# EHRI Annotator: A Web-Based Tool for Named Entity Recognition and Linking in Holocaust-Related Texts

**Maria Dermentzi**[1,2]
[1]EHRI-CZ, Prague, Czechia
[2]Toolbox 21 Single Member PC, Kavala, Greece
maria@toolbox21.com

**Abstract**

This paper presents the EHRI Annotator, a web-based tool for multilingual named entity recognition (NER) and entity linking (EL) in Holocaust-related texts. The tool was developed to support services provided by the European Holocaust Research Infrastructure (EHRI), primarily the digital scholarly editions published by EHRI (EHRI Online Editions) by streamlining the process of detecting named entities in documents and linking them to their unique identifiers in EHRI and third-party controlled vocabularies and gazetteers. The EHRI Annotator builds upon previous work on domain-specific NER, taking it a step further to support multilingual EL. The tool adopts a dual entity linking architecture that uses a different matching approach depending on the type of the named entity. It performs semantic matching for entities to be linked to EHRI vocabularies and authority sets which are modestly sized, and string-matching-based retrieval for locations to be linked to the extensive GeoNames gazetteer using a domain-specific relevance weighting. A preliminary evaluation on 264 entities from a manually annotated dataset of Holocaust testimonies in three languages (English, German, Hungarian) yields an Accuracy@5 of 77.7% when it comes to the linking component of the tool. User testing confirms the tool's usability but also highlights areas for improvement.

**Keywords:** named entity recognition, entity linking, Holocaust studies, digital humanities, cultural heritage, multilingual NLP, digital editions

## 1. Introduction

A core service of the European Holocaust Research Infrastructure (EHRI)[1] is the publication of *EHRI Online Editions*[2], which are digital scholarly editions of thematically curated documents. Since 2018, EHRI has supported the publication of seven[3] EHRI Online Editions. The documents included in an EHRI Online Edition primarily include Holocaust testimonies, as well as diplomatic reports and correspondence hosted by different archival institutions around the world. To be included in an EHRI Online Edition, these documents are manually annotated with Extensible Markup Language (XML) according to the Text Encoding Initiative (TEI) P5 guidelines. Specifically, subject matter experts affiliated with EHRI partner institutions enrich the documents with semantic annotations to highlight and give context about the people, locations, organizations, and topics mentioned in them. Where applicable, these annotations also include links to unique identifiers in the EHRI vocabularies [4] and authority sets[5], as well as to GeoNames[6] according to the annotation guidelines created by EHRI[7]. This manual annotation process is extremely resource-intensive because it requires not only close reading of the documents by domain experts but also strong familiarity with large knowledge bases used for linking the named entities found in texts with their associated unique identifiers. At the same time, the result of this annotation process is very useful because it produces documents interlinked with common access points based on semantic similarities regardless of the source collection, enabling research on a specific topic using diverse and often transnational material.

This paper presents the EHRI Annotator[8], a web-based tool that was primarily built to streamline the process of named entity recognition (NER) and entity linking (EL) for Holocaust-related texts being prepared for publication as EHRI Online Editions[9]. The tool builds on previous work on domain-specific NER (Dermentzi and Scheithauer, 2024), which included fine-tuning a multilingual language model for Holocaust-related entity recognition (the EHRI-NER model[10]) using a dataset compiled from the EHRI Online Editions, making it a fit-for-purpose NER model because it has *"learned"* and *"inherited"* the annotation patterns and conventions followed

[1]EHRI project website. Accessed 2/25/2026.
[2]EHRI Online Editions webpage. Accessed 2/25/2026.
[3]At the time of writing on February 25th, 2026.
[4]EHRI Vocabularies. Accessed 2/25/2026.
[5]EHRI Authority Sets. Accessed 2/25/2026.
[6]GeoNames website. Accessed 2/25/2026.
[7]TEI encoding and annotation documentation page. Accessed 2/25/2026.
[8]EHRI Annotator website. Accessed 2/25/2026.
[9]The tool is publicly available as a beta service. The source code is not publicly released at this time.
[10]Available on Hugging Face.



*Proceedings of The Second Workshop on Holocaust Testimonies as Language Resources (HTRes) @ LREC 2026*, pages 37–46
11 May 2026. ©ELRA Language Resources Association (ELRA), 2026

by EHRI Online Edition editors. In the EHRI Annotator, the EHRI-NER model is deployed as part of the NER component of the tool's pipeline, which is the first stage following the user's input of a text. The other major stage of the pipeline is the linking component which is triggered when the user decides that a match should be attempted for any of the entities detected by the NER component. For EL, EHRI Annotator employs a hybrid lexical and machine learning-based strategy to retrieve and suggest to the user the top five potential matches for the named entities that have been detected and deemed linkable by the user. Although the primary aim of the tool is to be a user-friendly platform that facilitates the annotation process when preparing new EHRI Online Editions, the tool can be easily modified and extended to support more use cases, such as metadata enrichment for enhancing the catalog of an archival institution or an aggregator like the EHRI Portal with more interoperable links according to the Findable, Accessible, Interoperable, and Reusable (FAIR) principles (Wilkinson et al., 2016), making the documents more findable and accessible and interlinking dispersed sources across institutions and languages. It can also be useful for enhancing a knowledge graph like the one described by García-González and Bryant (2023).

This paper's contributions include: a) the description of the EHRI Annotator pipeline; b) the hybrid architecture used for entity linking which employs a different strategy per entity type for more efficient candidate retrieval; c) preliminary evaluation results of entity linking accuracy on a manually annotated dataset of Holocaust testimonies; d) findings from user testing sessions. The NER model used in the EHRI Annotator is unchanged from Dermentzi and Scheithauer (2024).

## 2. Related Work

This work forms part of a broader effort to offer reliable named entity recognition and linking services in ways that support metadata enrichment of Holocaust-related archival material for the purposes of indexing and information retrieval but also for the contextualization of this material within an international landscape of dispersed, multilingual archival resources. Previous work (Dermentzi and Scheithauer, 2024) described why this is useful for EHRI and its services while also making the first step towards offering a multilingual NER model for the Holocaust domain. Having access to a reliable enough NER model was key to progressing towards finding an EL approach that is reasonably efficient and accurate for our use case. While in recent years there has been a lot of focus on NER and EL for historical documents and this research topic has been addressed as the target of shared tasks (Ehrmann et al., 2022b, 2020, 2022a), previous linking approaches typically target Wikidata as the knowledge base to link to, whereas domain-specific vocabularies like the ones used to link entities in the EHRI Online Editions are not prioritized as much. For a comprehensive survey of NER and entity disambiguation in historical documents, see Ehrmann et al. (2023).

The named entity linking approach followed in this paper was inspired by the work of Arora and Dell (2024) on the LinkTransformer package. LinkTransformer treats record linkage as a text retrieval task, where entities are encoded into dense vector representations using a transformer language model and then cosine similarity over these embeddings is being measured to retrieve the nearest neighbor in the target knowledge base. This approach supports multilingual matching without translation, as multilingual models such as LaBSE (Feng et al., 2022) map texts in different languages into a shared embedding space. Before developing the EHRI Annotator, the author experimented with the LinkTransformer package and found its approach to be an effective strategy for linking against the EHRI controlled vocabularies, which contain approximately 13,000 entries. However, the same was not true for linking against the GeoNames gazetteer, where the sheer volume of toponyms and aliases makes embedding-based retrieval impractical in terms of indexing cost and inference times. As described in Arora et al. (2024), the GeoNames gazetteer is so large and comprehensive that toponym disambiguation can be very accurate using string matching methods alone.

Following Arora et al.'s (2024; 2024) insights, for EHRI entities, the EHRI Annotator adopts the dense retrieval paradigm using LaBSE embeddings indexed in a vector database, while for geographic entities, we employ text-based retrieval with domain-specific relevance weighting. This dual-strategy design is explained by the differences among the target knowledge bases. The EHRI vocabularies and authority sets contain limited alias and translation coverage with entries appearing mostly in English under their preferred name, making semantic matching essential for cross-lingual recall. GeoNames, conversely, is rich in translations and alternative names across many languages, making string-based matching both sufficient and more scalable.

## 3. System Architecture

The EHRI Annotator (screenshot in Figure 1 below) consists of three components: a web-based user interface for document input and annotation verification and exporting; an NER processing backend that segments input texts and runs inference

using the EHRI-NER model via an asynchronous task queue; and an EL microservice powered by a vector database (Qdrant[11]). Having the NER and EL components decoupled, with the EL service operating as an independent application programming interface (API), was a deliberate decision to allow greater flexibility on how each component is maintained, scaled, and deployed, anticipating the need to eventually expand coverage and allow other EHRI services to query the EL API independently for metadata enrichment outside the EHRI Annotator.

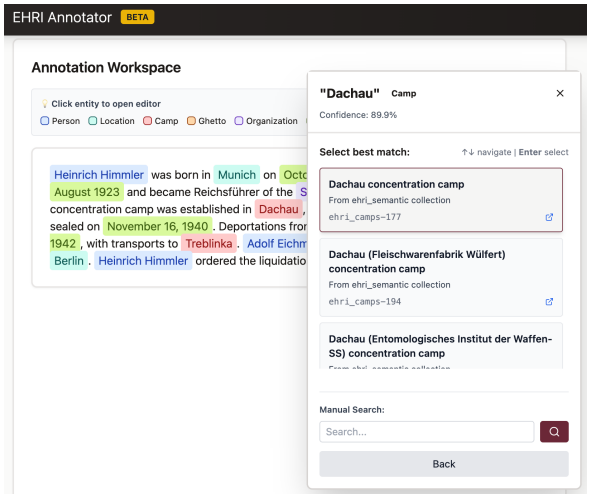


Figure 1: Screenshot of the entity linking process within the EHRI Annotator.

Upon entering the website, the user is prompted to input some text to a text box and click on a *"Process Text"* button which triggers the NER component of the application. As mentioned in the previous sections, for its NER component the EHRI Annotator currently relies on the EHRI-NER model, a detailed description of which can be found in Dermentzi and Scheithauer (2024). This model was developed by fine-tuning XLM-RoBERTa-Large for NER of six entity types, namely *PERSON*, *ORGANIZATION*, *LOCATION*, *CAMP*, *GHETTO*, and *DATE*. The NER backend splits the text into smaller chunks (if the text is too long for the model's context window to handle) and runs the model to return detected entities which are then displayed highlighted in the document. The user can then review each prediction to accept, reject or adjust its boundaries or label through an editing panel. For each entity detected (apart from *DATE* entities, where this is not applicable), the user can trigger the entity linking process by clicking on the *Link* button, which sends the query to the linking microservice and returns a ranked list of the top five candidates exceeding a certain threshold. The user can then select the correct match or perform a manual search against the index if no suitable candidates were returned. Any edits can be propagated to all mentions of the same entity with the same spelling within the same input text. Once all entities have been verified and linked, the user can export the annotated document as a TEI P5 XML file to process further with a text editor and prepare for publication. For linked entities with geographic coordinates (these could be *CAMP*, *GHETTO*, or *LOCATION* entities), linked places are additionally displayed on an interactive map.

Once the EL process of an entity is triggered by the user, the retrieval strategy depends on the entity's assigned label. This filtering strategy was inspired by what Arora and Dell (2024) describe as *"blocking"* in their paper and is used to make the tool as fast and efficient as possible. Therefore, *PERSON* entities get matched against the *EHRI Personalities* authority set, *ORGANIZATION* entities get matched against the *EHRI Corporate Bodies* authority set, *CAMP* entities against the *EHRI Camps* vocabulary, *GHETTO* entities against the *EHRI Ghettos* vocabulary and *LOCATION* entities against *GeoNames*.

Specifically, the linking component of the EHRI Annotator is supported by two Qdrant *"collections"*. The first collection concerns entities from the EHRI vocabularies and authority sets, which are indexed for semantic matching. These entities are encoded with LaBSE (Feng et al., 2022) into 768-dimensional vectors for approximate nearest neighbor search. Candidate retrieval returns the top 2,000 entries by cosine similarity, which are then re-scored using a function that combines semantic similarity, string similarity (a multi-stage scoring function from exact match through to fuzzy matching), context-aware scoring that incorporates the text of one neighboring entity to the left and one to the right of the target mention to aid disambiguation (e.g., the presence of the entity "Riga" in the context near the entity "Gestapo" helps rank "Gestapo Riga" higher in the candidate list compared to other Gestapo-related entities in the relevant EHRI authority set), and a multi-alias boost that rewards entities with multiple matching name variants. Exact string matches receive a much higher score to guarantee first-rank placement. Character-level fuzzy matching using the `rapidfuzz` library (Bachmann, 2025) with a similarity threshold of 0.7 is applied during candidate re-ranking for EHRI entities to handle minor spelling variations and Optical Character Recognition (OCR) errors common in archival texts. This multi-stage linking strategy was refined through extensive trial and error since there was no domain-specific dataset available for training a supervised ranking model at the time of building the EHRI Annotator. Semantic matching is

[11] GitHub repository for Qdrant vector search engine.

essential for matching against EHRI vocabularies, which have limited multilingual coverage. For example, matching *"malou pevnost"*, meaning small fortress in Czech, to the German equivalent *"Kleine Festung"* in the EHRI Camps dataset is only possible through semantic matching because the EHRI Camps vocabulary only lists the German name.

The second Qdrant collection concerns the GeoNames gazetteer, which contains over 13 million entries in total. Embedding every entry and its alias names would be prohibitively expensive in terms of storage and inference time. Instead, taking into account the conclusion reached by Arora et al. (2024) that toponym disambiguation against a comprehensive gazetteer can be highly accurate using non-neural methods, location entities are matched using text-based retrieval only. To enable this within Qdrant, which requires a vector for every entry, GeoNames entries are indexed with placeholder *"dummy"* zero vectors and rely exclusively on Qdrant's built-in text index of each entry's multilingual alias array for candidate retrieval. Moreover, we do not index the entire GeoNames dataset but rather filter it down to approximately 7.8 million entries based on a curated set of 102 GeoNames feature codes (e.g., populated places, historical sites, camps, railway stations, religious sites. The full list is provided in Appendix A.). Again, the selection of these feature codes was refined through iterative trial and error. When a good location match is not returned although it exists within the GeoNames dataset, the author examines whether a new feature code should be considered. Another concern was how to handle morphological variation in highly inflected languages. The solution to this was to lemmatize queries before matching using `Stanza` (Qi et al., 2020) with models trained on Universal Dependencies (Zeman et al., 2023) for Czech (cs), Polish (pl), Slovak (sk), German (de), Hungarian (hu), Russian (ru), Ukrainian (uk), Lithuanian (lt), Belarusian (be), Greek (el), and Hebrew (he). Candidate locations are ranked using a relevance score pre-computed during indexing that combines feature code importance weighted by domain relevance (e.g., camps and historical sites are weighted higher than generic buildings), country priority weights reflecting Holocaust and WWII geography (e.g., Poland, Germany, and Austria receive the highest weights), and a population factor. The final ranking for each candidate is determined by the product of this relevance weight and a text match quality score derived from comparing the query against the entry's primary name and aliases. The full feature code and country weight configuration as it currently stands is detailed in Appendix A.

The system architecture is designed to retrieve plausible candidates efficiently while leaving the final disambiguation decision up to the domain expert annotating the document. Verified annotations are exported as TEI P5 XML with `ref` attributes pointing to EHRI Portal Unique Resource Identifiers (URIs) or GeoNames URIs, pre-annotated for further processing as part of the Online Editions publication pipeline.

## 4. Evaluation

A preliminary evaluation of the EHRI Annotator has taken place both qualitatively through user testing sessions with Holocaust researchers who were asked to fill in feedback forms but also quantitatively through an automatic evaluation of entity linking accuracy against a gold-standard dataset which is still under active curation.

### 4.1. Dataset

The evaluation dataset was produced during the first EHRI–CLARIN Datathon, held on 26-27 February 2025 in Budapest, Hungary, co-organized by EHRI, CLARIN, ELTE University Research Center for Computational Social Science, and the Leibniz Institute for the History and Culture of Eastern Europe. During the event, participants annotated Holocaust testimonies using the INCEpTION platform (Klie et al., 2018). The source documents were provided by the Hungarian Jewish Museum and Archives and the Wiener Holocaust Library. *PERSON*, *ORGANIZATION*, *CAMP*, and *GHETTO* entities were linked to EHRI vocabularies and authority sets, while *LOCATION* entities were linked to Wikidata. At the time of writing, 40 (32 English, 7 German, 1 Hungarian) of the 140 documents processed during the event have been curated to ensure correct selection of URIs and consistent application of the annotation guidelines shared during the event. The full dataset and a detailed description of the annotation process will be part of a future publication once the curation process has been fully completed.

### 4.2. Entity Linking Evaluation

To evaluate entity linking, Wikidata identifiers for *LOCATION* entities were mapped to GeoNames via Wikidata property P1566. There were 30 location entities annotated with Wikidata identifiers that lacked a GeoNames mapping via property P1566. For non-location entities, EHRI vocabulary identifiers were compared directly. Entities annotated only with Wikidata identifiers for which no corresponding EHRI or GeoNames mapping exists were excluded. Entities were deduplicated by spelling, entity type, and document language, since the main evaluation (see Table 1 below) queries the linking service without document context and identical

queries produce identical candidate rankings. After deduplication and exclusions, 264 entities were retained for evaluation. Accuracy@1 (correct entity ranked first), Accuracy@5 (correct entity in top five), and Mean Reciprocal Rank (MRR) over the top 10 retrieved candidates are presented in Table 1 below. The results in Table 1 aggregate entities from all curated documents across all three languages.

Table 1: Entity linking performance by type.

| Type | N | Acc@1 | Acc@5 | MRR |
|---|---|---|---|---|
| LOCATION | 156 | 42.3% | 73.7% | 0.556 |
| PERSON | 30 | 76.7% | 100.0% | 0.883 |
| ORGANIZATION | 45 | 62.2% | 71.1% | 0.654 |
| CAMP | 23 | 78.3% | 87.0% | 0.811 |
| GHETTO | 10 | 80.0% | 80.0% | 0.800 |
| **Overall** | **264** | **54.2%** | **77.7%** | **0.641** |

The overall Accuracy@5 of 77.7% means that the correct knowledge base entry appears among the top five candidates in most cases. Since the EHRI Annotator is taking a human-in-the-loop approach in its design, this metric is useful because it shows that the system retrieves a good set of top five candidates for the domain expert to choose from. *LOCATION* entities show the lowest Accuracy@1 (42.3%) while Accuracy@5 remains reasonably high at 73.7% given that there are many locations with the same name which makes the ranking of the results harder. It is worth noting that the system returned no candidate matches for 23 out of 156 *LOCATION*-type entities under evaluation. These entities generally concern poor OCR or spelling mistakes (e.g., *"Heidelterg"* for Heidelberg, *"Theresienstad"* for Theresienstadt, *"Shtirotava"* for Šķirotava). *ORGANIZATION* entities show the lowest Accuracy@5 (71.1%).

An ablation study was conducted on all ***non-location*** entities in the evaluation dataset (N=108) to estimate the contribution of the context-aware scoring described in Section 3. For each entity, two requests were sent to the linking service: one with context derived from one neighboring entity to the left and one neighboring entity to the right (matching the deployed service's behavior); and one request without context. Context-aware scoring improved Accuracy@1 from 71.3% to 74.1% and Accuracy@5 from 83.3% to 84.3%, with MRR increasing from 0.764 to 0.782.

However, this quantitative evaluation of the tool is preliminary while the gold standard dataset is under curation. In particular, results for *PERSON* (N=30), *CAMP* (N=23), and *GHETTO* (N=10) entities should be interpreted with caution given the small sample sizes. Nevertheless, this small-scale quantitative evaluation, taken into account together with the user feedback expressed under the qualitative evaluation described below, shows that this tool can already be useful in supporting Online Edition Editors to annotate new documents.

### 4.3. User Evaluation

The tool was evaluated by 11 users after two hands-on testing sessions, a workshop organized by EHRI-CZ in Prague and an EHRI webinar. Participants tested the tool on texts in English, German, Czech, Slovak, and Italian. They were then asked to complete a feedback form covering overall usability, quality of NER and EL predictions, shortcomings, and things to improve. All 11 participants rated the overall experience as either *Excellent* (8) or *Good* (3). Eight participants found the interface *Very easy* to use and three rated it *Mostly easy*; 10 of 11 described the layout as clean and easy to navigate.

NER accuracy was rated *Very accurate* by six participants and *Mostly accurate* by five. Participants noted several recurring NER errors, namely nationality adjectives misclassified as locations (e.g., *"British"*), incomplete entity boundaries requiring manual correction (e.g., the text reads *"of the Swedish Red Cross"* but the model detects *"Swedish"* as a separate *LOCATION* entity and *Red Cross* as its own *ORGANIZATION* entity. This is a problem because if we try to link *"Red Cross"* directly, we get presented with the wrong match. The correct entity here would be Swedish Red Cross as one *ORGANIZATION* entity.), and difficulty with retrieving matches for misspelled names (e.g., *"Krakoff"* for *"Kraków"*). One participant noted that morphological inflection in certain languages caused recognition failures.

Entity linking quality was rated *Very good* (correct match almost always ranked first) by eight participants and *Mostly good* (correct match usually in the first few candidates) by three, verifying qualitatively the results of the quantitative evaluation (albeit limited). When it came to more constructive feedback, participants requested expansion of the knowledge base to include vocabularies that cover additional historical entities such as the Slovak State or the Protectorate. Additionally, participants requested additional export formats including CSV, JSON, and spreadsheets. Editing features were generally well-received, though boundary editing was rated the most difficult task (*Easy*: 5, *OK*: 6), suggesting room for interface improvement. Six participants exported TEI XML; of these, four rated the export quality as *Excellent* and two as *Good - Needed minor tweaks*.

## 5. Discussion and Conclusion

The EHRI Annotator is a work in progress requiring continuous refinement. The evaluation presented here is preliminary, the dataset covers only 40 of the 140 documents from the datathon and sample sizes for several entity types are small. Nevertheless, several limitations of the current system emerged from both the quantitative and the qualitative evaluation. Disambiguating *LOCATION* entities is very challenging when dealing with such a large gazetteer like GeoNames. The system fails to retrieve candidates for non-standard spellings, while embedding this resource to enable semantic matching is prohibitive in terms of resources needed. The main limitation of the current EL approach when it comes to *LOCATION* entities is that the way Qdrant has been set up for the GeoNames collection (zero vectors and reliance on text-based retrieval) limits candidate retrieval to exact token matching, meaning that even small variations in spelling or OCR errors can lead to zero candidates even before any ranking can take place. While Qdrant remains a very practical solution in terms of keeping the infrastructure as simple and easy to maintain as possible given that we are already using it for the vector database, migrating to a search engine with rich full text search features such as Elasticsearch could help address these retrieval failures.

Another observation that can be made is that common errors of the EHRI-NER model that have been documented in previous work (Dermentzi and Scheithauer, 2024) can spill over to the linking stage as noted in the user evaluation reported here. It was also noted that the EHRI vocabularies themselves are not as comprehensive as users would like them to be, with unlinked entities often reflecting gaps in the authority sets rather than system failures. User feedback expressed the need to expand vocabulary coverage to include additional historical entities.

Given the sensitivity of Holocaust materials, all models used by the EHRI Annotator are self-hosted on a dedicated server in the European Union (EU), with no data transferred to third-party APIs. Original texts are discarded after NER inference and only entity-level data appears in system logs. This feature is essential for working with archival institutions which are often bound by strict data protection and ethical obligations regarding the materials in their custody. However, it also imposes practical constraints. Since galleries, libraries, archives, and museum (GLAM) institutions typically lack the infrastructure and resources to deploy and maintain computationally intensive models, this limits the adoption of state-of-the-art (SOTA) approaches that rely on large language models (LLMs) or cloud-based third-party API services. The architecture described in this paper was designed with these constraints in mind, favoring lightweight models that can be self-hosted over more powerful but externally controlled alternatives.

Current development priorities include support for adding new entities directly through the interface (currently possible only by editing the TEI XML output), linking to the EHRI Terms vocabulary which could be combined with previous work on automated subject indexing (Dermentzi et al., 2025), and additional import and export formats to support use cases beyond the EHRI Online Editions, such as archival metadata enrichment and geographic visualization. Longer-term goals include allowing users to connect custom vocabularies for domain-specific linking.

Another area of future work is experimenting with other NER and EL approaches. In Section 3, we described the three-component architecture of the EHRI Annotator. While the current system is based on the EHRI-NER model, the modularity of the service allows for experimentation with alternative models and techniques. Comparative evaluation against such approaches is planned for future work once the full evaluation dataset is available. For example, testing how open-source LLMs or other architectures like GLiNER (Zaratiana et al., 2024) perform against the EHRI-NER model can inform the choice of model for the NER and EL components before the tool moves from prototype to production. The fully curated dataset could also be used to train custom ranking models.

In conclusion, the EHRI Annotator demonstrates that a hybrid entity linking architecture which combines semantic retrieval for smaller vocabularies with text-based retrieval and domain-specific weighting for large gazetteers can provide effective candidate retrieval in a human-in-the-loop setting even when dealing with very large knowledge bases, as long as the larger knowledge bases are comprehensive enough in terms of alternative names and translations of the entries. The tool is currently deployed as a working prototype but it has already received positive reception. The hope is that it contributes to making Holocaust-related archival material more accessible, interoperable, and discoverable across institutions and languages.

## 6. Acknowledgements

The work described herein has also been supported by the Ministry of Education, Youth and Sports of the Czech Republic, Project No. LM2023062 LINDAT/CLARIAH-CZ.

## 7. Bibliographical References

Abhishek Arora and Melissa Dell. 2024. LinkTransformer: A Unified Package for Record Linkage with Transformer Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 221–231, Bangkok, Thailand. Association for Computational Linguistics.

Abhishek Arora, Emily Silcock, Melissa Dell, and Leander Heldring. 2024. Contrastive Entity Coreference and Disambiguation for Historical Texts. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6174–6186, Miami, Florida, USA. Association for Computational Linguistics.

Max Bachmann. 2025. rapidfuzz/rapidfuzz: Release 3.13.0.

Maria Dermentzi, Mike Bryant, Fabio Rovigo, and Herminio García-González. 2025. Multilingual Automated Subject Indexing: a comparative study of LLMs vs alternative approaches in the context of the EHRI project. *DH Benelux Journal*, 7(Breaking Silos, Connecting Data: Advancing Integration and Collaboration in Digital Humanities).

Maria Dermentzi and Hugo Scheithauer. 2024. Repurposing Holocaust-Related Digital Scholarly Editions to Develop Multilingual Domain-Specific Named Entity Recognition Tools. In *Proceedings of the First Workshop on Holocaust Testimonies as Language Resources (HTRes) @ LREC-COLING 2024*, pages 18–28, Torino, Italia. ELRA and ICCL.

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2023. Named Entity Recognition and Classification in Historical Documents: A Survey. *ACM Comput. Surv.*, 56(2):27:1–27:47.

Maud Ehrmann, Matteo Romanello, Antoine Doucet, and Simon Clematide. 2022a. Introducing the HIPE 2022 Shared Task: Named Entity Recognition and Linking in Multilingual Historical Documents. In *Advances in Information Retrieval*, pages 347–354, Cham. Springer International Publishing.

Maud Ehrmann, Matteo Romanello, Alex Flückiger, and Simon Clematide. 2020. Overview of CLEF HIPE 2020: Named Entity Recognition and Linking on Historical Newspapers. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 288–310, Cham. Springer International Publishing.

Maud Ehrmann, Matteo Romanello, Sven Najem-Meyer, Antoine Doucet, and Simon Clematide. 2022b. Overview of HIPE-2022: Named Entity Recognition and Linking in Multilingual Historical Documents. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 423–446, Cham. Springer International Publishing.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT Sentence Embedding. ArXiv:2007.01852.

Herminio García-González and Mike Bryant. 2023. The Holocaust Archival Material Knowledge Graph. In Terry R. Payne, Valentina Presutti, Guilin Qi, María Poveda-Villalón, Giorgos Stoilos, Laura Hollink, Zoi Kaoudi, Gong Cheng, and Juanzi Li, editors, *The Semantic Web – ISWC 2023*, volume 14266, pages 362–379. Springer Nature Switzerland, Cham. Series Title: Lecture Notes in Computer Science.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics. Place: Santa Fe, USA.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. ArXiv:2003.07082 [cs].

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J. G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A. C. 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun

Zhao, and Barend Mons. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1):160018.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2024. GLiNER: Generalist Model for Named Entity Recognition using Bidirectional Transformer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5364–5376, Mexico City, Mexico. Association for Computational Linguistics.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Hamid Aghaei, Željko Agić, Amir Ahmadi, Lars Ahrenberg, Chika Kennedy Ajede, Salih Furkan Akkurt, Gabrielė Aleksandravičiūtė, Ika Alfina, Avner Algom, Khalid Alnajjar, Chiara Alzetta, Erik Andersen, Lene Antonsen, Tatsuya Aoyama, Katya Aplonova, Angelina Aquino, Carolina Aragon, Glyd Aranes, Maria Jesus Aranzabe, Bilge Nas Arıcan, Hórunn Arnardóttir, Gashaw Arutie, Jessica Naraiswari Arwidarasti, Masayuki Asahara, Katla Ásgeirsdóttir, Deniz Baran Aslan, Cengiz Asmazoğlu, Luma Ateyah, Furkan Atmaca, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Mariana Avelãs, Elena Badmaeva, Keerthana Balasubramani, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, Starkaður Barkarson, Rodolfo Basile, Victoria Basmov, Colin Batchelor, John Bauer, Seyyit Talha Bedir, Shabnam Behzad, Kepa Bengoetxea, İbrahim Benli, Yifat Ben Moshe, Gözde Berk, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agnė Bielinskienė, Kristín Bjarnadóttir, Rogier Blokland, Victoria Bobicev, Loïc Boizou, Emanuel Borges Völker, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Anouck Braggaar, António Branco, Kristina Brokaitė, Aljoscha Burchardt, Marisa Campos, Marie Candito, Bernard Caron, Gauthier Caron, Catarina Carvalheiro, Rita Carvalho, Lauren Cassidy, Maria Clara Castro, Sérgio Castro, Tatiana Cavalcanti, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomír Čéplö, Neslihan Cesur, Savas Cetin, Özlem Çetinoğlu, Fabricio Chalub, Liyanage Chamila, Shweta Chauhan, Ethan Chi, Taishi Chika, Yongseok Cho, Jinho Choi, Jayeol Chun, Juyeon Chung, Alessandra T. Cignarella, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Daniela Corbetta, Francisco Costa, Marine Courtin, Mihaela Cristescu, Ingerid Løyning Dale, Philemon Daniel, Elizabeth Davidson, Leonel Figueiredo de Alencar, Mathieu Dehouck, Martina de Laurentiis, Marie-Catherine de Marneffe, Valeria de Paiva, Mehmet Oguz Derin, Elvis de Souza, Arantza Diaz de Ilarraza, Carly Dickerson, Arawinda Dinakaramani, Elisa Di Nuovo, Bamba Dione, Peter Dirix, Kaja Dobrovoljc, Adrian Doyle, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Christian Ebert, Hanne Eckhoff, Masaki Eguchi, Sandra Eiche, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Olga Erina, Tomaž Erjavec, Farah Essaidi, Aline Etienne, Wograine Evelyn, Sidney Facundes, Richárd Farkas, Federica Favero, Jannatul Ferdaousi, Marília Fernanda, Hector Fernandez Alcalde, Amal Fethi, Jennifer Foster, Cláudia Freitas, Kazunori Fujita, Katarína Gajdošová, Daniel Galbraith, Federica Gamba, Marcos Garcia, Moa Gärdenfors, Fabrício Ferraz Gerardi, Kim Gerdes, Luke Gessler, Filip Ginter, Gustavo Godoy, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta González Saavedra, Bernadeta Griciūtė, Matias Grioni, Loïc Grobol, Normunds Grūzītis, Bruno Guillaume, Céline Guillot-Barbance, Tunga Güngör, Nizar Habash, Hinrik Hafsteinsson, Jan Hajič, Jan Hajič jr., Mika Hämäläinen, Linh Hà Mỹ, Na-Rae Han, Muhammad Yudistira Hanifmuti, Takahiro Harada, Sam Hardwick, Kim Harris, Dag Haug, Johannes Heinecke, Oliver Hellwig, Felix Hennig, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Marivel Huerta Mendez, Jena Hwang, Takumi Ikeda, Anton Karl Ingason, Radu Ion, Elena Irimia, Ọlájídé Ishola, Artan Islamaj, Kaoru Ito, Siratun Jannat, Tomáš Jelínek, Apoorva Jha, Katharine Jiang, Anders Johannsen, Hildur Jónsdóttir, Fredrik Jørgensen, Markus Juutinen, Hüner Kaşıkara, Nadezhda Kabaeva, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Neslihan Kara, Ritván Karahóǧa, Andre Kåsen, Tolga Kayadelen, Sarveswaran Kengatharaiyer, Václava Kettnerová, Jesse Kirchner, Elena Klementieva, Elena Klyachko, Arne Köhn, Abdullatif Köksal, Kamil Kopacewicz, Timo Korkiakangas, Mehmet Köse, Alexey Koshevoy, Natalia Kotsyba, Jolanta Kovalevskaitė, Simon Krek, Parameswari Krishnamurthy, Sandra Kübler, Adrian Kuqi, Oğuzhan Kuyrukçu, Aslı Kuzgun, Sookyoung Kwak, Kris Kyle, Veronika Laippala, Lorenzo Lambertino, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phng Lê Hồng, Alessandro Lenci, Saran Lertpradit, Herman Leung, Maria Levina, Lauren Levine, Cheuk Ying Li, Josie Li, Keying Li, Yixuan Li, Yuan Li, KyungTae Lim, Bruna Lima Padovani, Yi-Ju Jessica Lin, Krister Lindén, Yang Janet Liu, Nikola Ljubešić, Olga Loginova, Stefano Lusito, Andry Luthfi, Mikko Luukko, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Menel Mahamdi, Jean Maillard, Ilya Makarchuk, Aibek Makazhanov, Michael Mandl,

Christopher Manning, Ruli Manurung, Büşra Marşan, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Stella Markantonatou, Héctor Martínez Alonso, Lorena Martín Rodríguez, André Martins, Cláudia Martins, Jan Mašek, Hiroshi Matsuda, Yuji Matsumoto, Alessandro Mazzei, Ryan McDonald, Sarah McGuinness, Gustavo Mendonça, Tatiana Merzhevich, Niko Miekka, Aaron Miller, Karina Mischenkova, Anna Missilä, Cătălin Mititelu, Maria Mitrofan, Yusuke Miyao, AmirHossein Mojiri Foroushani, Judit Molnár, Amirsaeid Moloodi, Simonetta Montemagni, Amir More, Laura Moreno Romero, Giovanni Moretti, Shinsuke Mori, Tomohiko Morioka, Shigeki Moro, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Robert Munro, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Mariam Nakhlé, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Manuela Nevaci, Lng Nguyễn Thị, Huyền Nguyễn Thị Minh, Yoshihiro Nikaido, Vitaly Nikolaev, Rattima Nitisaroj, Alireza Nourian, Hanna Nurmi, Stina Ojala, Atul Kr. Ojha, Hulda Óladóttir, Adédayọ̀ Olúòkun, Mai Omura, Emeka Onwuegbuzia, Noam Ordan, Petya Osenova, Robert Östling, Lilja Øvrelid, Şaziye Betül Özateş, Merve Özçelik, Arzucan Özgür, Balkız Öztürk Başaran, Teresa Paccosi, Alessio Palmero Aprosio, Anastasia Panova, Hyunji Hayley Park, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Giulia Pedonese, Angelika Peljak-Łapińska, Siyao Peng, Siyao Logan Peng, Rita Pereira, Sílvia Pereira, Cenel-Augusto Perez, Natalia Perkova, Guy Perrier, Slav Petrov, Daria Petrova, Andrea Peverelli, Jason Phelan, Jussi Piitulainen, Yuval Pinter, Clara Pinto, Tommi A Pirinen, Emily Pitler, Magdalena Plamada, Barbara Plank, Thierry Poibeau, Larisa Ponomareva, Martin Popel, Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Robert Pugh, Tiina Puolakainen, Sampo Pyysalo, Peng Qi, Andreia Querido, Andriela Rääbis, Alexandre Rademaker, Mizanur Rahoman, Taraka Rama, Loganathan Ramasamy, Joana Ramos, Fam Rashel, Mohammad Sadegh Rasooli, Vinit Ravishankar, Livy Real, Petru Rebeja, Siva Reddy, Mathilde Regnault, Georg Rehm, Arij Riabi, Ivan Riabov, Michael Rießler, Erika Rimkutė, Larissa Rinaldi, Laura Rituma, Putri Rizqiyah, Luisa Rocha, Eiríkur Rögnvaldsson, Ivan Roksandic, Mykhailo Romanenko, Rudolf Rosa, Valentin Roșca, Davide Rovati, Ben Rozonoyer, Olga Rudina, Jack Rueter, Kristján Rúnarsson, Shoval Sadde, Pegah Safari, Aleksi Sahala, Shadi Saleh, Alessio Salomoni, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Ezgi Sanıyar, Dage Särg, Marta Sartor, Mitsuya Sasaki, Baiba Saulīte, Yanin Sawanakunanon, Shefali Saxena, Kevin Scannell, Salvatore Scarlata, Nathan Schneider, Sebastian Schuster, Lane Schwartz, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Syeda Shahzadi, Mo Shen, Atsuko Shimada, Hiroyuki Shirasu, Yana Shishkina, Muh Shohibussirri, Maria Shvedova, Janine Siewert, Einar Freyr Sigurðsson, João Silva, Aline Silveira, Natalia Silveira, Sara Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Haukur Barri Símonarson, Kiril Simov, Dmitri Sitchinava, Ted Sither, Maria Skachedubova, Aaron Smith, Isabela Soares-Bastos, Per Erik Solberg, Barbara Sonnenhauser, Shafi Sourov, Rachele Sprugnoli, Vivian Stamou, Steinþór Steingrímsson, Antonio Stella, Abishek Stephen, Milan Straka, Emmett Strickland, Jana Strnadová, Alane Suhr, Yogi Lesmana Sulestio, Umut Sulubacak, Shingo Suzuki, Daniel Swanson, Zsolt Szántó, Chihiro Taguchi, Dima Taji, Fabio Tamburini, Mary Ann C. Tan, Takaaki Tanaka, Dipta Tanaya, Mirko Tavoni, Samson Tella, Isabelle Tellier, Marinella Testori, Guillaume Thomas, Sara Tonelli, Liisi Torga, Marsida Toska, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Utku Türk, Francis Tyers, Sveinbjörn Þórðarson, Vilhjálmur Þorsteinsson, Sumire Uematsu, Roman Untilov, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Andrius Utka, Elena Vagnoni, Sowmya Vajjala, Socrates Vak, Rob van der Goot, Martine Vanhove, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Uliana Vedenina, Giulia Venturi, Veronika Vincze, Natalia Vlasova, Aya Wakasa, Joel C. Wallenberg, Lars Wallin, Abigail Walsh, Jonathan North Washington, Maximilan Wendt, Paul Widmer, Shira Wigderson, Sri Hartati Wijono, Seyi Williams, Mats Wirén, Christian Wittern, Tsegay Woldemariam, Taksum Wong, Alina Wróblewska, Mary Yako, Kayo Yamashita, Naoki Yamazaki, Chunxiao Yan, Koichi Yasuoka, Marat M. Yavrumyan, Arife Betül Yenice, Olcay Taner Yıldız, Zhuoran Yu, Arlisa Yuliawati, Zdeněk Žabokrtský, Shorouq Zahra, Amir Zeldes, He Zhou, Hanzhi Zhu, Yilun Zhu, Anna Zhuravleva, and Rayan Ziane. 2023. Universal dependencies 2.12. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).

## A. GeoNames Index Configuration

The GeoNames gazetteer (13.4 million entries) was filtered to 7.8 million entries by retaining only locations the feature code of which is included in the list below. Each indexed entry receives a precomputed relevance score combining a feature

code importance weight (Table 2), a country priority weight (Table 3), and a population factor.

Table 2: The 102 GeoNames feature codes retained for indexing, sorted alphabetically, with importance weights (W). Codes not listed are excluded from indexing.

| Code | W | Code | W | Code | W |
|---|---|---|---|---|---|
| ADM1 | 0.8 | FRST | 0.6 | PPLA5 | 0.6 |
| ADM1H | 0.8 | FT | 0.6 | PPLC | 0.9 |
| ADM2 | 0.7 | GRVE | 0.7 | PPLCH | 0.9 |
| ADM2H | 0.7 | HBR | 0.6 | PPLF | 0.6 |
| ADM3 | 0.6 | HSP | 0.6 | PPLH | 0.6 |
| ADM3H | 0.6 | HSPD | 0.5 | PPLL | 0.6 |
| ADM4 | 0.5 | HSTS | 0.8 | PPLQ | 0.7 |
| ADM4H | 0.5 | HTL | 0.5 | PPLS | 0.6 |
| ADM5 | 0.4 | INDS | 0.5 | PPLW | 0.7 |
| ADM5H | 0.4 | INSM | 0.6 | PPLX | 0.6 |
| ADMD | 0.3 | ISL | 0.8 | PRN | 0.7 |
| ADMDH | 0.3 | ISLS | 0.7 | QUAY | 0.6 |
| AIRB | 0.6 | LIBR | 0.6 | RGN | 0.9 |
| AIRQ | 0.5 | MFG | 0.5 | RGNE | 0.9 |
| BAY | 0.5 | MILB | 0.6 | RGNH | 0.9 |
| BAYS | 0.5 | MKT | 0.6 | RR | 0.5 |
| BDG | 0.5 | ML | 0.5 | RSTN | 0.6 |
| BNK | 0.6 | MN | 0.6 | RSTP | 0.6 |
| BRKS | 0.7 | MNMT | 0.7 | RVN | 0.8 |
| BTL | 0.7 | MSTY | 0.6 | SCH | 0.6 |
| CAVE | 0.6 | MUS | 0.6 | SCHC | 0.6 |
| CH | 0.6 | NVB | 0.6 | SEA | 0.7 |
| CMP | 0.9 | PCL | 1.0 | SQR | 0.7 |
| CMPLA | 0.9 | PCLD | 0.9 | STNB | 0.6 |
| CMPQ | 0.8 | PCLF | 0.9 | STNR | 0.6 |
| CMPRF | 0.7 | PCLH | 1.0 | STRT | 0.6 |
| CMTY | 0.7 | PCLI | 1.0 | SYG | 0.8 |
| CSTL | 0.6 | PCLS | 0.9 | THTR | 0.6 |
| CVNT | 0.6 | PIER | 0.6 | TMB | 0.7 |
| DCK | 0.6 | PPL | 0.6 | TMPL | 0.7 |
| DCKB | 0.6 | PPLA | 0.8 | TNL | 0.5 |
| DIP | 0.6 | PPLA2 | 0.7 | UNIV | 0.6 |
| EST | 0.5 | PPLA3 | 0.7 | WHRF | 0.6 |
| FRM | 0.5 | PPLA4 | 0.6 | WRCK | 0.5 |

Table 3: Country priority weights. All countries not listed receive a default weight of 0.3.

| Weight | Countries |
|---|---|
| 1.0 | DE, PL, CZ, SK, AT, HU, NL, BE, FR |
| 0.9 | GB, IT, RO, BG, HR, GR, IL, PS |
| 0.8 | RU, UA, BY, LT, LV, EE, LU, MT, YU, CS |
| 0.7 | NO, DK, SE, FI, CH, ES, PT, TR, RS, AL, SI, BA, CY, GI, US |
| 0.6 | CA, AU, NZ, ZA, BR, AR, CU, MA, DZ, TN, SY, LB, IQ, IE, ME, MK, MD, XK, HK |
| 0.5 | JP, CN, PH, ID, SG, MM, TH, VN, MY, KR, TW, LY, EG, IR, MX, UY, CL, BO, DO, IS, IN |
| 0.3 | All other countries (default) |