HASSO-PLATTNER INSTITUTE – BACHELOR PROJECT WITH JDCRP

2023-2024

In 2023-2024, a group of eight senior bachelor students at the Hasso-Plattner Institute at the University of Potsdam employed newly emerging artificial intelligence technology to develop a prototype for the JDCRP central digital archival platform on Nazi-looted art. They documented their project year to encourage other students, young adults, teachers and others with a general interest to learn more about the Holocaust through the use of AI and provenance research.

The students divided their observations on the project into five subject areas:

1. Project Documentation
2. Methodologies and Technologies
3. AI and Provenance Research
4. AI and Holocaust Education
5. Results and Recommendations

1. PROJECT DOCUMENTATION

Phase 1: Orientation

1. **Familiarisation with the historical context**
   Collection of domain knowledge, e.g. visiting the Brandenburg state archive, watching the film *The Monuments Men.*
2. **Introduction to knowledge graph technology**
   Setting up the Neo4j graph database, learning to work with Cypher, a query language specifically designed for use with 'labelled property graph' databases.

Phase 2: Modelling an ontology

3. **Modelling of entities & relations**
   Challenges included modelling historical uncertainty, contradicting source information, events in time, ambiguous locations—all while retaining the context provided by the source's original material.
4. **Establishing a hierarchy of cultural asset classifications and materials**
   Classifying cultural assets into a known hierarchical structure may help with finding matches across sources in the future.
5. **Implementation of the ontology as RDF**
   Implemented our modelled concepts as an RDF schema, set up a build pipeline to automatically infer SHACL shapes from the schema, allowing for automated validation against the schema.

Phase 3: Infrastructure for data preparation

6. **Scrapers for various data sources**
   In an early phase of the project, not much data was yet available to us. We thus scraped public sources in order to become familiar with the nature of historical data in the art domain.
7. **Iteratively developing ETL scripts for 4 data sources**
   Four ETL (extract transform load) scripts were developed, integrating data from the Linz, MCCP, WCCP, and ERR sources into the unified schema—a process during which also the ontology was further refined.
8. **ETL framework**
   A Python framework for rapid development of ETL scripts was developed, providing a developer-friendly abstraction for integrating new data sources or iterating on the existing ETL scripts (e.g. extracting more structured data).
9. **Parsing unstructured data using LLMs**
   Extracted structured names, pseudonyms, creation dates and "style of" information from the "authors" column of the WCCP data set using OpenAI's GPT4 large language model, employing few-shot training.

Phase 4: Web platform

10. **Requirements engineering**
    Conducted interviews with three provenance researchers, identified needs in everyday work, developed a mock-up prototype.
11. **Python backend**
    Implemented a backend using the FastAPI framework, able to retrieve data from the Neo4j database, reading it into type-annotated classes and serving it to the frontend via a REST API.
12. **Full text search**
    Setup full text search on relevant entities in the graph database using Apache Lucene.
13. **Frontend**
    Developed a frontend based on the mock-ups, already displaying real data that has passed through the entire data pipeline end-to-end, from importing and preparing the source to rendering it the frontend.

Phase 5: Further improvements to the platform

14. **Timeline**
    Extraction of historic event chains from free text using LLMs. Resulting structured data is displayed in a "timeline" view on the platform and allows users to quickly gain an overall understanding of a cultural asset's known provenance.
15. **Advanced search**
    In addition to the existing full text search functionality, an "advanced mode" has been implemented. Especially targeting provenance researchers rather than the general public, it allows for precise querying across all attributes of the integrated schema.
16. **Similar cultural assets**
    A single cultural asset may have been observed at multiple occasions throughout history, each resulting in a separate record in the integrated datasets. Various approaches to matching similar cultural assets were thus explored, including ML-based techniques such as DITTO.
17. **Comparison table across data sources**
    User interviews with our target audience showed a strong objection against automatic deduplication and merging of corresponding entities from different data sources. A "comparison mode" was implemented to enable users to conveniently compare matched entities attribute-by-attribute.
18. **Collection & person page**
    An earlier version of the platform only allowed users to navigate to cultural assets via the search feature. In an effort to facilitate exploration of the integrated data, all cultural assets of a specific artist or collection can be now listed.
19. **Improvements to the result page**
    Search results now show additional details about the records shown, such as the originating dataset and dataset-specific identifiers (e.g., Munich number). The display of results without available images was also improved.

20. **Display of original values**
    In provenance research, traceability of any transformations applied to the data is of utmost importance. At the same time, our many data preparation measures greatly improve readability and searchability of the integrated data. The platform ensures that the original, unprepared values can be traced and reviewed at all times.

Phase 6: Additional data integration

21. **Marburg CCP data and photographs from bildindex.de**
    The Marburg Central Collection Point dataset was integrated into the platform. Additionally, corresponding photographs found on bildindex.de were linked to cultural assets, reducing manual research effort.
22. **Pictures of property cards from Bundesarchiv**
    Scanned imagery of the property cards from the German Federal Archives (Bundesarchiv) were downloaded and linked to the respective Marburg and Wiesbaden CCP datasets.
23. **AI preparation of complex values like author column**
    Since all datasets contain information on looted cultural assets, most of them feature one column in which data about the assets' artist were recorded, containing unstructured information. AI-based SSE (structured entity extraction) techniques were applied, extracting structured information from free-text using LLMs.
24. **Improve data quality of existing sources**
    Various incremental improvements were implemented for the data preparation of the existing four data sources from Phase 3, utilizing the additional domain knowledge gained in the meantime.
25. **Match information about classification and material to the taxonomy**
    All datasets contain information about the classification and material of the cultural assets (e.g., painting, oil on canvas). We developed a rule-based algorithm for mapping the values to the entities of our taxonomy, reacting flexibly to typos using the Levenshtein distance.

Phase 7: Finishing touches

26. **Documentation**
    README files were authored, documenting our code base for further usage.
27. **About the database page**
    An "About the Database" page on the platform was added, providing insights into the development process and listing the sources integrated so far.
28. **Final presentations**
    We presented our results on different occasions, like the Bachelorpodium and our final presentation.

29. **Reflection on the process**
    We reflected on the project itself and our teamwork in internal retrospectives and this blog article.
30. **Codebase handover to JDCRP**
    We will hand over our code base to JDCRP on June 29th during a walk-through session with representatives from all involved subteams.

2. METHODOLOGIES AND TECHNOLOGIES

How were artificial intelligence methodologies and machine learning programs used in creating the prototype?

AI technologies were especially used within the data preparation process (Extract Transform Load, ETL). Since free-text data, particularly the provenance data, can hardly be parsed using algorithms, AI provided a reliable form of extracting data into a fixed and neat schema. AI was also used for the event extraction for the timeline, to transform free-text provenance data into a structured timeline containing a series of events. These events denote what happened to a particular cultural asset, rendering the information more accessible overall.

For writing code, GitHub Copilot was used by many of our project members to accelerate development. GitHub Copilot is an AI assistant that is tightly integrated into the code editor. It provides dynamic code suggestions and answers programming questions.

Which AI models and algorithms were used in the project?

For both the data extraction and event extraction, we settled on the reliable GPT-4 model from OpenAI, which is one of the most reliable Large Language Models for text generation as of today's standard. These were combined with several rule-based techniques in our ETL process.

How would you evaluate the reliability and precision of the AI models you used?

OpenAI's GPT-4 proved to be a reliable solution to tackle our challenge of extracting the data from free-text columns into a defined set of arranged columns. Our project partners at JDCRP were impressed with the results of the data extraction through AI.

What role did design thinking play in creating a methodology for developing the prototype?

We designed the platform to be accessible to both provenance researchers and non-professional users. To achieve this, we talked to users and iterated based on their feedback. We conducted interviews with various provenance researchers to understand their processes and the challenges they face. Through these conversations, we learned how important it is not to make assumptions about the data.

Our goal is to build a platform that's usable for non-professional users as well. At first, this seemed contradictory because we need to show both the unprocessed, raw data and the prepared data points to ensure the best user experience. We resolved this by displaying the raw values only upon request. This feature has been well received by provenance researchers. Without iterating and talking to the ultimate users of our platform, we would have likely overlooked the need for this feature, resulting in a platform that only partially meets the needs of our core user groups. Therefore, incorporating elements of design thinking into the creation of our MVP was essential.

What were the most challenging aspects of the project for you, and how did you address and overcome these challenges?

One major challenge was that our project consists of multiple, nearly independent parts: The frontend, backend, and ETL part. Since we had members involved in only one or at most two of the code bases, communication was an important part in our project. We therefore conducted weekly internal meetings to discuss upcoming challenges and the work that needed to be done across those code bases. The communicational overhead oftentimes felt a little bit disproportional, but coordination in this project was severely important.

3. AI AND PROVENANCE RESEARCH

Was it possible by working on the project to develop new methodologies and concepts to identify works of looted art?

The project is solely based on archival artifacts dealing with looted art and looted cultural assets. While we have developed concepts that can be used to cross-reference these looted cultural assets across different archival sources, we have not found ways to identify works of looted art.

Were there ethical concerns relating to the use of artificial intelligence in dealing with often sensitive personal information in provenance research? If so, which ones and how were the concerns met?

The primary concerns we had about using LLMs and artificial intelligence for the platform were about wrong responses. For example, at the beginning we wanted to unify representations of cultural assets that our AI-models predicted to be the same. In talking to provenance researchers, we learned that this is a dangerous assumption. Therefore, we decided to only link similar representations, combining the usability gains of connecting similar assets with the requirements of provenance researchers.

Moreover, we have to ensure that the LLMs we use do not hallucinate, i.e. generate responses that are not based on the sources we have provided them with. We have incorporated the requirement into the prompts we use for the LLMs.

Did the technical aspects of this project contribute to your skill development? If yes, in what way?

The technical aspects of this data integration platform, as well as collaboratively working on them, significantly contributed to our skill development in software engineering, machine learning engineering and teamwork.

The large number of diverse and faulty data sources presented us with a huge challenge but offered us an excellent opportunity to get to know and apply different machine learning concepts.

We have worked with large language models, text embeddings, clustering algorithms and entity matching algorithms to find similar cultural assets across data sources and to prepare their information. Furthermore, we learned how to integrate our machine learning results with a knowledge graph and to present the data in a frontend that is usable for provenance researchers. This process deepened our understanding of AI-powered natural language processing, entity matching and its practical applications in data integration.

Working with a team of eight students moreover highlighted the importance of clear and consistent communication. We held regular stand-up meetings, maintained detailed documentation, and used collaborative tools. These practices improved our communication skills, ensuring we could articulate ideas clearly and keep the team aligned and informed. We managed our project versions with Git. Managing version control with Git taught us the importance of a structured branching strategy and regular merges. This practice was very important for maintaining clean code. Furthermore, we regularly conducted code reviews and pair programming sessions for maintaining high code quality, but also to learn from one another.

4. AI AND HOLOCAUST EDUCATION

Before we worked on Nazi cultural looting in the course of our bachelor's project, we were not aware of its extent. We did not consider the looting of cultural property as a central element of the Nazis' attempted eradication of Jewish identity, culture and history.

In order to delve deeper into the topic, our team read many articles and books, talked to provenance researchers, visited an archive and even watched a theater play on the subject together.

It was very valuable to learn about current research on looted cultural assets. We realized how much provenance research has already changed through digitalization, but also what potential still remains.
Although some sources have been digitized, they are scattered across different platforms in inconsistent formats. To find a comprehensive solution, we had to consider the data's political, legal, historical, and scientific significance.

The creation of a central, joint repository of knowledge could overcome many of the obstacles currently facing provenance research on Nazi-looted cultural property. This would improve the exchange and clustering of information for research and generally enhance the visibility of the scope of the topic.

Especially today, it is of immense importance that the systematic persecution and extermination of Jewish people in the Holocaust is not forgotten. The theft of Jewish cultural property as part of the attempted eradication of Jewish identity must be especially emphasized.

By making data on Nazi cultural looting more accessible, a central archival data repository promotes research on the subject and aids public awareness of the systematic erasure of Jewish identities. As an educational database, it has the potential to reveal patterns of antisemitism and counter the distortion and trivialization of the Holocaust.

5. RESULTS AND RECOMMENDATIONS

Could students in other academic disciplines benefit from engaging with the JDCRP central digital platform on Nazi-looted cultural property? If so, how?

Teachers, for example, can benefit from this (especially if more educational tools are added). As an educational tool, I can imagine that teachers and students can register and teachers can create a kind of classroom on the platform. This would include teachers being able to decide which entries students can (or cannot) see. In this way, younger students can be introduced to the topic without being confronted with too much cruelty. In this sense, it would probably also make sense if the entries on the platform were already labeled according to their "'appropriateness", even if only visible to teachers. A second tool in this classroom would be a shared comment function, so that everyone in the class can see the comments of everyone else in the class. Comments can be created on detailed pages of cultural assets and persons to document findings or questions.

Based on your experience, what would you recommend are the next steps for JDCRP?

Do you have any specific recommendations for future projects, initiatives or areas of focus that can help engage more young individuals with the topic?

Implement a Map View, so that you can navigate similarly to Google Maps, and display icons for things like artworks and people associated with that location. Similarly, you could display the chronological history of an entity (person, cultural object) like a video on the map (or possibly for multiple entities at once, to visualize the extent of looting and trafficking).

Bring the platform to interactive screens in thematically appropriate museums.

Classroom concept: Teachers can control which aspects of the platform their students can see, and there are commenting features that allow all members of a classroom to share ideas.

What impact did this project have on you?

The project had a remarkable impact on my final year at HPI. To begin with, it prompted me to learn more about art history - both generally, and specifically as it comes to the Nazi era and their art looting. As a team, we watched theater plays and movies, visited several museums, and met with provenance researchers and various art historians. For me, personally, I simply did not know the lengths to which the Nazis used art lootings to demonstrate their power and oppress the Jewish people and other groups.

Beyond that, the project allowed me to dive deep into a variety of technologies that I hadn't had worked with before. The technological challenges we had to overcome were profound and manifold - from dealing with typographical errors in the results of the digital scans (OCRs), to using Large Language Models to parse artwork data into structured tables. It was fun and exciting to apply such modern technology to a domain where most other software applications seemed dated or under-developed.

Lastly, our diverse team and the interaction with JDCRP researchers and developers helped me grow as a team member and sharpen my teamwork skills. Being part of an eight-person team at HPI, supervised by one professor and one PhD candidate, while simultaneously communicating with JDCRP stakeholders and external advisors and partners, meant that we had to communicate our plans and progress a lot. In our team, we all shared the responsibility to write stakeholder emails, prepare meetings, or communicate status updates to HPI and JDCRP.

What did you enjoy the most when it comes to the development of the platform?

I enjoyed the hands-on work of creating a good user experience. There were always challenges specific to provenance research that had to be taken into account, such as displaying the original values, ensuring that the images are always visible on the detail page even when scrolling down, or things that are almost taken for granted, such as being able to zoom in on images.

Of course, it's also fun to have your own creative freedom flow into the development of the platform, such as the choice of background images on the landing page or the many layout decisions you make on a daily basis.

Before the project, I hadn't done much front-end development, but now I feel pretty comfortable with it and I always enjoy working on new issues.

What was the most important lesson you learned working on the project?

Teamwork makes the dream work :)