From source to structure
Extracting knowledge graphs with LLMs

Raphael Schlattmann, Aleksandra Kaye, and Malte Vogl

Introduction

Extracting structured information, such as knowledge graphs (KGs), from unstructured historical texts presents significant challenges for researchers in the History, Philosophy and Sociology of Science (HPSS). While manual annotation is often considered precise, it scales poorly. Recent advancements in Large Language Models (LLMs) offer potential for automating parts of this process, yet require careful integration to handle the often hidden nuances of historical sources and research questions (see also Boulanger, 2026; Meding and Daugs, 2026).

To address this challenge, we developed a computational pipeline[1] aimed at extracting KGs from historical data. It aims to connect large, semi-standardized, but previously isolated sources like biographical lexicons and make them structurally analysable. The goal is to supplement classical historical analysis, with a structural perspective that focuses on relations. Such an approach makes it possible to ask new kinds of questions from these sources and, at the same time, forces a new engagement with their nature (e.g., Weiß et al., 2025). The pipeline employs a two-stage architecture grounded in the principles of Task Decomposition, Human-in-the-Loop (HiL) validation, and a research question & data-driven approach.

Stage 1 implements ontology-agnostic Open Information Extraction (OIE), extracting factual triples (subject-predicate-object) without explicitly defining an ontological structure. The process begins with data preprocessing, loading, and chunking. Already at this stage, we utilize multimodal capabilities of LLMs to extract textual input data for later KG construction directly from initial sources, often characterized by complex layouts. Raw SPO triples (plus time as a property) are then extracted from these text segments using a first LLM. These triples undergo ensemble validation, with a second LLM

1 The code will be made available under https://github.com/raphschlatt/Knowledge_Graph_Extraction

antagonistically scoring, refining, and normalizing them ("LLM-as-a-judge"). Stage 1 includes systematic evaluation against manually created gold-standard samples to ensure extraction quality meets research-question-specific thresholds. This ontology-agnostic approach enables the discovery of unexpected relationships, mitigating premature constraints based on researcher assumptions about the data, while complementing the research-question-driven focus of the second stage.

Stage 2 focuses on ontology-driven knowledge graph construction. The process begins with an LLM analyzing a representative sample of validated triples from Stage 1 to draft potential Competency Questions (CQs) – research questions that the final knowledge graph should be able to answer. These machine-generated, data-oriented questions are refined, discarded, or expanded by a human expert, ensuring alignment with the actual research questions. The resulting expert-validated CQs and the same sample of triples then guide an LLM in drafting a domain-specific ontology designed to address these questions. After manual ontology review in another HiL step, validated OIE triples are mapped onto this ontology, and an LLM drafts SHACL shapes based on the ontology and validated triples, which are again subsequently manually refined. Entity linking and disambiguation then resolves textual entity labels to persistent identifiers from authority files through a multi-source strategy: asynchronous searches against Wikidata with VIAF as fallback, followed by LLM-based contextual disambiguation of entity candidates. Entities without matches undergo character-embedding-based candidate blocking, again followed by LLM-based contextual disambiguation. Successfully matched entities receive authoritative labels, and for locations, geographic coordinates. The pipeline concludes with RDF-star output generation and SHACL validation.

Human-in-the-Loop integration occurs at multiple points: gold-standard creation for OIE evaluation, competency-question refinement, ontology and SHACL shape review and correction, manual disambiguation verification, and quality assessment at each pipeline stage. This tries to balance LLM-driven automation for speed and scalability with expert validation for accuracy and domain appropriateness. The modular design facilitates component replacement as technology evolves, while comprehensive metadata logging at each input/output step (model details, parameters, prompts, timestamps, costs, providers, etc.) ensures partial reproducibility, given the non-deterministic nature of LLM outputs and transparency. This human oversight is crucial both for validating the accuracy of extracted relations and for iteratively refining the ontology based on encountered complexities.

Rather than presenting the intricacies of a finalized tool, this contribution centers on the process of its creation and adaptation. The theoretical framework for the utility of knowledge graphs or networks in general is discussed elsewhere (see e.g., Kaye et al., 2024). Likewise, the specific application of the KG for historical analysis is not the subject of this article. The adopted presentation format here is that of an expert interview, presenting an edited interview transcript (Part 1) featuring the research team. This serves as a research vignette tracing the pipeline's evolution. The interview explores the rationale behind the pipeline's design, discusses concrete problems encountered (e.g., optimizing OIE prompts, designing effective validation steps, balancing automation and control), and details the adaptive strategies employed.

Following this detailed account (Part 1), the paper includes a machine feedback component (Part 2). LLMs were tasked to analyse the introduction and interview transcript and provide critical commentary from two distinct viewpoints: that of a HPSS researcher considering epistemological and methodological implications and that of a Computer Scientist assessing technical choices and implementation.

By combining detailed process documentation via an interview with AI-generated commentary, this work aims to contribute practical insights into the challenges and strategies involved in leveraging LLM capabilities within toolbuilding workflows for HPSS, fostering methodological discussion on the responsible and effective integration of such technologies in humanities research.

The choice of this unconventional, two-part format – an expert interview complemented by LLM-generated commentary – is deliberate. We believe it offers a transparent and reflexive lens through which to examine the often "messy", iterative, and collaborative realities of developing and integrating these new tools within humanities research. Instead of presenting a final result, we aim for a process documentation, revealing the complex decision-making, unforeseen challenges, and adaptive strategies that characterise the actual research process. By contrast, the interview format allows us to capture these nuances, offering a more authentic research vignette and to reflect on our own choices and decisions. Furthermore, juxtaposing the human narrative of the development process with LLM feedback, aims to provoke deeper methodological reflection on the nature of such human-AI collaborations and the epistemological implications of using these new tools in fields like HPSS not only for automatable tasks but also processes of critical interaction.

Part 1: Expert Interview

Notes on format

Why a full-length interview? We present the interview answers in full because Part 2's LLM commentaries are derived from and should be read against this complete transcript. Trimming would firstly compromise this comparative analysis and make some LLM remarks unintelligible. Secondly, as this contribution serves as a research vignette it is meant to narratively capture human experiences in real life settings, giving "the reader a sense of being there in the scene" (Erickson, 1985: 150), in this case the development process.

How was the Interview conducted? The interview questions were collaboratively developed through iterative discussion and redaction. We internally reviewed and refined the questions, then answered them individually in written, asynchronous rounds, iterating until all authors confirmed clarity and completeness. The transcript presented here is slightly edited (light copy-editing for readability, without substantive changes).

Questions and answers

Motivation and challenges

Question: *Let's start at the very beginning. What was the initial problem that you aimed to address with an LLM-based pipeline for knowledge graph extraction, and what did your first conceptual sketch of such a pipeline look like?*

RS: I've been working with automated network data creation for a while now, and in 2023 I participated in a workshop on Heterogeneous Referencing where I had already started using LLMs experimentally to map person names referenced in publications (implicit citations) to explicit citations. That worked "reasonably" well with GPT-3.5 back then, and since then there have been several ideas for using the contextual "understanding" of LLMs to extract otherwise difficult-to-extract semantic information. In my work on the history of science of gravitational research in the GDR, for example, I repeatedly consulted the biographical database *Who was who in the GDR? A lexicon of East German biographies*[2], and each time I had to think about how much information is contained in this lexicon, while being completely disconnected internally (it is a classic lexicon, originally printed, then got digitized but only text-based, so you have individual entries without linking). I always thought there's so much structural information in there, that could be utilized additionally to the way it was used already. So to be able to combine micro and macro perspectives. That was the initial idea: to automatically extract networks from it, before the next step became the idea of using knowledge graphs. The first version was also pretty rudimentary and straightforward: Define a few entity classes with a few relations and then write a prompt. And because this simple experiment already seemed to work so "well", it became a side project and then later eventually a larger project that also took quality control and source criticism in historical research more seriously and tried to mitigate the known problems of LLMs.

AK: For a while now, I've been interested in better understanding the role of migrants in the transnational circulation of scientific information and methods in the nineteenth century. What knowledge and in what formats did migrants contribute towards production and exchange? What underlying conditions were key for facilitating knowledge circulation? I have been researching this in the context of exchanges between various Latin American countries and the territories of the partitioned Polish Lithuanian Commonwealth. It was in relation to this work that I came across a particularly interesting source: Stanisław Zieliński's "Mały słownik pionierów polskich kolonialnych i morskich"

2 "Who was who in the GDR? A lexicon of East German biographies" is part of the biographical databases of the Federal Foundation for the Study of the Communist Dictatorship in Germany. Together with the "Biographical Handbook of German Communists", the database contains around 5,400 structured short biographies of central actors from the GDR and the communist movement. The entries offer fact-based life stories with information on origin, function(s), party affiliation, and professional career (Müller-Enbergs et al., 2010; Weber and Herbst, 2013). Available online at: https://www.bundesstiftung-aufarbeitung.de/de/recherche/kataloge-datenbanken/biographische-datenbanken.

[The Little Dictionary of Polish Colonial and Maritime Pioneers][3], where the author listed over 1070 Polish 'travellers, discoverers, achievers, researchers, explorers, and migrant diarists, activists and writers' (Zieliński, 1933: 1). If the information Zieliński included in his work could be extracted and structured, then digital humanities methods could be applied, allowing us to reconstruct the social networks of Poles working abroad in the nineteenth century, not just in Latin America, but globally. Such an approach could enrich the research I was already conducting in relation to historical Polish migration, which, among other things, involved careful appraisal of "The Little Dictionary" as a historical source. At that point, Raphael had already been applying LLMs in his work for OCR correction with good results, and it seemed like these models could also be useful for retrieving and processing information from "The Little Dictionary".

MV: I had already worked with structured sources previously in my role as a research software engineer, e.g. a list of all ancient Greek settlements in the Mediterranean sea or a three-volumed dictionary of analysed handwritings of medieval manuscripts written all over Europe. For these sources I had built regular expression search methods that gave sufficient results, but required a lot of tweaking and inclusion of edge cases. In the end the actual historical research with these newly organized and more accessible sources was not possible because the time to build the tool to work with the sources was taking so long. The additional aim of extracting the content of a source as a knowledge graph was inspired by another project where colleagues had constructed such a resource after literally years of discussion on the ontology. In the end their knowledge graph was great to describe a specific source but not easily adjustable to other historical questions or source material. My hope with this LLM-based approach was to get results of better quality and shorten the amount of time necessary to verify and clean the results to have more time for the actual historical work with this type of source material.

Question: *When you began implementing these initial ideas, what were the most significant and perhaps unexpected challenges you encountered, particularly concerning the nature of historical (biographical) sources and the then-current capabilities of LLMs for extraction tasks?*

RS: The biggest and initially underestimated challenges were actually three things in retrospect: **First**, precisely operationalizing what we actually want. Because biographical encyclopedia entries contain more information than you might initially think. Take the sentence: "X was a professor of theoretical physics at University Y from 1950–1960." Depending on how explicitly or implicitly you ask and define it, many statements are contained in that one sentence. Just a few examples. Explicitly, for instance:

1. Person X existed (at least between 1950–1960).
2. University Y existed (at least during that time).
3. X held a professorship (at least during that time).
4. X specialized in *theoretical physics* (at least during that time).

3 Available online at: https://polona.pl/preview/ed8bca19-0311-43db-b535-f8c160a449b2

Implicitly (via logic, world knowledge, or domain expertise):

1. X was alive and professionally active (at least between 1950–1960).
2. X taught and/or conducted research at University Y (at least during that time).
3. University Y had a theoretical physics department (at least during that time).

And that's just a small selection of possible inferences. Since we want to ask structural questions, we need standardization/operationalization for that. So: which statements like the ones above actually interest us, and which are actually present in the data? That usually happens through ontologies, and creating them took quite some time, as it does in many KG projects. Later, we reversed the process and began by formulating the questions we wanted the final KG to answer. **Second**, the precision of extracting those statements. At first glance, the results looked surprisingly good. But as so often in digital humanities, the devil is in the details. The more data we tested and the closer we looked, the clearer it became that we needed to proceed very granularly to achieve the desired quality, including quality checks at multiple stages. **Third**, the merging and disambiguation of "identical" entities. That turned out to be one of the hardest problems. Sometimes the same entity is written very differently, abbreviated or spelled out, like "HU Berlin" vs. "Humboldt University Berlin". Sometimes entirely different things have the same name, like two people or places. Or there are OCR/transcription errors. You have to resolve all that if you really want only the "correct" connections in your graph. And of course, across all three tasks, you can observe large differences between time points and model versions in LLM development. The large frontier models are naturally better at extraction with naive prompting and zero-shot approaches. But if you take the right model, prompt cleverly and/or combine techniques, the open, smaller, faster models are often good enough.

AK: Like Raphael already pointed out, the disambiguation challenge stemming from the historical sources themselves was particularly troublesome. In the Little Dictionary, instances are common of the same publications, individuals or institutions being referred to in multiple ways across different dictionary entries, making it difficult to determine whether or not the extracted names all refer to the same person or entity. Additional manual and thus time-consuming checking and matching before further analysis can be conducted was necessary. A separate underestimated challenge relates to the fact that it is difficult to know what the LLMs have been trained on. Likely, it was not nineteenth-century documents in Polish. A slight inconvenience, hence, was ensuring that when LLMs were used for the extraction of information or OCR correction, the spelling and grammar were not 'corrected' to the present-day format. What I also found concerning at the beginning and to some extent continue to do now, is that it is so difficult to ascertain the level of trust we can have in the accuracy of the results we obtain when employing LLMs for information extraction.

MV: I was actually surprised by the quality of the results already at the start. But it was also clear that open-source LLMs did not reach the same quality as SOTA for-pay services. I am still struggling with the idea that paid-for services should be part of a scientific software pipeline. On the one hand because I like to make software reusable and findable but also because a piece of closed-source software will make the research less reproducible. One other thing that is still challenging to me is the design of the actual prompt. To me the results seemed very sensitive to the concrete language and structure of the prompt. In the beginning of the implementation of this idea it was not clear to me how to deal with this sensitivity. How could one be sure that the current prompt gives the correct results? I accept this part of the development work as more of a craft instead of a science at the moment. Maybe I will change my position on this later.

Question: *Could you describe a specific early-stage component or process that didn't work as anticipated? What were the failure points, and what did you learn from that setback regarding the complexity of the task?*

RS: In the beginning, we first defined an ontology and then extracted statements based on its structure. But it quickly became clear that some types of subject-predicate-object (SPO) triples were extracted much more reliably than others. For instance, relations like Person ->has_role -> Role worked very well, almost no relevant cases were missed. Others, like Person -> achieved -> Achievement, performed not as good or inconsistently, which is of course not surprising. Roles are often short, well-defined terms with clear semantics – "physicist", "mayor", "professor" – and are in our sources usually prominently mentioned, often in the first sentence. Achievements, by contrast, are more complex and diffuse, often embedded in longer, narrative-style sentences like: "Through groundbreaking work on gauge symmetry, X helped redefine approaches to unifying gravity and quantum theory." There's no simple verb like "received" here, and the contribution is conceptual, not tied to a concrete title or award. But this also reflects a more fundamental limitation of the graph approach itself: when structuring unstructured text, I trade narrative nuance for operationalization. This inevitably reduces interpretive flexibility, but in return, I gain scaled comparability and the ability to detect structural patterns. These patterns have to be treated with care and rather often point to tendencies, offering interpretation at a different, macro layer. But that's why combining micro- and macro-level interpretation is so crucial. Meaningful analysis requires both. This also means, when designing an ontology, distinctions between classes and relations must be as clear and non-overlapping as possible. The more precise the concept of a statement, the better the extraction, but this applies to humans and machine extractors alike. Also, extraction works better when the area of attention is smaller. Sentence-level extraction, but with the whole paragraph as context is significantly more accurate than directly paragraph-level extraction. This again is no surprise, but the combination of a narrow focus and a broad (or broader) context is an important quality factor. And an insight, that is also common knowledge: LLM ensembling proved highly effective. If one LLM extracts a triple and a second LLM reviews both the sentence and the proposed output of the first, the quality improves noticeably.

AK: One process that did not work as well as anticipated was simply switching between the two biographical dictionaries during testing, from "Who was who in GDR" to "The Little Dictionary". Initially, we were getting much fewer and less accurate results from the Polish source than from the German source, so we had to adapt and adjust the process to the different input dictionaries, for example, by providing context-specific prompts and examples. As Raphael mentioned, another issue we encountered in the early runs was that certain entities were frequently miscategorised; in both cases, it was the persons' achievements that proved particularly difficult to get right.

MV: The earliest version of the approach gave a list of JSON files that included co-mentioned names in each entry. Aleks then had to provide a list of names that were not persons, or differently spelled names and so on. This required several cycles of data cleaning. In later variants we therefore changed the prompt for the LLM to e.g. expand abbreviations or correct capitalizations and so on. This made things easier. Another aspect of solving this was the inclusion of disambiguation parts in the pipeline. Again all of this is just a way to solve the data generation and leave more time for the scientific work, which of course contains a critical reflection of the way the data was generated.

Adaptive Solutions

Question: *The introduction highlights a two-stage architecture and, more generally, Task Decomposition. How did you arrive at this specific design? Was it a gradual evolution or a more distinct pivot based on the early challenges?*

MV: In my memory the two-stage architecture emerged gradually as a result of the data quality. Even with quite good initial results, it was always necessary to reconcile the data e.g. concerning name disambiguations. To me the current architecture is also a result of the question of how to judge the sensitivity of a result to the structure of the prompt. By having several stages with clear target goals, it becomes possible to have an answer to the question of what is the best prompt. We just compare it to the hand-created small datasets for each step.

RS: I would agree with that. It was a gradual process where more and more intermediate steps were added. The more precisely we tried to extract certain things, the more obvious the problems became. And the more granular we made the sequence of steps, the clearer it also became which points in this ladder were problematic. And as said before, current models are usually particularly good when the task is clearly defined. If you want to simultaneously extract statements regarding an ontology and return the extracted things in original Polish form and normalized (e.g., directly in English and spelled out), that's usually worse than first extracting in Polish and then normalizing in a further run. The more granular, the better you usually can formulate few-shot examples, for instance what you exactly mean by normalization, since you only provide examples for very specific tasks, which is more unambiguous. Quality control is easier as well, if you have more steps. Also in terms of modularity and the choice of special models for special tasks (also regarding the future), it became increasingly clear that radical decomposition makes sense.

AK: Quite early on it became apparent to us that to limit the black-box effect and increase our understanding of the overall process, it would be helpful to break it down into smaller tasks and thoroughly interrogate the impact of the individual steps. As Malte and Raphael said, this realization came about as a result of the unexpected peculiarities we were observing in the extracted data.

Question: *Human-in-the-Loop (HiL) is also central to your approach. Could you elaborate on the different HiL stages? For instance, how does the HiL for validating LLM-extracted triples differ from the HiL involved in refining the domain ontology, and what considerations were driving the process?*

AK: We keep tweaking the different steps of the pipeline, so I would say Human-in-the-Loop is omnipresent. Pretty early on in our journey of utilising LLMs for extracting information from historical sources, we had manually created detailed and accurate examples of triples, in the form of subject-predicate-object expressions, based on excerpts from our source material. Based on the close reading of the biographical dictionaries, we also came up with a series of competency questions. Deciding on these questions helped us both to think through the types of information we can actually expect to find in these specific sources, and later to assess the accuracy of the triples extracted using LLMs. The HiL validation of these triples is pretty straightforward and involves returning to the source material and cross-referencing to determine whether the triples are correct or not. We are not expecting complete accuracy, so the difficulty lies rather in deciding what level of accuracy is sufficient and what requires further refining of the obtaining process. Refining the domain ontology is a more involved process, as the complexities of the base language really come to the fore.

RS: There are several points in the current pipeline where HiL is central. These are actually either quality control, i.e., human feedback loops, or questions that are crucial for later analysis. The first central HiL point is the Competency Questions generation step. These are questions that the final Knowledge Graph (KG) or the ontology should be able to answer when properly designed and filled with the appropriate data. In our design, these points are mostly co-working tasks, so in this step, for example, a SOTA reasoning model designs about 25 questions based on the OIE-extracted triples that can be asked of the available data at all.[4] The domain expert then goes through these questions, corrects them, discards them, or adds their own, and this iteratively, until it's clear: these are (1) the questions I want to be able to answer and (2) they are actually askable of the available data via KG. The second important HiL step is the ontology creation. This works basically similar to creating the CQs. A reasoning model makes a proposal based on the already approved CQs and the OIE-extracted triples. So we take on one side what statements are really present, and on the other, what our ontology should be able to answer, and have a first draft for an ontology created from that. A domain expert then goes through this proposal again and revises until the ontology fits the questions and the real data (or sometimes you have to go back to the questions). There are other interfaces, but these two are probably the most relevant. Because that's where it's decided what and how the KG contains information.

4 This idea is based on (Kommineni et al., 2024a, 2024b).

MV: I would say that the LLMs are just tools that help human researchers to perform certain tasks more quickly. So actually I could maybe call this LLMs-in-the-Loop. Similar to traditional research methods the results of LLMs in each stage have to reach a quality that is good enough for the current research question and source material. That is to me the core principle: Each new source material and research question concerned with it requires a new tweaking of the method. There may not be one LLM prompt or one ontology that is useful for all questions. The process of the triple validation was interesting to me since it requires a careful reflection of what implicit statements are contained in everyday language. To give a good example we need to express these, which was not always straightforward to me. While the first step requires careful reading of the sources, the second step requires reflection again but this time on the current research question. To be able to validate the usefulness of an ontology, we had to ask ourselves what information we would need to be able to answer the question. Since I am interested in applying methods from historical network research, I had to go even one step further and consider the potential relation that I would like to query based on the future knowledge graph. This added a level of complexity to me. All of these are examples of the hermeneutic circle of doing research. LLMs are a new tool that requires new ways of reflecting. I guess this proposed pipeline is one way of making these reflections possible.

Question: *Focusing on Stage 1, the LLM-driven Open Information Extraction: Firstly, why did you choose to use OIE? And secondly: What were some of the key learnings in prompt engineering or model selection to get "useful" factual triples from your historical texts, and how iterative was this process?*

RS: We only introduced OIE relatively late, when it became increasingly clear that we wanted to make KG construction seriously data- and research-question-driven, not ontology-driven. There are tons of comprehensive ontologies or even specialized ones for person data, for example. In our experience, the problem was always: if we focus directly on the ontology, we lose sight of what questions we actually want to answer and also what the data potentially provides at all. Moreover, it was clear from the beginning that we wanted to extend the project to other data and not just use it for the two case studies. So at some point we stumbled upon OIE, because it's about structuring all statements contained in text without a prefixed structure. If you then apply this to a representative sample of the original dataset, for example, it often only becomes apparent which statements are actually present or relevant. So before you start constructing a sophisticated ontology, it might become apparent in advance that the structure you wanted to ask about or model is only covered by a minimal fraction of the statements. For instance, out of 40,000 statements, there are actually only 8 that represent a certain relation. That can of course also be interesting and produce a historically meaningful narrative, but to track structural change over time, that's probably too little and likely makes the ontology unnecessarily complex. Regarding prompt engineering, I had already said something about this earlier. The most important learning in my opinion is: the more specific and clearly delineated or defined the task is, the better the output. Good few-shot examples make a huge difference. And model selection should actually be empirically accompanied from the get-go, even though we unfortunately didn't do that consistently. Because sometimes small changes already make a big difference. So classically iterative, always optimizing quality/cost/speed and open-source vs. proprietary. In the newer versions of the pipeline, we consistently include metadata so that every output logs all relevant parameters (model, model parameters, prompt, etc.), so we can systematise what config gives the best output.

MV: I can not say much about the prompt engineering since Raphael did a brilliant job with that. However, the core reason for me to introduce this step was really the point that it became possible to validate the usefulness of each prompt. There was a hugely reduced space left to hallucinations since text units were smaller and it became easier to validate the output quality.

AK: I think a learning outcome regarding prompt engineering that is worth mentioning relates to the fact that we have been working multilingually on this computational pipeline for knowledge graph extraction. The sources I study here are all in Polish, while the sources Raphael has been working with are in German. Our common language is English, and so this is the language we communicate in and discuss the research challenges we face. While Polish, German and English are all Indo-European languages, they differ in their grammars, structure and vocabulary, so necessarily the HiL interventions needed to be specific to the language of the source. This, on the one hand, makes things more difficult, as we can't just freely apply the same fixes across contexts, but on the other hand, it is helpful as it encourages us to interrogate more thoroughly the possible difficulties.

Question: *Regarding Stage 2, structure imposition and validation: How did you balance the need for a domain ontology with the complexities of the source texts? How did the HiL feedback loop contribute to evolving the ontology itself?*

MV: To me the crucial shift in perspective was to get away from the idea of a central knowledge resource that is gradually filled by the triples generated in diverse research projects. Coming from this old perspective I first had in mind to build and extend an ontology based on CIDOC-CRM, since I had used it in previous projects. However, the event-based approach made the resulting structure quite complex, which of course reflected in the necessary SPARQL queries to actually do any analysis. To my mind, by speeding-up the process of triple extraction and ontology building by LLMs we now have the chance to see each ontology and each source as unique, while keeping the option to link to existing knowledge through Linked-Open-Data paradigms. This to me is giving the benefit of structured output while staying flexible to source material and research questions. The validation by SHACL constraints is again an elegant and useful way of limiting the potential of the LLMs to hallucinate. The circular nature of our approach seems natural to the hermeneutic reasoning of traditional research. The full approach is based on the source material and the research questions concerned with it. If the ontology is not able to provide answers to central aspects of the research, we have to adapt the prompts. This process gives opportunity to reflect and explicate, which is very useful in interdisciplinary teams.

RS: I agree with Malte. The goal is a combination of standardization and source focus. I had previously worked with other ontologies and my association was often: modeling data with CIDOC-CRM, for example, uses enormous resources and often seems to be the main task compared to analysing the data modelled that way. But coming from network research, we were actually interested in the analysis. And for that, it's also helpful if the whole process, data creation and analysis, can run iteratively. The potential scale effects of LLMs allow us to proceed more experimentally (without downplaying the necessary historiographical rigour or the significance of environmental impacts). With our current case studies, for example: running the Polish data completely through the current pipeline (with all steps, including linking and disambiguation) takes about 2 hours and costs around $ 8–10. That means if you have to, you can also adjust something afterwards if necessary and do it again. That's not the goal and we of course always test extensively and evaluate the intermediate steps, but what I want to express with this is: you don't need to design a perfect ontology for 2 years that covers all eventualities if you only want to answer three specific questions for the current project.

AK: There is a limit to what we can learn from any given historical source. If the ontology is good enough, it will help the researcher learn not only what is in the source but also identify what is missing. Navigating between the source and developing a domain ontology that connects the information contained in the source with a broader context (through, for example, manually matching entities to their Wikidata item identifiers) can be viewed as having a function of meaning-making, as we make sense of the content in this reiterative process. Through the HiL feedback loop, we have been taking stock of what we can learn from the two biographical sources alone and have identified directions for the necessary further research to be conducted as we proceed to analyse the obtained structured information.

Decision-Making & Trade-offs

Question: *Looking back, what were one or two of the most critical design decisions that significantly shaped the pipeline's current form? What alternatives did you consider, and what were the primary factors or trade-offs (e.g., accuracy vs. scalability, cost vs. control, automation vs. interpretability) that tipped the scales?*

AK: All these factors (accuracy vs. scalability, cost vs. control, automation vs. interpretability) merit careful consideration, and indeed they all played into the decisions we made. For example, it was important for us to keep the costs, both monetary and environmental, as low as possible, but in some cases, the more 'expensive' models worked better for our needs. It was a process of experimentation and balancing different criteria. I would say, however, that perhaps the aspect that often ultimately decided for us was time – whether we had sufficient time to properly implement the design decisions we were considering. We'd ask, given the time constraints we face, what are the most efficient changes and improvements that we can make?

MV: The design of the pipeline in different stages allowed us to select different LLM models depending on the task at hand, some of them quite small-scale. This is crucial to keep the costs low, which would otherwise be a big problem concerning the reproducibility of the results. I would say we gradually got more away from the idea to automate everything and reintroduced the human component more and more. Of course this also means that I can not simply analyse a corpus without the domain knowledge providing the human in the loop aspect. Like I said earlier this also limits the scalability of the approach, but I do not consider that an issue any more.

RS: As I mentioned in other questions: I do think that the three core principles that crystallized, Research-Driven & Data-Oriented, Human-in-the-Loop, and Task Decomposition, were the critical design decisions. So the iterative discovery and then making it a guiding principle, that was critical for the current form. The good thing is that this allows everything to be built modularly, and if there's a better/cheaper/faster model for one step, like OIE extraction, you can simply integrate it there and all other steps aren't affected by it. Alternatively, we thought at the beginning that we might be able to use something out of the box or adapt it slightly, because there are theoretically some packages or other projects that supposedly do something similar, like the Neo4j GraphBuilder (neo4j-labs/llm-graph-builder 2025), but they either do something completely different or have very limited possibilities to control the result, which speaks rather against them in a scientific environment, which ideally incorporates transparency and traceability.

Question: *The project involves team members with different expertise/domain knowledge. How did this interdisciplinary collaboration influence the problem-solving process and the design choices? Were there specific instances where these different perspectives led to a particularly innovative solution or a necessary compromise?*

RS: Interdisciplinary collaboration has definitely influenced how the pipeline looks now and will look in the future. The whole project was designed from the beginning to bring together different domain experts with similar source types (Polish biographies and GDR biographies). And also that it must be extensible to other sources. I think the multilingualism of the sources was decisive in some design decisions, for example. For instance, the separation of extraction and normalization (e.g. from "Biblj. Uniw. Warsz." becomes "University of Warsaw Library") might not have been emphasized so much otherwise, as well as the approach to entity linking and disambiguation (taking a lot of context from the original and the normalized language and emphasizing, for example, diacritics or other language peculiarities for disambiguation). But also the emphasis on what's important or which relationships had to be modelled emerged essentially collaboratively.

AK: I am grateful to have been able to collaborate on this and other projects with Malte and Raphael, because I think together we come up with interesting research directions and creative solutions to the questions we seek to answer. Combining expertise from different domains enables us to identify viable avenues more quickly, thoroughly and effectively than if we were to work on a similar project alone. For example, we can determine together how much pre-processing work would be necessary to attempt to use digital methods for analysis and whether the effort would be worthwhile given the likely output. We can also discuss which questions are well-suited to be answered, and what sorts of arguments we can substantiate, using the historical sources and methods we plan to employ. That is to say that we can work out the necessary compromises early on in the research process. Due to the division in expertise, it does mean that we rely on one another to take the lead in different sections of the project – good teamwork is key. Regarding our current project in relation to deriving knowledge graphs from biographical dictionaries, I believe that our approach is fresh and innovative. I have worked with similar historical documents for more than fifteen years, consulting individual entries, but with the approach we are implementing, I will for the first time be able to use such sources more holistically, and observe the bigger picture they paint.

MV: I find this hard to answer, since interdisciplinarity was such an essential part of the whole process. I am sure, if I would have done something similar alone I would have taken several shortcuts and just looked at the source material as is. The collaboration forced me to become more reflective in my work, which makes things harder but leads to more substantial results in the long run. So I could say the compromise on my side was to take the sources seriously.

Question: *Reflecting on the entire development journey, what advice would you give to other humanities researchers or teams looking to integrate LLMs for complex analytical tasks with historical or nuanced textual data? What are the key "lessons learned" about the "messiness" and iterative nature of such projects?*

RS: I've kind of answered this in other questions: Especially while working with LLMs, proceed iteratively and experimentally. Take the model of the hermeneutic figure eight seriously, that is, the image often used in digital humanities to visually represent the iterative dynamic loop between parts and wholes, data and context, as well as analysis and self-reflection (even if it is sometimes partly generated, see Part 2: LLM Commentary). And only use LLMs in your research, if you are also interested in reflecting their strengths and weaknesses.

AK: My advice would be to remember that while it might be comfortable to think of LLMs as something separate from our messy humanness, they are, after all, fundamentally built on human languages, and humanities scholars, historians included, have much experience in the domain of studying human expression. As Nick Petrić Howe put it, "the 'language' part of the large language model, fundamentally impacts how they can function, because language is intrinsically linked with culture, history, belief, identity, wellbeing, inequity and progress". Moreover, I also think those looking to use LLMs for complex analytical tasks with historical or nuanced textual data, and to bridge the gap between human and computer language, would do well not to neglect the earlier approaches to language analysis from the humanities and Natural Language Processing.

MV: Similar to Aleks, I would say that LLMs should be treated with a grain of salt all the time. Similar to the critical reading of a source, the answer of an LLM requires reflection. And since the LLMs language can be quite convincing, I think we have to be even more alert in this reflection. The basic idea of the pipeline to break down a large task into much smaller tasks to make them understandable and verifiable is key to me. Reflecting on the academic journey that led to this current project as a whole, I want to highlight that although our disciplinary backgrounds and fields of expertise differ, we share an interest in finding ways to address research questions related to knowledge evolution, and I would say that some sort of question or problem that interests everyone involved in a project is vital to the success of a project such as ours. Coming from different disciplines can result in a different vocabulary and focus, so what is also necessary is the willingness to communicate, listen, learn, and teach one another.

Question: *Thinking about the process of developing this pipeline, how did the choices you made in terms of what data to extract, how to structure it, and where to implement HiL, reflect or even shape your own epistemological assumptions about constructing historical knowledge from these biographical sources?*

RS: I would say that, as with other projects, I naturally went into this process with a certain idea of why the data might be interesting and why connecting it could add value. Behind this lies, perhaps, an internalized structuralist epistemology, the assumption that structures matter for historiographical interpretation. That historical persons or phenomena can be better understood, if you look at them both individually, on a micro level, but also on the meso/macro level through their positions in graphs, for instance. The whole project is guided by this idea: that creating connections reveals structures that can help answer structural questions. The problematic thing about the whole matter is somewhat that you potentially reproduce or perpetuate biases (such as collection, gender, training data or authority file bias etc.) because you're investigating structures that are based on a biased data basis or a biased model/algorithm. Sometimes, what you uncover tells you more about editorial practices, like how biographical entries were written, than about the historical setting of the described persons. It's a kind of structural enrichment of the original layer of production, but one that can offer valuable historical insights.

AK: Similarly when evaluating sources as historical evidence — considering their context, purpose, significance, biases, strengths, and limitations — thinking about the process of developing this pipeline has encouraged me to reflect on my role and positionality within the research process. I am keenly aware that the choices we made about what data to extract and how to structure it will contribute to the sort of results we observe. As we ask questions, explain, and argue, our choices contribute to shaping and creating new knowledge. The outputs that we produce will need to be scrutinised by future researchers before they choose to use the results we produce or to otherwise follow in our footsteps. To help with that, it is vital that we not only observe the FAIR principles but also be transparent about our decision-making that went into developing the pipeline. I hope this interview is a step towards that.

MV: One thing that comes to mind is the writing of an entry in the source and the splitting in small facts in the pipeline. How far did the author of e.g. the dictionary consider what information is entailed in a sentence? Could we then potentially enrich the sentence beyond what was intended by the author and would that be problematic? Or are we just making use of the general structure of language to extract new knowledge that the author of the dictionary did not even have? One central point is in addition the fact that we can look at structures of entries. On the other hand if I think e.g. of the index and abbreviations introduced by the author, it seems to reflect a networked knowledge by the author of the considered person. However these details get somehow lost to the readers of a dictionary. In that sense our approach reflects the process of writing the dictionary. Another aspect is our background in research on historical epistemology and what is called socio-epistemic networks, i.e. that knowledge is carried by people, expressed with time-dependent semantics and also has material conditions. I would say this background has strongly influenced our implementation.

Question: *Beyond the specifics of this pipeline, what broader methodological questions or potentials do you see for HPSS in adopting such LLM-driven, HiL-based approaches to knowledge extraction? How might this change the way researchers in HPSS interact with, analyze, or even conceptualize their source material in the future?*

AK: I see in LLM-driven, HiL-based approaches to knowledge extraction a potential for pre-processing information contained in the historical documents. If we develop robust, reliable methods of extracting knowledge graphs we would be able to use them to quickly identify sections of the documents, or indeed entire big corpora that contain the information we are interested in, that we can then proceed to interrogate using our chosen methods of analysis, whether that is network analysis, close reading, or some other well-established historical research method. Essentially, I see in these approaches a way of structuring information contained in historical sources that allows for further analysis using digital methods.

MV: Maybe my answer is a bit on the pessimistic side, but I think there is not really a fundamental change in how HPSS will work with source material. All the practices of critical reflection of source material are still necessary and structural questions have been asked on more limited sets of sources in the past. There might be a real shift in scale however and maybe a chance to include less spoken of actors like lab assistants or collectors of data. That would be a real benefit of such methods.

RS: I agree with the others in part, but would go a bit further. I think scale is essential here. These kinds of extractions and structurings can "unlock" corpora on a structural level much faster, cheaper, and at larger scale than ever before, including previously untapped sources, which also potentially changes the types of questions we ask. For example, with RAG techniques you could locate very specific passages for close reading across many more documents than was previously feasible, while the kind of data creation we do here also makes longue durée, data-driven distant reading analyses on large datasets more accessible. This scaling alone already brings in new methodological approaches from other fields like Complexity Sciences or Computational Linguistics into HPSS. Take what Aleks mentioned: with these techniques, one could for example analyse how certain text components, like introductions or methods sections and so on, change across thousands or millions of scientific publications in a certain field over extended periods. The same applies to other specific text sections across large representative samples of any other text category. So yes, I think LLMs and other ML techniques will and already are changing both how we interact with sources and how we conceptualize research questions and programs around the central question of "what to do with which sources".

Part 2: LLM Commentary

In **Part 1** we documented the development of a two-stage pipeline for knowledge-graph extraction. **Part 2** now utilizes LLMs as *role specific, digital sparring partners* to analyse that process from two standpoints that we think are relevant for (HPSS) humanities toolbuilding: (a) epistemological adequacy within History, Philosophy & Sociology of Science (HPSS) and (b) technical rigour within Computer Science & AI/ML engineering. The following two sub-sections present a commentary generated by different LLMs after being provided with the complete transcript of the interview and the introduction above.

Drawing on the concept of *AI-as-Sparring-Partner* (SP), we treat these commentaries not as decorative "add-ons," but as dialogical instruments designed to challenge and extend our own analytical stance. A related discussion of LLMs as "thinking partners" in humanities research is offered by Wagner and Hermes (2026). Herrmann (2025: 170) defines an AI sparring partner as an artefact that

> interacts with users in a combination of a cooperative as well as a challenging, competitive mode, [...] which either expands the range of ideas and viewpoints considered [...] or fosters discussion and critical thinking about and reflecting on [...] the quality of ideas or problem solving [...], the skills and knowledge of the user [...] or the performance of the AI system itself.

Such a dual mode of cooperation and challenge aligns with Human-Centered AI's call for keeping the "human-in-the-loop" while still provoking cognitive stretch (Shneiderman, 2022; Dellermann et al., 2019). A well-designed sparring session can offer several potential benefits (Herrmann, 2025: 166): (1) perspective broadening: exposing blind spots by injecting counter-arguments or alternative framings (2) critical-thinking: obliging the human team to reflect and/or justify theoretical or design decisions, echoing work on generative AI and higher-order reasoning (Kasneci et al., 2023) and (3) reflexive benchmarking of one's own ideas and methods, testing whether the pipeline's narrative can withstand critique from systems that were not involved in its construction, thus functioning as an externalised assessment loop.

However, these same benefits can constitute fundamental limitations. LLM-generated commentary often lacks broader situated or very specific "understanding" of research contexts, cannot draw on embodied experience, may be factually incorrect and may reproduce biases embedded in their training data while appearing neutral.[5] For this reason, we do not outsource final evaluation or conceptual decision-making to machines. Rather, we solicit structured, externalised critiques that can expose blind spots in the messy, iterative, and collaborative realities of integrating AI into humanities scholarship.

On the implementation side we created an OpenRouter[6] chatroom querying eight state-of-the-art LLMs in parallel with synchronised prompts and sampling parameters. All logs (model versions, system prompts, model parameters, and full chat logs) are accessible unedited via Zenodo[7], alongside their summaries for both disciplinary roles.

HPSS research perspective analysis

The first commentary prompt tasked LLMs to provide a critical analysis from the perspective of a senior HPSS researcher. This persona was chosen to critically interrogate the project's theoretical assumptions and methodological choices, handling of source nuance, and underlying epistemological assumptions, thereby challenging the core of the historical research design. The initial prompt given to each of these models is shown below. Subsequently, a summary meta-commentary was created from the eight resulting analyses, for the purpose of reducing text length and to highlight consensus and unique standpoints of the different models.

Prompt Used: HPSS Perspective Analysis

Critically analyse the preceding interview transcript and introduction from the perspective of a senior researcher in the History, Philosophy, and Sociology of Science (HPSS). Focus on the following aspects:

- Implications for historical methodology.
- How the team handled source nuance, ambiguity, and potential biases.
- The epistemological assumptions seemingly underlying the project.
- The alignment of the project's goals and methods with core concerns in HPSS.
- How the described process reflects the "messiness" of integrating computational tools in humanities research.

Be critical. Be concise. Don't overhype. No bloat.

From the eight resulting analyses, a summary/meta-commentary was generated using the synthesis prompt below:

5 For a broader discussion of the ethical and epistemic implications of using LLMs in humanities and HPSS research, including issues of bias, data labour, and the situatedness of machine "understanding", see Lang (2026).

6 OpenRouter is an AI model router that provides an API endpoint to access a diverse array of models from various providers, and also features a chat interface for simultaneous, comparative prompting of multiple models using consistent parameters. See: https://openrouter.ai/chat.

7 You can find the materials under https://doi.org/10.5281/zenodo.16737269

Prompt Used: Synthesis Prompt

Instructions and Rules:

Role: You are an AI assistant specialised in synthesising multiple complex text sources into a coherent meta-commentary.
Context: Your sources are the eight available LLM responses that provide a critical analysis of an interview from the perspective of an HPSS researcher.
Task: Produce a thematically ordered meta-commentary that summarises the core statements, points of consensus, and unique critiques across the eight responses. The text must emulate their critical, analytical tone and use a clear, paragraph-based structure.

1. Headings and Structure: Use the following headings **verbatim and in this order**.

- Implications for Historical Methodology
- Handling Source Nuance, Ambiguity, and Bias
- Epistemological Assumptions
- Alignment with Core Concerns in HPSS
- The "Messiness" of Integrating Computational Tools
- Brief Summary

2. Content Synthesis:

- For each thematic heading (the first five), provide a substantive synthesis in prose (< 200 words recommended). In your paragraph(s), first identify Consensus points, then highlight Distinctive perspectives.
- For the final "Brief Summary" heading, provide a single, more detailed paragraph distilling the core critique (< 100 words).
- Include short illustrative quotations (< 25 words) where helpful.

3. Citation Requirements (Strict):

- Every claim—quoted or paraphrased—must carry an in-text citation in this form: (Model-Name).
- For multiple models, separate names with semicolons, e.g., (Grok-4; O3-Pro; DeepSeek-RI).
- Use these keys exactly: Gemini-2.5-Pro, Claude-Opus-4, Grok-4, DeepSeek-R1, O3-Pro, Qwen3-235B, Magistral-Medium, Llama-4-Maverick.

4. Style and Limits:

- Write in a scientific style adhering to the principle: "Be critical. Be concise. No over-hype. No bloat."
- Avoid first-person narration.
- Do not introduce information or opinions not present in the sources.

Keep the total length under ∼1100 words.

LLM-Response:

Summary

Implications for Historical Methodology

A broad consensus across the analyses posits that the pipeline's core methodological implication is a trade-off that exchanges "narrative nuance for operationalization" (Gemini-2.5-Pro; Claude-Opus-4). This process of "methodological flattening" (DeepSeek-R1; Qwen3-235B) inherently privileges quantifiable, relational facts over ambiguous or contextual meaning, risking the reinforcement of a positivist bias in historical inquiry (Grok-4; Qwen3-235B). The models agree that while the shift to a data-driven, ontology-last approach is an intelligent adaptation, the very act of structuring sources into subject-predicate-object triples imposes a framework that may not align with the historical context (Claude-Opus-4).

Distinctive critiques further refine this point. The primary scholarly act is seen as shifting from interpretive reading to the design and validation of an "extraction and structuring apparatus" (Gemini-2.5-Pro). This introduces a new methodological risk: the research agenda can be subtly shifted from the historian to the model, particularly when using LLM-generated "competency questions" to guide the inquiry, which may naturalise accidental patterns in the training data (O3-Pro). The loss of nuance is thus not merely an acceptable trade-off but a "historical problem to be theorised" in its own right (O3-Pro).

Handling Source Nuance, Ambiguity, and Bias

The analyses concur that while the project team demonstrates awareness of surface-level ambiguities like name variants and OCR errors, its handling of deeper, systemic biases is underdeveloped (Grok-4; DeepSeek-R1). A central point of consensus is that the pipeline risks laundering or reifying the inherent biases of the source material—such as editorial selection or gender representation—by converting them into structured, seemingly objective data (Gemini-2.5-Pro; Qwen3-235B). The strategy of linking to authority files like Wikidata is identified as particularly problematic, as it imports contemporary, often Western-centric, categorisations and their own latent biases into historical contexts (DeepSeek-R1; Grok-4).

Unique perspectives sharpen this critique. One analysis argues that nuance, such as the “diffuse” nature of an achievement, is treated as “noise to suppress” rather than as historically significant data (Qwen3-235B). Another notes that ambiguity is not embraced as a feature of the source but is instead “forced to be determinate” by the extraction process, which atomises sentences and strips them of their rhetorical function (O3-Pro). The most critical missing step is identified as a robust analysis of how source biases and model biases interact and amplify each other, a dynamic that the current technical mitigations do not adequately address (Gemini-2.5-Pro).

Epistemological Assumptions

There is a strong consensus that the project is grounded in a “structuralist epistemology” (Gemini-2.5-Pro; Grok-4) or a form of “structuralist positivism” (Qwen3-235B). This stance assumes that historical knowledge is structurally extractable and that relational patterns revealed in graphs constitute meaningful insight. This approach is seen to carry a “latent positivism” (Claude-Opus-4) and a degree of “techno-optimism” (Grok-4), where language is treated primarily as a carrier of information rather than as a form of social action (O3-Pro). The pipeline’s design, therefore, implicitly de-prioritises forms of knowledge that are not easily structured or quantified (Gemini-2.5-Pro).

Distinctive critiques elaborate on this foundation. One analysis frames the core epistemological problem as a risk that “what can be easily structured” becomes a proxy for “what is important to know” (Gemini-2.5-Pro). Another challenges the iterative refinement process as implying a “quasi-asymptotic notion of truth”, which is at odds with constructivist understandings of how facts are produced within situated practices (O3-Pro). The claim of a “data-driven” ontology is also challenged as illusory, since the LLM prompts used in the initial extraction stage already impose implicit, non-neutral categories on the source material (Qwen3-235B).

Alignment with Core Concerns in HPSS

The analyses generally agree that the project’s stated goals—exploring socio-epistemic networks and knowledge circulation—align with established HPSS interests (Grok-4; Claude-Opus-4). The reflexive documentation of the research process is also seen as a positive contribution to the field (O3-Pro). However, this alignment is consistently described as partial or incomplete. The consensus is that the project’s methods do not sufficiently engage with critical HPSS concerns regarding power asymmetries, the sociology of knowledge tools, or post-colonial critiques of data and archives (Grok-4; DeepSeek-R1; O3-Pro). Divergent points specify this gap. One critique warns that the focus on formal, documented affiliations may inadvertently reinforce “great-man narratives” by overlooking the informal networks and practices central to STS scholarship (Claude-Opus-4). Another argues that the project reduces the practice of source criticism to “accuracy validation”, failing to interrogate how the sources themselves function as ideological or knowledge-making artifacts (Qwen3-235B). A more profound HPSS framing, it is suggested, would use the pipeline not merely to extract data from sources but to ask critical questions about the construction of those sources (Gemini-2.5-Pro).

The “Messiness” of Integrating Computational Tools

All analyses commend the interview format for successfully capturing the technical “messiness” of the research process, including the iterative design pivots, the “craft” of prompt engineering, and the pragmatic negotiation of trade-offs between cost, accuracy, and control (Gemini-2.5-Pro; Grok-4). This transparent documentation of a non-linear workflow is considered a valuable contribution to the digital humanities community (Gemini-2.5-Pro; O3-Pro).

However, several of the more critical analyses argue that this documented messiness is largely confined to technical implementation, while the deeper epistemological messiness is “somewhat sanitized” or glossed over (Grok-4; Qwen3-235B). A distinctive critique is that the narrative presents an underlying belief in technical solutions for what are “fundamentally interpretive problems” (Claude-Opus-4). The pipeline’s modularity is also seen as masking how each stage embeds its own non-neutral assumptions, which are not subject to the same level of critical reflection as the final output (Qwen3-235B). The real messiness—how computational constraints actively reshape the historical questions being asked—remains under-examined (Qwen3-235B).

Brief Summary

The collective critique finds the project to be a methodologically transparent and technically sophisticated effort that nonetheless remains epistemologically shallow. Its core weakness, identified across the analyses, is the treatment of fundamental historiographical challenges as technical problems amenable to computational solutions. The pipeline operationalises historical inquiry by trading narrative depth for quantifiable structure, a process that risks reifying source biases and flattening interpretive complexity. While the Human-in-the-Loop framework is praised for managing surface-level errors, it is seen as insufficient for confronting how the computational apparatus itself co-produces historical meaning. The project thus serves as a valuable case study, but one that highlights the persistent gap between computational capability and deep historical critique.

Technical engineering perspective analysis

In the second round of the sparring process, the LLMs were prompted to take the role of a senior AI/ML engineer. This perspective shifts the focus of the critique towards the pipeline’s technical architecture, scalability, and maintainability. The aim here is to “stress-test” the functionality and design choices from a practical implementation standpoint, providing a complementary, engineering-focused evaluation. As with the HPSS perspective, the outputs from multiple models were synthesized into a meta-commentary.

Prompt Used: Technical Engineering Perspective

Critically analyse the preceding interview transcript and introduction from the perspective of a Senior AI/ML Engineer specializing in Natural Language Processing and Digital Humanities. Focus on the following aspects:

- Evaluation of the technical design choices mentioned. Potential optimisations or alternative approaches.
- Assessment of the task decomposition strategy.
- Effectiveness and implementation details of the Human-in-the-Loop (HiL) components from a systems view.
- Considerations regarding scalability, modularity, and maintainability.
- How the described process aligns with the known capabilities and challenges of LLMs in scholarly workflows.

Be critical. Be concise. Don't overhype. No bloat.

LLM-Response:

Summary

Evaluation of the technical design choices mentioned. Potential optimisations or alternative approaches.

There is a broad consensus that the two-stage architecture—commencing with ontology-agnostic Open Information Extraction (OIE) before imposing a schema—is a methodologically sound and pragmatic design choice (Gemini-2.5-Pro; Grok-4; O3-Pro; Claude-Opus-4). This approach grounds the final knowledge graph in the source material's content rather than researcher preconceptions (Gemini-2.5-Pro). Similarly, the use of an "LLM-as-a-judge" for ensemble validation is recognised as a common pattern for improving quality, though it is criticised for increasing cost and complexity (Gemini-2.5-Pro; Grok-4; DeepSeek-R1; Qwen3-235B). The reliance on proprietary models is consistently identified as a primary weakness that compromises reproducibility and inflates costs (Grok-4; O3-Pro; Claude-Opus-4; Qwen3-235B).

Distinctive critiques and optimisations were offered. Several models suggest that fine-tuning smaller, open-source models on domain-specific data would be more efficient and cost-effective than relying on prompt engineering with large, general-purpose APIs (Grok-4; O3-Pro). One analysis proposes that modern "closed-IE" models could potentially render the second mapping stage redundant by extracting ontology-ready triples directly (O3-Pro). The entity linking strategy was also scrutinised, with one critique dismissing the character-embedding-based blocking as "over-engineered" for historical text and suggesting sourcespecific gazetteers as a more practical alternative (Qwen3-235B). Finally, the claim of using "multimodal capabilities" was deemed an overstatement for what is essentially a standard document processing task (Gemini-2.5-Pro).

Assessment of the task decomposition strategy.

The "radical decomposition" of the workflow into granular, single-purpose steps is praised across the board as an exemplary design principle (Gemini-2.5-Pro; Grok-4; Llama-4-Maverick; Qwen3-235B). This modularity is seen as a core strength that enhances debuggability, allows for targeted model selection, and represents a logical evolution from the initial, less successful monolithic approach (Grok-4; Claude-Opus-4). The strategy of focusing on "sentence-level extraction with paragraph context" is noted as a fundamental and effective NLP practice (Claude-Opus-4).

However, this decomposition is not without its flaws. A distinctive critique identifies a "critical flaw" of circular validation, where Competency Questions (CQs) derived from OIE output may fail to reflect genuine research needs if the initial extraction was incomplete (Qwen3-235B). Other analyses warn of the high risk of error propagation and increased latency across the long chain of tasks (O3-Pro; Claude-Opus-4). From a systems perspective, the decomposition lacks a formal orchestration layer (e.g., Airflow) to manage dependencies (Grok-4; O3-Pro), and one model warns of accruing "integration debt" without a strict, version-controlled data schema to define the contracts between the discrete stages (Gemini-2.5-Pro).

Effectiveness and implementation details of the Human-in-the-Loop (HiL) components from a systems view.

The models concur that the integration of Human-in-the-Loop (HiL) is essential and well-conceived, correctly focusing expert validation on high-leverage, epistemologically critical tasks such as refining the ontology and Competency Questions (Gemini- 2.5-Pro; Grok-4; Llama-4-Maverick; O3-Pro). Despite this conceptual strength, the implementation is uniformly assessed as ad-hoc, synchronous, and a significant scalability bottleneck that relies on unstructured expert availability (Gemini-2.5-Pro; Grok-4; DeepSeek-R1; Qwen3-235B).

A strong consensus emerged around a key distinctive critique: the failure to implement active learning loops. Such loops would prioritise human review on the most uncertain or ambiguous LLM outputs, thereby systematically reducing the expert workload (DeepSeek-R1; Claude-Opus-4; O3-Pro; Qwen3-235B; Grok-4). Another pointed critique is the missed opportunity to use the valuable human feedback to fine-tune the models, which means the system augments the human but does not "learn from them systemically" (Gemini-2.5-Pro). The lack of structured protocols, such as inter-annotator agreement metrics or dedicated annotation UIs, was also highlighted (Grok-4). One analysis aptly re-frames the process as "LLM-in-the-Loop", a term that better captures how human expertise remains the ultimate driver of the workflow (Grok-4).

Considerations regarding scalability, modularity, and maintainability.

There is a clear consensus that while modularity is the pipeline's greatest strength, scalability is its most significant weakness (Gemini-2.5-Pro; Grok-4; Llama-4-Maverick; Magistral-Medium). The modular design allows for individual components to be updated or replaced as technology evolves (Gemini-2.5-Pro). However, the current implementation is deemed unscalable due to the combination of high proprietary API costs, long processing times, and the non-scalable dependency on expert time for HiL validation (Gemini-2.5-Pro; Grok-4; Claude-Opus-4; Qwen3-235B). The practice of logging metadata is considered a necessary but insufficient step toward true maintainability and reproducibility (Gemini-2.5-Pro; Grok-4; Claude-Opus-4).

Distinctive critiques focus on the lack of MLOps rigor. The description of prompt engineering as a "craft" is heavily criticised as unmaintainable, with multiple models calling for systematic prompt versioning and automated testing frameworks (Gemini-2.5-Pro; Qwen3-235B; O3-Pro). The claim of achieving "partial reproducibility" is pointedly dismissed as "inadequate" and "reproducibility theater" given the use of non-deterministic, closed-source models (Claude-Opus-4). To achieve genuine modularity, a formal contract, such as a versioned I/O schema, is required between each component (O3-Pro).

How the described process aligns with the known capabilities and challenges of LLMs in scholarly workflows.

The analyses agree that the project aligns well with emerging best practices for using LLMs in the humanities: as a tool for augmentation and exploration, not blind automation (Gemini-2.5-Pro; Grok-4; Llama-4-Maverick). The team's thoughtful reflection on the trade-off between "narrative nuance" and "operationalization" is praised as astute and central to digital humanities discourse (Gemini-2.5-Pro). The pipeline's architecture correctly acknowledges and attempts to mitigate known LLM challenges like hallucination through its extensive HiL components (Grok-4; Magistral-Medium).

However, several distinctive critiques argue the approach underestimates key LLM limitations. One sharp analysis contends that the "ontology-agnostic" OIE stage is misleading because it inevitably "embeds the prompt designer's ontology" (Qwen3-235B). Others note that while the team reflects on source and model bias, the pipeline itself lacks technical components to actively measure, audit, or mitigate it (Gemini-2.5-Pro; O3-Pro). A "critical gap" is the absence of confidence scoring for LLM outputs, which would provide more transparent error analysis (DeepSeek-R1). Finally, the risk of LLMs "correcting" or modernizing historical language is highlighted as a significant challenge that erases crucial period-specific nuance (Grok-4; Qwen3-235B).

Brief Summary

The collective analysis assesses the described pipeline as a methodologically thoughtful and pragmatically designed research prototype that successfully integrates scholarly rigour with LLM-driven automation. Its core strengths lie in its modular, two-stage architecture and its principled centering of human expertise. However, from a systems engineering perspective, the project is critically flawed by significant operational weaknesses. The ad-hoc implementation of Human-in-the-Loop components creates an unscalable bottleneck, while the heavy reliance on costly, non-deterministic proprietary APIs undermines reproducibility and long-term viability. The pipeline lacks standard MLOps practices such as active learning, prompt versioning, and automated regression testing. Consequently, it functions more as a well-documented proof-of-concept for an exploratory process than as a robust, scalable, or maintainable piece of scholarly infrastructure.

Conclusion

This contribution sought to document the development of a knowledge graph extraction pipeline through an experimental format: an expert interview complemented by LLM-generated commentary. The approach highlights tensions that are fundamental to computational humanities work in general.

Our pipeline emerged from a straightforward goal, namely to unlock the structural information latent in biographical dictionaries. The resulting two-stage architecture attempts to balance data-driven discovery with research-question-driven structuring, incorporating Human-in-the-Loop steps that reflect our conviction that scholarly decision-making requires human supervision and/or intervention. The LLM commentaries highlighted fundamental tensions embedded within the proposed pipeline, such as narrative nuance vs. operationalization, interpretive flexibility vs. scalable comparability, and reproducibility vs. practical feasibility, among others.

In this sense, the experimental use of LLMs as disciplinary "sparring partners" proved fruitful. Their critiques targeted several supposedly undervalued/underdeveloped points or blind spots. A Whiggish tendency in our technological framing that implied inevitable progress through a focus on structure and automation, the circularity between structuring data and presuming a structuralist epistemology, or the risk of naturalizing design decisions that are, in fact, contingent. They also questioned the privileging of disambiguation over ambiguity and how our claims to transparency clashed with the black-box nature of the AI systems used. However, these analyses also exposed the models' limitations, specifically their context quality-dependence. For instance, the epistemic tension between narrative nuance and operationalization, a well known debate in Historical Network Research that we hinted at in the interview, was treated by the models as a blind spot rather than a deliberate design choice. A human reviewer might have weighted this differently while possibly overlooking other issues. This mirrors the potential and risk of these tools: they sometimes can augment scholarly analysis well while subtly reshaping the focus of our critique in other cases.

Three insights, while not new, were highlighted by this exercise. First, the "messiness" of integrating computational tools in humanities research is not merely technical but fundamentally epistemological, each design decision embeds assumptions about what historical knowledge is and how it can be represented. Second, transparency about this process matters since documenting not only what worked but also what failed, as well as the potential solution and its iterations, reveals the interpretive work that computational methods often obscure. Third, the structuralist epistemology inherent in knowledge graph approaches is not a bug but a feature, one that demands theoretical framing rather than technical "workarounds".

This experiment also suggests that the focus of tool-building in computational humanities lies not only in mitigating the tension between technical capabilities and interpretive nuance, but in rendering that tension productive by making the underlying negotiation processes transparent. The pipeline extracts structures that enable new forms of analysis, but only by imposing frameworks that themselves require critical reflection. In this sense, our tools and their theoretical framing are active participants in the construction of historical knowledge, a role that, as this experimental documentation might show, benefits from both human reflection and machinic provocation.[8]

References

Dellermann D, Calma A, Lipusch N et al. (2019) The future of human-AI collaboration: A taxonomy of design knowledge for hybrid intelligence systems. In: *Proceedings (HICSS 2019)*. DOI: 10.24251/HICSS.2019.034.

Boulanger C (2026) The potential of LLMs for constructing a socio-legal knowledge graph. In: Simons A, Wüthrich A, Zichert M, et al. (eds) *Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science*. Bielefeld: transcript, part-4.

Erickson F (1985) *Qualitative Methods in Research on Teaching*. East Lansing, MI: Institute for Research on Teaching, Michigan State University. (Occasional Paper, Institute for Research on Teaching, Michigan State University).

Herrmann T (2025) AI as a sparring partner – An HCAI approach to promote human capabilities. In: Degen H and Ntoa S (eds) *Artificial Intelligence in HCI*. Cham: Springer Nature Switzerland, pp. 162–176. DOI: 10.1007/978-3-031-93415-5_10.

Kasneci E, Sessler K, Küchemann S et al. (2023) ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences* 103: 102274. DOI: 10.1016/j.lindif.2023.102274.

Kaye A, Schlattmann R, Vogl M, et al. (2024) Socio-epistemic networks: A framework for history of knowledge. In: Lerg C et al. (eds) *Experimental spaces: knowledge production and its environments in the long nineteenth century*. Berlin: De Gruyter Oldenbourg. pp. 253–274. DOI: 10.1515/9783111291383-012.

8 This chapter was written with support from large language models (LLMs). All model-generated text was reviewed and, where necessary, rewritten by the authors, who remain fully responsible for the final version. For details on the use of LLMs in this volume, see the statement in the volume's introduction.

Kommineni VK, König-Ries B and Samuel S (2024a) From human experts to machines: An LLM-supported approach to ontology and knowledge graph construction. Preprint. DOI: 10.48550/arXiv.2403.08345.

Kommineni VK, König-Ries B and Samuel S (2024b) Towards the automation of knowledge graph construction using large language models. In: *Proceedings of the 3rd NLP4KGC International Workshop on Natural Language Processing for Knowledge Graph Creation*, pp. 19–34. Available at: https://ceur-ws.org/Vol-3874/paper2.pdf (accessed 29 October 2025).

Lang S (2026) Critical concerns for using LLMs in the (computational) humanities and beyond. In: Simons A, Wüthrich A, Zichert M, et al. (eds) *Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science*. Bielefeld: transcript, part-1.

Meding H and Daugs A (2026) On the use and limitations of large language models in historical scholarship. In: Simons A, Wüthrich A, Zichert M, et al. (eds) *Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science*. Bielefeld: transcript, part-1.

Müller-Enbergs H, Wielgohs J, Hoffmann D et al. (2010) *Wer war wer in der DDR? Ein Lexikon ostdeutscher Biographien*. 5th ed. Berlin: Links.

Neo4j Labs (2025) *LLM Graph Builder*. Available at: https://github.com/neo4j-labs/llm-graph-builder (accessed 29 October 2025).

Shneiderman B (2022) *Human-Centered AI*. Oxford and New York: Oxford University Press.

Wagner A and Hermes J (2026) Encoded humanities, or: not everything has to be generative. A dialogue on AI tasks and roles. In: Simons A, Wüthrich A, Zichert M, et al. (eds) *Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science*. Bielefeld: transcript, part-1.

Weber H and Herbst A (2013) *Deutsche Kommunisten: biographisches Handbuch 1918 bis 1945*. Berlin: Dietz.

Weiß L, Von Welczeck L, Vogl M et al. (2025) Past, present, and future of HNR: Reflections on the practices and methods in historical network research based on a quantitative survey. *Journal of Historical Network Research* 11(1). DOI: 10.25517/JHNR.V11I1.103.

Zieliński S (1933) *Mały słownik pionierów polskich kolonialnych i morskich: podróżnicy, odkrywcy, zdobywcy, badacze, eksploratorzy, emigranci – pamiętnikarze, działacze i pisarze migracyjni*. Warszawa: Instytut Wydawniczy Ligi Morskiej i Kolonialnej.