damage and deterioration of the scan and illustrates some of the challenges encountered during text layer extraction and OCR-based analysis of digitized historical periodicals. These features severely impair standard OCR systems. In particular, Yiddish texts, often printed in Hebrew script using non-standard fonts, pose unique recognition difficulties.

Figure 2: *Front page of Russkij Jewriej, no 49, 23.12.1883, St. Petersburg.*

Figure 3: *Front page of Jüdische Frauenwelt, no 13.06.1902, Kraków*

Figures 2 and 3 present different physical issues for appropriate scanning and OCR of old periodicals – mixed font types within one page and blurred, sepia-toned page. All those irregularities are real difficulties for proper text recognition and extraction.

2.2. Adaptive OCR framework

Within mentioned above projects we are continuously developing a robust OCR pipeline using open-source tools such as Tesseract [TSRC], with custom-trained language models for underrepresented languages. The workflow includes:

- Image preprocessing

We employed a flexible, modular approach to preprocessing, iteratively tuning techniques based on the quality and structure of each document. To enhance OCR performance on degraded or historically complex scans, we used a suite of Python libraries:

  - OpenCV [OCV] for grayscale conversion, adaptive binarization (Otsu), noise reduction (median blur), and layout-aware processing,
  - Pillow [PILL] and pdf2image [PD2I] for image extraction and format conversion,
  - NumPy [NUMP] for efficient array manipulation,
  - PyMuPDF (fitz) [PMPD] for page rendering and annotation.

By experimenting with combinations of preprocessing steps and dynamically adjusting parameters such as blur kernel size or thresholding method, we were able to adapt to variations in scan quality, typography, and print artifacts. This adaptability was crucial in achieving high OCR fidelity on underrepresented languages like Yiddish, especially when dealing with faded ink, skewed layouts, or archaic fonts.

- OCR customization**:**

To address the challenges of working with multilingual historical sources lacking searchable text layers, we developed an adaptable OCR pipeline capable of handling underrepresented and typographically diverse languages, including Yiddish, Polish, and German. We leveraged [TSRC], [SMIT07] OCR with appropriate

language models (e.g., yid.traineddata, pol.traineddata, deu.traineddata), depending on the document language. However, standard models alone were insufficient due to the poor print quality, archaic orthographies, and inconsistent fonts in the source material.

We used Tesseract's image_to_data output to extract not only textual content but also detailed spatial metadata (bounding boxes, line information, and confidence scores). This metadata enabled us to locate text instances accurately on the page and map keyword matches directly onto visual representations of the PDFs.

Additionally, we constructed lightweight, domain-specific vocabularies and heuristics for post-processing, tailored to the orthographic characteristics of each language. These were not directly integrated into the OCR engine but were applied during post-recognition normalization and fuzzy matching. This allowed us to recover misspelled or partially recognized keywords, compensating for OCR noise through a combination of lexical filtering and approximate string comparison.

- Text cleaning and normalization

After initial OCR processing, the raw text output often contained significant noise—particularly in low-quality scans or documents with historical fonts and multilingual content. To ensure reliable keyword matching and semantic analysis, we implemented a multi-step cleaning and normalization pipeline.

This process began with the removal of non-target glyphs, such as stray punctuation, misrecognized characters, or fragments resulting from image noise. Using regular expressions and Unicode filtering, we retained only characters belonging to the expected script (e.g., Hebrew for Yiddish, Latin for Polish and German).

Next, we applied diacritic normalization, collapsing variant forms of characters (e.g., decomposed vs. composed Unicode forms) to their canonical representations. This step reduced inconsistencies in the output, especially in languages like Polish, where diacritic use is frequent and prone to OCR distortion.

In addition, we addressed orthographic variation by mapping historical or alternate spellings to standardized forms. For example, older Yiddish spellings using obsolete ligatures or Germanized orthography were harmonized with contemporary transliterations. In Polish and German texts, normalization included adjusting outdated spelling conventions (e.g., "th" → "t") or standardizing compound formations.

Finally, we filtered out isolated symbols, single-character OCR "hallucinations," and extremely short or low-confidence segments. Combined with fuzzy matching and language-specific stopword filtering, this cleanup step significantly improved the precision of downstream keyword detection and visual annotation.

2.3. Fuzzy matching for error-tolerant keyword search

In many cases, the quality of OCR output was too poor to support exact keyword searches—even after preprocessing and text normalization. To overcome this, we implemented an adaptive fuzzy matching mechanism across all language variants encountered in the documents, including Yiddish, Polish, and German.

The approach was built on Python libraries such as fuzzywuzzy [FZWZ] and rapidfuzz [RPDF], which compute string similarity using the Levenshtein distance. We prepared a curated list of approximately 70 keywords, each manually expanded to include relevant morphological and inflected forms specific to each language. This included pluralization, verb conjugations, and orthographic variants, ensuring that a wide semantic net was cast over the noisy OCR data.

During execution, each word detected by OCR was compared against the entire list of keyword variants. We applied a dynamic similarity threshold, typically around 85%, to strike a balance between recall (retrieving as many valid hits as possible) and precision (avoiding false positives caused by OCR noise or short words). Matches falling below this threshold were discarded, and additional filters were introduced to remove:

- Single-character results
- Matches involving only diacritics or partial glyphs
- Redundant detections of the same root across nearby lines

The process was **language-agnostic**, allowing consistent keyword extraction from multilingual corpora using the same logic and libraries. This enabled robust searching even in documents without embedded text layers, supporting a wide range of historical print quality and linguistic variation

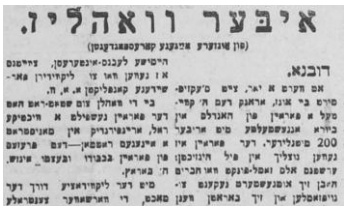

Figure 4: *A piece of text from Ẏoliner lebn, no 343, 13.01.1928, pg. 4*

Figure 4 presents picture of original source, while the following texts are examples of different approaches for OCR of the source periodical.

(a mizre yiyjri kfrnssgdnngsn-ről) Dubna.

*Itt werp egy év" Tzip B'lpy'o• Sidt ny Ogu, Aragk yr□ H. Kzii■ Mel a ®arain ®un kereskedik nidra angcsstmlp? Sif több mint 200 simglirei. paia® iy*i is pnya newlych in ®il Hinzihpnu nrmfgm alm uml-f'«kf uzouhgeim h-.nn nich img/pmert Diypyj «u• guipumlen and goi-open pyi Himyae lenns-inferesen, zeuimnm o W? konfliktus a. A, pp. Ni a választások zum spam-rap hap der ®p aia nunschfilp a vinfike role, ariguirgdi? In monimtrat a iig«nem rapfan-d»p ®a rjn ®in ®p aia gnnudu jaryr i1««. H Badain. MIP RER LICIOAZIN DR*

דובנא.)מיגזערע פון yiyjri קפרנססגדנגסן(

ציפ יאר" א ווערפ אט B'lpy'o • אראגק ,אוגו ניי מ׳ידט yr □ א מעל ■קזיי ה׳

מ׳סגלירע׳י200 אריבער מ׳פ ?אנגצשטמלפ נידרא אין האנדלס ון® אראין®.

קפ»'פ-ואסל אלם ורמפגם חינזיכפון יל® אין נוזליך paia® iy*i pnyai׳ט

ויך נן.-ח וזאוחגיים Diypyj »גאי־אפן זיך און גוויפואסלען •yii אראיערענן® ליקווירירן זו וואו פסערט/אימג ,אינפערעסן-לעננס הימיאש ?] Wa

ראפ-שפאם זום וואהלן די ני .זז ,א .א קאנפליקפן®p aia א נוקשפילפ או זיויימנם
נעמ»אייג א מאניטראט אין ?ארגוירגדי ,ראל ווינפיקנען®a rjn דער חאפ
מאכפ ר/ד דורך ליקייויאזין רער סיפ .באדאין ח1»«. אי פ«ד—ראפפאן
פראל«««» ר>זוארשאוו די< ®ין®p aia גגנודז jary

There are noticeable differences between original material and OCR results, thus shape matching approach is applied in order to detect and indicate searched keywords.

Figure 5: Keyword indication using visual shape matching with threshold of: 70, 80 and 90 respectively
Source: *Folks Gezunt, 15.01.1928, pg. 3*

2.4. Visual annotation and data export

To enhance both the transparency and usability of the results, we implemented a robust system for visual annotation and structured data export. Once keywords were detected via fuzzy matching, each instance was mapped back to its precise page-level coordinates using metadata extracted from Tesseract's image_to_data function. This provided access to word-level bounding boxes (x, y, width, height) along with line and block hierarchy, which were then scaled appropriately to match the resolution and layout of the original PDF (usually from 300 DPI images to 72 DPI PDF coordinates).

Using PyMuPDF (fitz), each identified keyword was then highlighted directly within the original PDF. These highlights allowed users to visually inspect where keywords occurred—even in OCR-degraded texts—and provided immediate contextual verification. For pages with multiple keyword hits, all instances were simultaneously marked, creating an intuitive visual layer over the scanned material.

Figure 6: *Page 3 of Folks Gezunt, no. 2 (86), published on 15 January 1928. Keywords in the text related to health (the article on this page is about tuberculosis) were found using visual annotation and marked clearly for further research*

In parallel, we exported structured metadata about each keyword hit to support further computational analysis:

- Keyword matched (including the actual OCR text matched)
- Target keyword from the list
- Page number
- Line number (if available)
- Coordinates of the bounding box
- Match confidence score

This data was saved in multiple formats depending on the use case:

- A **CSV** or **Excel** file suitable for quantitative analysis or corpus comparison
- A plain-text log *(keywords_locations.txt*) for quick review or debugging
- Optionally, JSON-structured output to integrate with further annotation tools

This multi-channel export strategy ensured **interoperability** with both humanistic and computational research workflows. It also allowed for **cross-validation** between visual annotations and data records, significantly improving the reliability and auditability of keyword detection in low-quality historical documents.

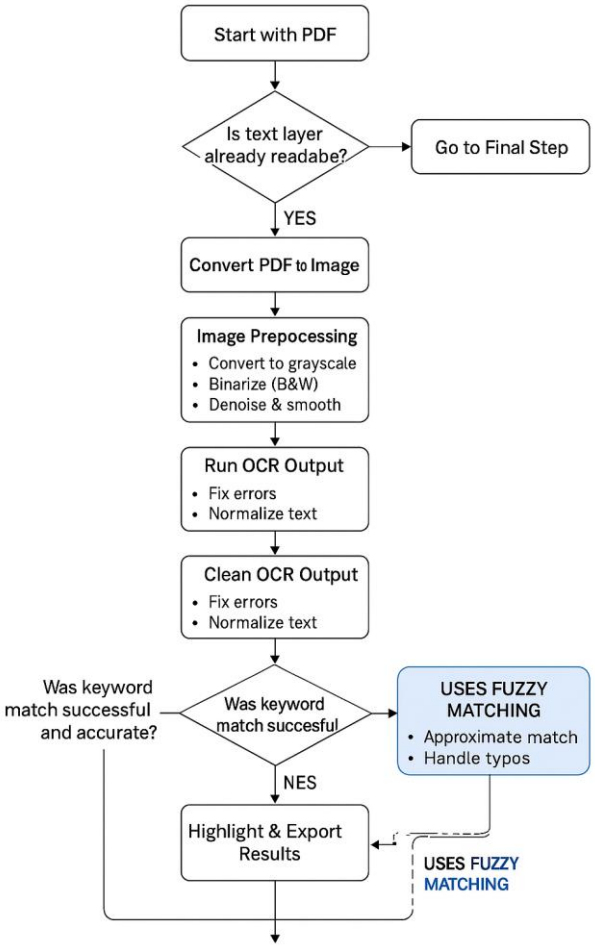

Figure 7: *Adaptative OCR framework flowchart*

3. NLP and Data Mining for cross-temporal and cross-subject analysis

Beyond text extraction, our research focuses on semantic enrichment and structural analysis of the digitized content. Drawing from the Tsadikim and Two Enlightenments platforms, we implemented layered analytical tools that uncover intellectual, linguistic, and socio-political trends.

3.1 Named Entity Recognition (NER) and Authority Record Resolution

To semantically structure the multilingual and historically complex corpus, a multi-step Named Entity Recognition (NER) pipeline was implemented. The goal was to extract and categorize entities such as:

- Personal names (e.g., rabbis, thinkers, historical figures),
- Geographic locations (e.g., towns, shtetls, regions),
- Institutions and publications (e.g., yeshivot, journals, printing houses),
- Culturally significant works referenced implicitly - as in times of our sources origin there was no references standards, thus calling by common (on those times) author name or part of the title.

NER Framework and Tools

We utilized Python-based libraries, particularly **spaCy**, and implemented a **custom supervised training pipeline**. Annotations were performed using **Doccano**, an open-source annotation platform that allowed trained domain specialists to manually tag entities using **XML-style inline tags** (e.g., <PERSON>, <LOC>, <TITLE>). These annotations were essential not only for identifying direct mentions but also for detecting **implicit cultural references** to well-known texts or figures—often cited by shorthand or title alone, without standard bibliographic formatting.

For example, one commonly referenced author from the Two Enlightenments project bibliography is **Salomon Maimon**. His philosophical works, such as *Autobiography*, frequently appear in texts with minimal citation. Annotators were instructed to mark such mentions as follows:

The reflections in <TITLE>Autobiography</TITLE> by <PERS>Salomon Maimon</PERS> contributed to the <TREND>Haskalah</TREND> debates across <LOC>Berlin</LOC> and <LOC>Vilnius</LOC>.

Authority Control and Disambiguation

After entity recognition, a semi-automated resolution pipeline for **authority control** mechanisms was applied in order to standardize and link entities, e.g.:

- Disambiguation and Standardizing of name variants (e.g., “S. Maimon” to “Salomon Maimon”)
- Linking to external authority files and internal project records:
  - **VIAF** (Virtual International Authority File),
  - **YIVO Encyclopedia**,
  - Project-specific curated authority files.
- Using probabilistic matching and thresholding to handle uncertainty in ambiguous or partial references. To manage ambiguity, a semi-automated verification interface was developed:
  - Candidate matches were suggested to human editors based on string similarity and contextual frequency,
  - Editors confirmed or corrected links within a lightweight review dashboard,
  - All actions were logged, enabling iterative refinement and future model training.
- Flagging low-confidence entities for expert review, ensuring human validation in critical cases

This combination of supervised NER, culturally informed annotation, and structured authority resolution enabled precise and context-aware semantic enrichment of the corpus, providing a foundation for further automated analysis and visual exploration.

3.2 Cross-Indexing and Topic Modeling

The indexing strategy was designed to facilitate semantic, thematic, and contextual access to diverse digital humanities (DH) corpora. The material, drawn from digitized historical journals, rabbinic treatises, and Enlightenment-era pamphlets, was annotated and classified using a multi-layered scheme:

- **Multi-Level Tagging:** Each text segment (paragraph or article) was manually or automatically tagged according to primary and secondary themes — such as theology, medicine, mysticism, education, polemics, or Jewish law.
- **Latent Topic Discovery:** To uncover implicit themes and semantic groupings not immediately visible through surface analysis, we employed **Latent Dirichlet Allocation (LDA)**. This unsupervised probabilistic model was applied across the corpus to identify clusters of co-occurring words, enabling us to map latent thematic structures within and across documents. Example: Texts that discuss bodily purity, religious rituals, and illness treatment often clustered into topics associated with *corporeality and religious law*, even if those exact phrases were absent.
- **Cross-Referencing by Metadata:** Index entries were enriched with detailed provenance: document title, source project (e.g., *Two Enlightenments*, *Tsadikim*), page number, paragraph number, date of publication, and all keyword matches. This enabled researchers to trace a concept not only by keyword but through its contextual evolution in time and discourse.

Interoperability: The classification schema was harmonized across platforms (e.g., *tsadikim.uwr.edu.pl* and *twoenlightenments.uwr.edu.pl*), allowing users to correlate a mystical 18th-century rabbinic opinion with a 19th-century Haskalah-era critique, or to explore how health and the human body were perceived in Hasidic versus Enlightenment-oriented writings.

- **Cross-Referencing by Metadata:** Index entries were enriched with detailed provenance: document title, source project (e.g., *Two Enlightenments*, *Tsadikim*), page number, paragraph number, date of publication, and all keyword matches. This enabled researchers to trace a concept not only by keyword but through its contextual evolution in time and discourse.
- **Interoperability:** The classification schema was harmonized across platforms (e.g., *tsadikim.uwr.edu.pl* and *twoenlightenments.uwr.edu.pl*), allowing users to correlate a mystical 18th-century rabbinic opinion with a 19th-century Haskalah-era critique, or to explore how health and the human body were perceived in Hasidic versus Enlightenment-oriented writings.

3.3 Time Series and Trend Detection

To investigate how language and ideology evolved over time within the Jewish intellectual tradition, we implemented temporal analytical techniques over extracted entity and keyword data:

- **Term Frequency Over Time:** We measured and visualized how often key concepts appeared in specific historical intervals. For instance, the frequency of terms such as **"Haskalah"**, **"emancipation"**, **"orthodoxy"**, and **"Zionism"** was calculated per decade or year.
- **Smoothing and Trend Modeling:** Time series were smoothed using moving averages to reveal longer-term trends rather than short-lived anomalie. This allowed clearer interpretation of intellectual shifts across generations.
- **Historical Correlation:** Peaks and troughs in term usage were aligned with historical events:
    - A notable spike in Enlightenment-related vocabulary during the **1860s** aligns with Polish uprisings and the intensification of the emancipation debate.
    - Increased mentions of **Zionism** in the **late 1890s** coincided with early Jewish nationalist discourse in Eastern Europe.
    - Temporal overlays also illustrated shifts from rabbinic legalism to secular ethics or education during interwar years.

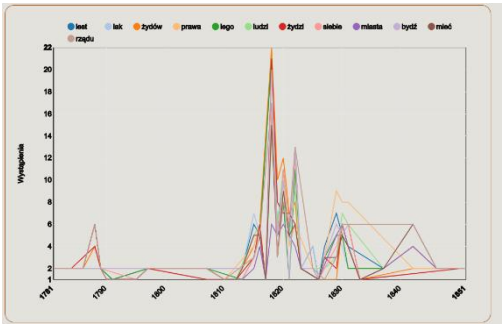

Figure 8: *Example timeseries chart for 12 top frequent words (excluding stop-words) in Two Enlightenments corpora.*

Figure 8 presents an example timeseries chart for 12 top frequent words (excluding stop-words) in Two Enlightenments corpora. The high values of most frequencies around year 1820 are result of corpora structure reflecting publicity of those times.

Interactive Dashboards: The *Two Enlightenments* platform integrates these visualizations with faceted search and filtering (e.g., by publication or author), enabling scholars to contextualize linguistic changes over time.

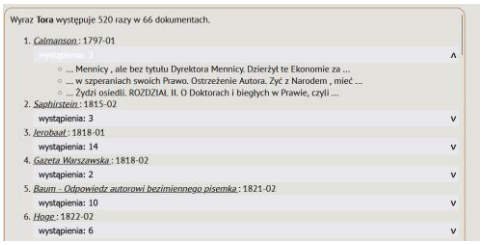

Figure 8: Occurrences of word “Tora” within analyzed corpora

3.4 Social Network Analysis (SNA)

To map intellectual influence and trace dialogic threads across fragmented corpora, we utilized Social Network Analysis based on entity co-occurrence and explicit intertextual relationships. We developed a custom visualisation pipeline entirely in Python, integrated with real-time interactive web visualizations:

- **Graph construction:**
  - **Nodes** represented key individuals (e.g., *Nachman Krochmal*, *Salomon Maimon*), institutions, or publications.
  - **Edges** denoted textual relationships, including citations, polemics, dedications, and co-occurrences within contextually proximate segments (paragraphs, sections).
  - Relationship types and strengths were derived from annotated data and rule-based co-reference detection, distinguishing direct references from indirect cultural allusions.
- **Processing pipeline:**
  - Implemented using standard Python libraries such as:
    - networkx for constructing and analyzing graphs
    - pandas and numpy for preprocessing and statistics
    - custom modules for name disambiguation and temporal slicing of relationship data
  - Edge weighting was based on relationship type and frequency, e.g., a repeated citation of a work strengthened the tie between two nodes.
- **Data Output and Visualization:**
  - Instead of static or Python-native visualizations (matplotlib, plotly, etc.), relationship graphs were serialized into **JSON streams** formatted for web rendering.
  - Graphs were rendered client-side using **D3.js**, enabling:
    - Dynamic, interactive visualizations with zoom, pan, and tooltips for detailed exploration of node and edge properties.
    - Real-time filtering (e.g., by period, region, topic, or document type) to study evolving intellectual networks.
    - Temporal and ideological analysis through adjustable timelines and visual representations of shifts in network density and centrality over time.
- **Network Analysis Features:**
  - **Community Detection:** Unsupervised clustering methods identified interpretive communities (e.g., Galician Maskilim, Lithuanian Mitnagdim).
  - **Centrality Metrics:** Betweenness, closeness, and degree metrics highlighted hubs of influence or discourse bridges.
  - **Evolving Networks:** Time-aware subgraphs allowed tracking of ideological evolution and key influencers across decades, tied to specific events (e.g., partitions of Poland, pogroms, rise of Zionist rhetoric).
  - **Named and Unnamed Relations:** The system accounted for indirect cultural references (e.g., works mentioned by title or trope rather than formal citation), reflecting how knowledge circulated non-bibliographically.
- **Integration with DH Platforms:**
  - Final visualizations were embedded in the *Tsadikim* and *Two Enlightenments* portals, allowing users to explore dense relationship graphs in-browser.
  - This offered an intuitive interface to study how Jewish intellectual thought was shaped across time, text, and territory.

This architecture enabled a lightweight but powerful SNA system that combines backend analytical rigor with frontend exploratory utility.

Figure 8: *example network visualization for dynasty Vizhints (Kosów-Vizhnitz)*

4. Example use case

All described above operations, combined in one, universal old-text research pipeline provide, among others:

- **Multilingual corpora** in Polish, German, and Hebrew, annotated for named entities, rhetorical markers, and ideological references.
- **Topic modeling**, using co-occurrence-based clustering, to detect zones of convergence (e.g., educational reform, natural law) and divergence (e.g., secularism vs. traditionalism).
- **Timeline analytics**, enabling users to explore how specific concepts (e.g., "reason," "nation," "Torah") gained or lost prominence across decades, revealing trends tied to broader socio-political change.
- **Geo-referenced text mapping**, showing how Enlightenment discourses spread through publication centers, correspondence networks, and translational flows between Galicia, Berlin, Vilnius, and beyond.

The main features are:

- browsing of Hasidic data by **dynasty, author, location, and topic**.
- Integrated with metadata such as **date, publication history, lineage**, and cross-references to other Hasidic figures.
- A geo-visual representation of the **Hasidic world**, mapping places of origin, influence, and textual production.
- Clicking a location reveals **associated figures and texts**, including excerpts and cross-links to related records.
- Supports temporal filtering to track the **geographical diffusion of dynasties and teachings** over time.
- Embedded relationship graphs visualize **inter-author influences**, **citation trails**, and **textual reuse**.

Together, these portals form an integrated digital humanities environment that allows scholars to **move seamlessly from close reading to distant viewing**, combining philological accuracy with quantitative, data-driven insights into the Jewish Enlightenment and Hasidic textual spheres.

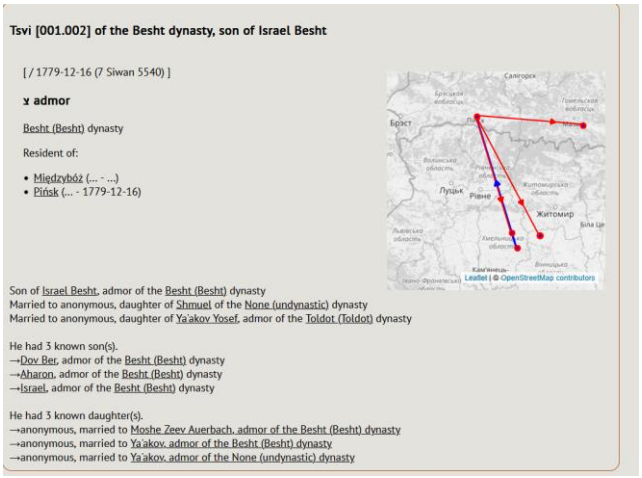

Figure 9: *example details of tsadik Tsvi of the Best dynasty with arrows indicating his relocations and relocations of his sons*

5. Conclusions

Our interdisciplinary approach—blending ML, NLP, and humanities—demonstrates how historical texts can be transformed into structured, analyzable data. The combination of OCR customization, fuzzy matching, and advanced data visualization enables:

- Efficient and scalable access to historical archives
- Rich metadata extraction and authority control
- Discovery of thematic, linguistic, and social patterns
- Interoperable data for use across multiple DH platforms

Future work includes developing deep learning models tailored to historical Yiddish and Polish prints, implementing semantic web standards for cross-institutional data sharing, and creating user-friendly UIs for non-technical researchers.

Our flexible, language-agnostic strategy proved effective across various historical publications, making it possible to extract meaningful textual data and annotate documents visually, even when traditional OCR methods failed.

References

[CRRL] Corporeality. Discurses on body, available online: https://corporeality.uwr.edu.pl/start_en

[FZWZ] fuzzywuzzy · PyPI: https://pypi.org/project/fuzzywuzzy/

[NUMP] NumPy: https://numpy.org/

[OCV] OpenCV - Open Computer Vision Library https://opencv.org/

[PD2I] pdf2image · PyPI: https://pypi.org/project/pdf2image/

[PILL] pillow · PyPI: https://pypi.org/project/pillow/

[PMPD] PyMuPDF 1.26.3 documentation: https://pymupdf.readthedocs.io/

[RPDF] RapidFuzz · PyPI: https://pypi.org/project/RapidFuzz/0.1.0/

[SMIT07] Smith, R. (2007). An Overview of the Tesseract OCR Engine. ICDAR.

[TENL] Dwa Oświecenia. Polacy, Żydzi i ich drogi do nowoczesności, available on-line: https://twoenlightenments.uwr.edu.pl/

[TSAD] Tsadikim - networking Charisma, available on-line: https://tsadikim.uwr.edu.pl/tsadikim

[TSRC] GitHub - tesseract-ocr/tessdoc: Tesseract documentation: https://github.com/tesseract-ocr/tessdoc

Acknowledgements

This article was supported by the National Science Centre, Poland (NCN) under grant no 2022/47/B/HS2/01522.

This article was supported by the National Science Centre, Poland (NCN) under grant *Dwa Oświecenia. Polacy, Żydzi i ich drogi do nowoczesności*, grant no 2019/35/B/HS3/00434.

This article was supported Polish Minister of Science within the National Programme for the Development of the Humanities, Poland (NPRH) under grant *Chasydyzm. Leksykon przywódców religijnych*, grant no 11H 20 0154 88.