



**European Holocaust Research Infrastructure  
H2020-INFRAIA-2019-1  
GA no. 871111**

**Deliverable 11.6**

**Report on Standards**

**Hugo Scheithauer**  
Research team ALMAAnCH, Inria, Paris

**Sarah Bénérière**  
Research team ALMAAnCH, Inria, Paris

**Maria Dermentzi**  
King's College, London

**Floriane Chiffolleau**  
Research team ALMAAnCH, Inria, Paris

**Alix Chagué**  
Research team ALMAAnCH, Inria, Paris

**Laurent Romary**  
Inria, Directorate for Scientific Information and Culture, Paris

**Start: September 2020 [M1]**

**Due: June 2024 [M46]**

**Actual: July 2024 [M47]**



EHRI is funded by the European Union

## Document Information

Project URL	<a href="http://www.ehri-project.eu">www.ehri-project.eu</a>
Document URL	<a href="https://www.ehri-project.eu/deliverables-ehri-3-2020-2024">https://www.ehri-project.eu/deliverables-ehri-3-2020-2024</a>
Deliverable	D11.6 Report on Standards
Work Package	WP11
Lead Beneficiary	Inria
Relevant Milestones	MS4
Dissemination level	PU
Contact Person	Laurent Romary (Inria)
Abstract (for dissemination)	This report presents the work carried out within task 11.4 (standards) on the definition of a standards-based workflow for the creation, management and dissemination of the EHRI digital edition. In particular, it identifies the subset of the TEI guidelines that is appropriate for integrating the output of an HTR process within a TEI document, transforming it as a generic scholarly digital edition and publishing it within a standard platform such as TEI Publisher. The work described here paves the way for a stable workflow to be deployed within the future EHRI ERIC.
Management Summary	n/a

## Table of Contents

1.	5	
1.1.	5	
1.2.	5	
2.	6	
2.1.	6	
2.2.	6	
3.	6	
3.1.	7	
3.2.	7	
3.3.	7	
3.4.	8	
4.	9	
4.1.	9	
4.2.	10	
4.2.1.	10	
4.2.2.	11	
4.2.3.	12	
4.2.4.	14	
4.2.5.	Discussion	20
5.	20	
5.1.	20	
5.2.	24	
5.3.	25	
5.3.1.	Introduction	25
5.3.2.	Related Work	26
5.3.3.	EHRI Online Editions	28
5.3.4.	The EHRI-NER Dataset	28
5.3.5.	Experimental Setup	31
5.3.6.	Conclusion	37
6.	38	
6.1.	38	
6.2.	38	
Conclusion		42
Bibliographical References		44
Appendix		49
BeGrenzte Flucht Edition		49

Early Holocaust Testimonies Edition	49
Diplomatic Reports Edition	49
Von Wien ins Nirgendwo: Die Nisko-Deportationen 1939 Edition	49
Documentation Campaign Edition	49
Uzavřít Hranice Edition	50

## 1. Introduction

### 1.1. What is an EHRI digital edition?

“EHRI [European Holocaust Research Infrastructure] Online Editions consist of annotated digitized documents from various sources gathered around a theme, and are a new way of presenting digital archival content.”<sup>1</sup> They gather digitized documents kept in various archives, which are edited and annotated by EHRI researchers using the Text Encoding Initiative (TEI) P5 standard, an XML vocabulary and schema dedicated to editing digital texts, in order to publish them online. They also aim at giving contextual information for the edited documents by linking them to EHRI vocabularies, descriptions and by using interactive maps. They are fully searchable and can also be filtered using specific thematic or spatial facets thanks to their encoding in TEI.<sup>2</sup> All documents within an edition have a transcription, in its original language, a translation, or both, and provide a link to the corresponding digital facsimile. They are published without following a consistent schedule. A scholarly discussion concerning the theme of the edition is also included alongside the documents.

At the time of writing this report, the EHRI Consortium has supported the development and publication of six<sup>3</sup> Holocaust-related digital scholarly editions. EHRI began releasing curated documents during its second phase (EHRI-2 2015-2019).

### 1.2. Creating an EHRI Online Edition with the EHRI digitization pipeline: processes and digital standards

Digital editions act as a valuable bridge between archival institutions of various sizes and researchers, allowing them to curate the documents they preserve and bringing them together around various themes.

The EHRI Consortium requires a robust infrastructure to support the publication of edited documents online. It must also accommodate the storage of these documents and adhere to standards that ensure they are FAIR (findable, accessible, interoperable, reusable)<sup>4</sup>. Standards are implemented at two levels within EHRI’s processes:

- **Edition:** The content and metadata of the documents are managed using TEI XML, with metadata expressed in Dublin Core.
- **Access:** Edited archival materials are accessible directly from their home institutions through finding aids written in Encoded Archival Description (EAD)<sup>5</sup> and accessible through the EHRI portal.

The EHRI digitization pipeline is centered around (1) the annotation of digital documents, (2) their conversion into TEI XML using the open-source software Odette<sup>6</sup>, and (3) their publication

---

<sup>1</sup> See EHRI Online Editions webpage: <https://www.ehri-project.eu/ehri-online-editions>. Accessed June 2024.

<sup>2</sup> See EHRI vocabularies: <https://portal.ehri-project.eu/vocabularies>. Accessed June 2024.

<sup>3</sup> As of July 2024, a seventh edition is about to be published: *The Sunflower: History and Reception of a Literary Holocaust Testimony*.

<sup>4</sup> See the [FAIR principles](#) online. Accessed June 2024.

<sup>5</sup> See the [EAD official website](#). Accessed June 2024.

<sup>6</sup> See <https://github.com/oeuvres/odette>, and Odette documentation (in French): <https://resultats.hypotheses.org/267>. Accessed June 2024.

online using Omeka along with an EHRI-specific plugin for publishing TEI files. The editorial process is based on the utilization of the TEI standard, which acts as a pivot for accessing the metadata of a document, displaying contextual information related to it, and displaying its content.

The objective of this report is to assess the current status of the EHRI digitization pipeline and the resulting online editions it generates. Additionally, it will examine processes that have not yet been incorporated, such as automation. It will also underline the significance of the standards, TEI for instance, used at each step of the pipeline. These standards play a vital role in ensuring the sustainability of the data produced for EHRI Online Editions. The report will also provide general recommendations and potential trajectories for the evolution of online editions produced under the EHRI framework.

This report will discuss (1) how EHRI rely on archival institutions for acquiring source materials, (2) the process of transcribing textual documents, (3) how raw transcriptions are transformed into TEI files, (4) how the latter can be enriched with the semantic information contained in the documents, and finally (5), it will assess the software used for the dissemination of digitized documents to online users.

## **2. Archival institutions and digitization**

### **2.1. Helping archival institutions in digitizing archival materials and integrating them into the EHRI Portal**

EHRI does not store archival materials itself. Instead, it acts as a centralized platform where institutions with Holocaust-related materials can upload their finding aids, thereby enhancing the discoverability and accessibility of these resources. This integration is achieved using the EAD 2002 and EAD-3 XML standards, ensuring consistency and interoperability across different archives<sup>7</sup>.

### **2.2. Curating EHRI Online Editions: a theme-based approach**

An online edition created by EHRI compiles sources on the history of the Holocaust, which are stored in one or multiple archival institutions. Its goal is to provide a scientific discourse on a specific aspect of the history of the Holocaust. This discourse is crafted by specialists from various institutions within the EHRI network. The project involves extensive editorial work and is composed of two main parts: the edited documents themselves and the historical and research discourse. The latter includes explanations about the purpose of the edition, thorough historical research, interpretations of the sources, and definitions. Additionally, it features indexes that are directly linked to the documents, facilitating a comprehensive understanding of the material presented.

## **3. Digitally transcribing Holocaust archival materials**

---

<sup>7</sup> To address the varying capacities of archival institutions, EHRI offers additional support through its 'Mobile Lab' initiative. This service extends assistance to institutions requiring help with data integration or digitization, which is particularly beneficial for smaller archives lacking the necessary infrastructure. The Mobile Lab's support can range from technical guidance to providing on-site digitization services, making it feasible for these institutions to participate in the EHRI network. See more information on [EHRI's website](#). Accessed June 2024.

### **3.1. A manual approach**

The documents selected for each edition are, as of Summer 2024, transcribed by hand. Although this approach is time-consuming and can become tedious depending on the volume of data to be processed, it ensures a high level of quality control when done by historians and archivists.

However, the documentation of the transcription process (omissions, typographical errors, and handwritten notes transcription) varies significantly across different editions, and still lacks a clear methodology. Currently, there is no standardized transcription guide to follow. This inconsistency creates challenges for keeping homogeneity, consistency and transparency across all EHRI Online Editions.

### **3.2. Archiving the transcription process: creating transcription guidelines**

Creating transcription guidelines for all future EHRI Online Editions would significantly benefit both the EHRI consortium and the broader research community. Transcription guidelines help guide this process itself. As illustrated in (Pinche and Camps 2022), such guidelines help "minimize the collective cost, including that of training people." They also facilitate the creation of shareable, reusable, and durable ground truth datasets.

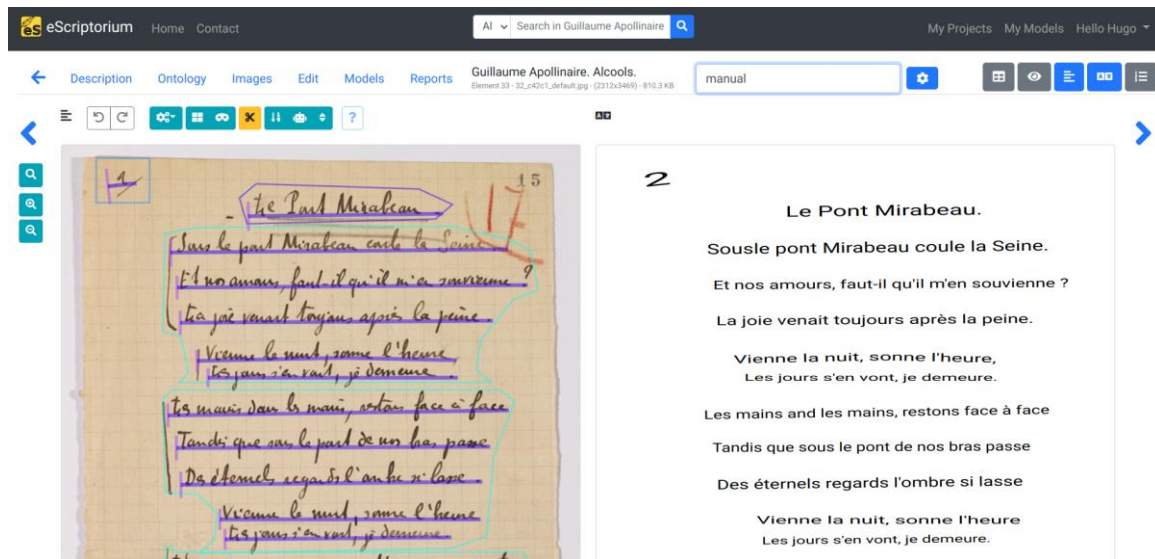
Maintaining a record of the transcription process also ensures full transparency for future researchers working with materials transcribed in the context of EHRI Online Editions. These guidelines enable a clear understanding of how a document was transcribed and allow for easy navigation between the original document and its digital version.

Following the recommendations established by (Stutzmann 2011), EHRI needs to create transcription guidelines that discuss what is transcribed (either everything present in the document, or only parts of it), and how is it transcribed (what are the conventions used for transcribing abbreviations, for example), both with precise scenarios and illustrations. By doing so, EHRI would ensure that the transcriptions produced within the context of the consortium remain interoperable.

### **3.3. Streamlining the transcription process with automatic text recognition**

Implementing a dedicated transcription interface such as such as the open-source graphical user interface eScriptorium (Stokes et al. 2021) allows for a comprehensive archival record of the transcription process, even when manually transcribing. It indeed uses ALTO XML and PAGE XML standards, which not only capture the results of the transcription but also keep the coordinates of each text lines and text regions, therefore aligning the digital transcription to its facsimile.

These standards can be leveraged to archive the digitization of documents for future editions. A transcription interface such as eScriptorium can be used to ergonomically and efficiently manually transcribe a document, or to apply an automatic text recognition (ATR) model to automatically transcribe it, often both at the same time.



The eScriptorium transcription interface.

As ATR technology becomes increasingly accurate and powerful for both printed and handwritten materials, it offers a significant advantage for creating EHRI Online Editions more efficiently and quickly. The learning curve for these technologies is now more accessible with the right tools and interfaces<sup>8</sup>. Partner institutions, like ALMANaCH, possess expertise in this area and could assist in training or even establishing an automatic transcription pipeline. EHRI could aim to develop transcription models for both printed and handwritten documents related to the history of the Holocaust, ready to use for future online editions. These models would improve progressively as more documents are transcribed and then used to retrain the transcription models, enhancing their accuracy and efficiency over time.

EHRI can also leverage existing transcription models, such as Manu McFrench (Chagué et al. 2023) a generic handwritten text recognition model for French modern and contemporaneous texts, which can be fine-tuned on other languages based on the Latin alphabet. By proposing to create transcription models, EHRI also expands its research outcomes beyond just editions, and could offer and share valuable resources for other projects related to the research on Holocaust, on the online ground truth catalog HTR-United (Chagué and Clérice 2022) for instance.

### 3.4. A multilingual ATR model for EHRI Online Editions

In Spring 2024, Floriane Chiffolleau, a Ph.D student from the ALMANaCH team at Inria, Paris, aligned with the help of Sarah Bénérière, research & development engineer in the same team, the existing transcription of documents already edited by EHRI with their facsimile to create ground truth data for training an ATR model.

<sup>8</sup> The ALMANaCH team wrote an online documentation for eScriptorium, accessible [here](#). Accessed June 2024.



252 typescripts documents in seven different languages (Czech, Danish, English, German, Hungarian, Polish, and Slovak) were sampled, the text lines were segmented and then aligned on eScriptorium. The resulting dataset was made available online on Github<sup>9</sup>.

A multilingual model was then trained on the dataset. It achieved 97.20% accuracy. It works well on the languages it was trained on, and achieves good results on unknown languages with the same script, although it would need more training data to achieve a higher accuracy. The main problem of the model remains the recognition of diacritics and uppercase letters. The dataset and the model's accuracy is described in the following table:

Language	Source	Number of documents	Number of lines	ATR model accuracy
German	BeGrenzte Flucht (BF); Die Nisko-Deportationen (Nisko); Early Holocaust Testimony (EHT)	56	2287	97.9%
English	BF; EHT; Diplomatic Reports (DR)	54	1989	97.5%
Czech	BF; EHT	46	1713	96.7%
Danish	DR	36	1007	97.8%
Hungarian	EHT	30	1334	95.7%
Polish	EHT	15	468	93.1%
Slovak	BF	15	395	93.7%
Multilingual	BF; Nisko; DR; EHT	252	9193	97.2%

Such a model could help transcribe more efficiently and more quickly new documents for future EHRI Online Editions.

## 4. Encoding sources: transforming a text into a database with the TEI XML standard

### 4.1. EHRI editorial process

The EHRI Consortium uses the TEI (Text Encoding Initiative) XML standard to semantically markup the transcribed documents that were curated to create an online edition. TEI XML is precisely tailored for digital editions, offering the advantage of seamlessly managing both the intellectual arrangement of the content and its physical representation, as well as its metadata<sup>10</sup>. The EHRI Consortium documented their editorial process online<sup>11</sup>. It relies on three main steps:

- **Annotation using controlled vocabularies.** For instance, the EHRI controlled vocabularies for Holocaust-related entities, or Geonames for geographic information. Documents are annotated in common text editors, and annotations are linked with URLs that act as unique identifiers in the different vocabularies.
- **Conversions to TEI and enriching TEI headers.** As mentioned in the introduction, the annotated documents are then converted into TEI using the

<sup>9</sup> <https://github.com/FloChiff/ehri-dataset>. Accessed June 2024.

<sup>10</sup> See the Text Encoding Initiative [website](#). Accessed June 2024.

<sup>11</sup> See the editorial process [documentation online](#). Accessed June 2024.

open source tool Odette. A TEI enhancement utility has been developed to help develop linked control vocabularies<sup>12</sup>. It is a command-line interface written in PHP, allowing for a possible integration into Omeka, which fetch all entities annotated in the <body>, create entity lists accordingly in the <teiHeader>, and performs rule-based enrichment of the entity by fetching metadata using external resources such as EHRI vocabularies (places, camps, ghettos and terms) and Geonames (geographic coordinates, and further resources such as Wikipedia articles). The tool adds a normalized version of the metadata in conformance with a Dublin Core to TEI mapping. These two first steps lack a controlled TEI schema, or ODD (One Document Does it all)<sup>13</sup>, to ensure consistency and validation, but also a precise documentation on all annotation and encoding choices that were made.

- **Documents are then uploaded on a front application based on Omeka.**

The TEI encoding used within the EHRI Online Editions is briefly documented online<sup>14</sup>.

## 4.2. TEI Specifications for a Sustainable Management of Digitized Holocaust Testimonies

During Spring 2023, the ALMAnaCH research team at Inria, Paris, hired a Master's student intern, Sarah Bénière, to work and homogenize the EHRI TEI encoding. She was supervised by Floriane Chiffolleau. After a thorough analysis of the existing editions, Sarah Bénière's report argued that the encoding was not homogeneous throughout the editions, especially in the metadata. The TEI conformance could also be improved with an appropriate and customized TEI schema that could be created for EHRI Online Editions. Her research resulted in an article published in the context of the First Workshop on Holocaust Testimonies as Language Resources (HTRes) at the LREC-COLING 2024 conference, held in May 2024 in Turin, Italy (Bénière, Chiffolleau, and Romary 2024), whose content has been the basis for the current section.

### 4.2.1. Introduction

Research in the humanities has taken a turn with the advent of computational methods. The TEI—*Text Encoding Initiative*, or *Text Encoding for Interchange* ((Unsworth 2011); (Holmes 2016))—has been involved in the processing of textual data since 1988 (Schmidt 2014) and has become a widely used standard in Digital Humanities for structuring textual documents at large ((Burnard 2014); (Burnard 2018)). In 2018, the European Holocaust Research Infrastructure<sup>15</sup> (EHRI) published its first online edition of Holocaust testimonies: *BeGrenzte Flucht*, or “Bordered Escape”, encoded according to the general TEI All schema, which we will discuss later in this section.

Numerous digital scholarly editions of textual documents have been published, mainly of historical and literary texts (Schmidt 2014), and have generally contributed to the advancement of Digital Humanities. Since the 1990s, the TEI has evolved and expanded greatly in a desire

<sup>12</sup> See the [documentation online](#). Accessed June 2024.

<sup>13</sup> See what is an ODD on the TEI [website](#). Accessed June 2024.

<sup>14</sup> See the TEI encoding and annotation [documentation online](#). Accessed June 2024.

<sup>15</sup> <https://www.ehri-project.eu/>. Accessed June 2024.

to meet the needs of the research community as much as possible ((Bauman 2011); (Holmes 2016); (TEI Consortium 2023)). For example, the development of the Shelley-Godwin Archive project (Muñoz and Vigiante 2015) coincided with the improvement of Chapter 11 of the TEI Guidelines “Representation of Primary Sources” (TEI Consortium 2023), which proved incredibly useful to the community having to deal with legacy material.

The issue of standardizing encoding practices for specific purposes, such as the publication of Holocaust testimonies, remains to be addressed. Our corpus, the EHRI Online Editions, is a great test-bed for doing so. In the course of taking up the existing editions with the purpose of providing a stable publishing environment for them, we observed disparities and inconsistencies in the encoding from one edition to another due, in particular, to the improvement of the encoders’ skills over time. As a result, the need for normalization within the EHRI Online Editions emerged, as well as a broader reflection on the standardization of the encoding of Holocaust-related documents.

We first present the TEI customization that we developed for the EHRI Online Editions, and how it can be extended to standardize the encoding of Holocaust-related textual documents. We then present the EHRI Online Editions, followed by data structuration in TEI, and a focus on the EHRI TEI customization<sup>16</sup>. Finally, we discuss the extension of the EHRI specifications to all encoding projects dealing with Holocaust-related documents.

#### 4.2.2. The EHRI Online Editions

EHRI is a transnational consortium funded by the European Union (EU) with partnering institutions all across Europe, Israel, and the United States. It is coordinated by the NIOD Institute for War, Holocaust and Genocides Studies based in Amsterdam, Netherlands. EHRI is currently in its third phase (EHRI-3, 2020-2024), organized in twelve work packages (WP), among which the WP10 “New Approaches to Holocaust Research and Archiving”.

Within the framework of WP10, EHRI has already published six online editions<sup>17</sup>. These digital editions are collections of archival documents held by EHRI’s various partnering institutions, gathered together and processed by EHRI’s editors and made available online<sup>18</sup>. The description of each EHRI Online Edition is available in the Appendix of this document.

---

<sup>16</sup> [https://github.com/SarahBeniere/EHRI-Workflow/blob/main/ENCODING/Guidelines/ODD\\_EHRI.xml](https://github.com/SarahBeniere/EHRI-Workflow/blob/main/ENCODING/Guidelines/ODD_EHRI.xml). Accessed June 2024.

<sup>17</sup> <https://www.ehri-project.eu/ehri-online-editions>. Accessed June 2024.

<sup>18</sup> When unavailable on EHRI’s website, the translations of the titles of the editions in English are our own.

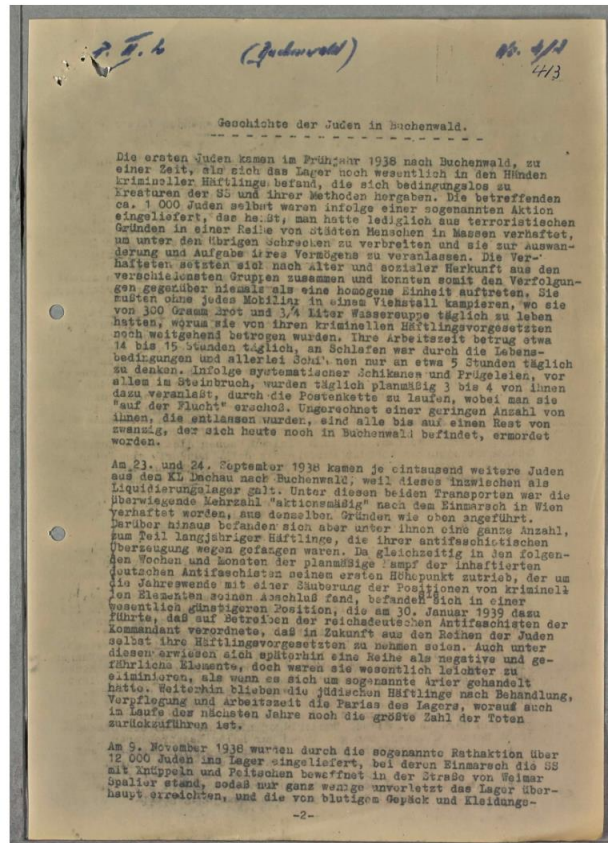


Figure 1: Example of testimony

#### 4.2.3. Structuring Data in TEI

##### A Standard for Structuring Textual Documents

As briefly mentioned in the introduction, the TEI Guidelines are a widely adopted standard for structuring textual documents in, among other applications, digital scholarly edition projects. They are based on the W3C XML recommendation, and provide “a highly interoperable format” (Schmidt 2014) with a set of recommended elements that come with a precise syntax and documentation. These recommendations are compiled in the TEI infrastructure as both a technical specification and extensive prose (TEI Consortium 2023), thus ensuring a common knowledge on the encoding of textual data for research in the humanities. In the case of the EHRI Online Editions, choosing the TEI instead of developing their own arbitrary EHRI tagset has two main advantages:

1. Using the TEI gives relevance to the project, because it aligns with the values and practices of a wider community ((Burnard 2014); (2018)) and thus facilitates the integration of the outputs within a wider corpus, as well as it increases the possibility to reuse existing editing, query, or publishing tools.
2. It also aligns with the pre-existing practices of EHRI as an infrastructure, given that their system already relies on XML technology, in particular on EAD-XML ((Alexiev, Nikolova, and Hateva 2019); (Levy 2019); (Romary and Riondet 2019)).

According to Lou Burnard ((2014); (2018)), the success of the TEI in Digital Humanities projects lies in its three main characteristics:

1. Contrary to typical word processors like Microsoft Word or LibreOffice Writer—which

tend to focus on the aesthetic rendering of the text—a TEI encoding is semantic. It is particularly useful for named entities disambiguation tasks. For example, the character string “Warsaw” could either refer to the city and capital of Poland Warsaw, or to the Warsaw Ghetto (Figure 2).

```
<placeName type="city">Warsaw</placeName>  
<placeName type="ghetto">Warsaw</placeName>
```

Figure 2: Disambiguation of “Warsaw” in TEI

2. A TEI-XML file, like all XML files, is a succession of characters that both humans and machines can read and understand. As a result, the action of opening and reading the content of a TEI-encoded text is independent of any software, whereas a Microsoft Word document (.docx), for example, requires at the very least a word processor to open.
3. The TEI recommendations are sustained by the TEI Consortium and improved by the continuous involvement of the TEI community. In addition, because the Guidelines are available online and the community is active, it makes it an accessible technology for beginners.

### Best Practices and Standardization

When encoding a text in XML, the encoder is free to use whatever tags they want and to give them a meaning of their own. In his article on TEI conformance, Lou Burnard (2018) gives the example of the <p> tag. Generally speaking, <p> is used to encode a paragraph, but we could decide that in the case of our encoding it means “potato”. This example highlights the relevance of a standard like the TEI. Nevertheless, criticism has been expressed toward the TEI as being too wide and too restrictive at the same time, or the choice of the tags being guided by human interpretation of the text, thus leading to an impediment of interoperability ((Bauman 2011); (Schmidt 2014)).

While we agree with the fact that the encoders choosing which element they want to draw attention to makes interchange difficult *per se*, because it implies that everyone is aware of the purpose of said encoding, we argue that a solution could be the implementation of a schema and documentation by means of an ODD. The ODD—for *One Document Does-it-all*—is a TEI-XML file which contains both a customization of the TEI and its associated documentation. From the ODD file, we can derive a RelaxNG validation schema with the customized TEI specifications, but also the prose documentation for the human reader to understand the purpose and extent of the project’s encoding. In addition, an ODD established by an experienced TEI user can help a beginner to make sure their encoding is valid.

We previously alluded to a few inconsistencies in the TEI encoding of the EHRI Online Editions. This is due, on the one hand, to the improvement of the encoders’ skills over time, and on the other hand to the fact that the declared validation schema was “TEI All”. As the name suggests, the TEI All schema encompasses all elements and attributes from the TEI. However, no project would ever use them all, thus emphasizing the relevance of a TEI customization, which “expresses how a given project has chosen to interpret the general principles enumerated by the Guidelines, as well as formally specifying which particular component of the Guidelines it uses” (Burnard 2018). In addition, this profusion of TEI elements can easily lead to confusion between several elements (typically <bibl>, <biblFull> and <biblStruct>), especially for encoders who might not yet be familiar with TEI-XML.



The TEI customization and specifications associated can help define a framework within which the encoders can work and apply best practices. For example, a good practice in TEI-XML consists in structuring the <body> of the <text> with at least one <div> (division) element (Figure 3). We decided to make this a mandatory rule in the EHRI specifications (Figure 4).

```
<body>
  <div type="transcription" xml:lang="de">
    <pb n="1"/>
    <p>[...]</p>
  </div>
</body>
```

Figure 3: Minimal template for the <body>

```
<schemaSpec ident="body" mode="change">
  <!-- div is mandatory in the body -->
  <content>
    <elementRef key="div" minOccurs="1" maxOccurs="unbounded"/>
  </content>
</schemaSpec>
```

Figure 4: Schema specification for <body>

This framework applies to both published and future editions. For editions that have already been published, we wrote a Python script to automatically apply the new RelaxNG schema to all the XML files<sup>19</sup>. For future editions, the schema should be applied instead of “TEI All” from the beginning.

As a final general good practice, we recommended using international norms like ISO to fill in the value for an attribute. The ISO norms we included in the EHRI specifications are:

1. ISO 639<sup>20</sup> codes for the representation of languages;
2. ISO 3166<sup>21</sup> codes for the representation of names of countries;
3. ISO 8601<sup>22</sup> standard for dates (YYYY-MM-DD).

#### 4.2.4. TEI Customization for Holocaust Testimonies

##### Normalizing the EHRI Online Editions

Until now, the texts selected by the editors were transcribed and encoded manually (Frankl et al. 2018), which raised two main issues:

1. It is an extremely time-consuming and tedious task;
2. It is a source of encoding mistakes.

In order to write the ODD for the EHRI Online Editions, we needed to analyze the encoding practices of the encoders for the editions that had already been published: “Bordered Escape”, “Early Holocaust Testimony”, “Diplomatic Reports”, and “Nisko”. We noticed, for instance, recurring mistakes in the spelling of attribute values (i) or the usage of different languages (ii): (i) @type=“subejct” or (ii) @type=“subjekt” (German) instead of @type=“subject”. Even though they may refer semantically to the same entity—a term (<term>) for example—the

<sup>19</sup> <https://github.com/EHRI/ehri-online-editions>. Accessed June 2024.

<sup>20</sup> <https://www.iso.org/iso-639-language-code>. Accessed June 2024.

<sup>21</sup> <https://www.iso.org/iso-3166-country-codes.html>. Accessed June 2024.

<sup>22</sup> <https://www.iso.org/iso-8601-date-and-time-format.html>. Accessed June 2024.

machine will consider them as different instances. This leads to an incorrect count of the occurrences and to referencing mistakes that are not easily detectable.

One of the normalizing aspects for the EHRI Online Editions which we considered important is the language chosen for encoding the metadata. In an edition gathering documents from different holding institutions, the metadata should be filled in thoroughly. In a spirit of data reuse, we thought that all metadata should appear at least in English. Some metadata can be translated, like the title of the document (Figure 5) or the name of its holding institution. For example, the original name for the Jewish Museum in Prague is “Židovské muzeum v Praze” (Czech), but we estimated that the most commonly understood language among EHRI partners would be English. Therefore, we established English as the main language for encoding the metadata.

```
<title xml:lang="en">List of Viennese Nisko deportees who died in Kamensk-Uralski</title>
<title xml:lang="de">Liste von Wiener Nisko-Deportierten, die in Kamensk-Uralski verstarben</title>
```

Figure 5: Encoding of the title of a document

Normalizing the EHRI Online Editions is the first step toward TEI specifications for the standardization of Holocaust-related documents in TEI-XML. Indeed, the ODD for the EHRI Online Editions serves three purposes:

1. Avoiding encoding mistakes as much as possible;
2. Setting up good encoding practices in general, especially in case any of the encoders is not yet familiar with TEI-XML;
3. Establishing a validation schema particularly suitable for Holocaust-related textual documents, derived from the ODD, insofar as simultaneously harmonizing the previously published EHRI digital editions and ensuring the consistency of the future ones.

## Points of Interest in the EHRI TEI Specifications

### *Language Codes (ISO 639)*

Even though this is a mistake that was rapidly corrected in the second edition, we found some inconsistency in the codes chosen for the representation of languages as values for the @xml:lang attribute. It is naturally tempting to use a code that would be correct in one's own native language, which can result in referencing mistakes like the misspelling of “subject” we mentioned earlier. A very common example of such bias is the representation of the German language: we could imagine either “de” for “Deutsch” (German), “ger” for “German” (English), and even “all” for “Allemand” (French). While all these codes are correct representations of the German language, they must not be used all at once within the same edition. As a result, we recommended that the encoders use the codes provided by ISO standard 639 (Figure 6), available through the IANA Language Subtag Registry<sup>23</sup>.

```
<valList mode="add" type="semi">
  <valItem ident="cs">
```

<sup>23</sup> <https://www.iana.org/assignments/language-subtag-registry/language-subtag-registry>. Accessed June 2024.

```

    <desc>Czech</desc>
  </valItem>
  <valItem ident="da">
    <desc>Danish</desc>
  </valItem>
  <valItem ident="de">
    <desc>Deutsch</desc>
  </valItem>
  <valItem ident="el">
    <desc>Modern Greek</desc>
  </valItem>
  <valItem ident="en">
    <desc>English</desc>
  </valItem>
  <valItem ident="es">
    <desc>Spanish</desc>
  </valItem>
  <valItem ident="fr">
    <desc>French</desc>
  </valItem>
  <valItem ident="he">
    <desc>Hebrew</desc>
  </valItem>
  <valItem ident="hu">
    <desc>Hungarian</desc>
  </valItem>
  <valItem ident="it">
    <desc>Italian</desc>
  </valItem>
  <valItem ident="ja">
    <desc>Japanese</desc>
  </valItem>
  <valItem ident="nl">
    <desc>Dutch</desc>
  </valItem>
  <valItem ident="pl">
    <desc>Polish</desc>
  </valItem>
  <valItem ident="ru">
    <desc>Russian</desc>
  </valItem>
  <valItem ident="sk">
    <desc>Slovak</desc>
  </valItem>
  <valItem ident="uk">
    <desc>Ukrainian</desc>
  </valItem>
  <valItem ident="yi">
    <desc>Yiddish</desc>
  </valItem>
</valList>

```

Figure 6: Language codes used by EHRI

### **Implementing Controlled Vocabulary**

The EHRI Portal<sup>24</sup> presents itself as one of the main resources about the Holocaust as it gathers information on archival sources from across the world. One of its primary

<sup>24</sup> <https://portal.ehri-project.eu/>. Accessed June 2024.



achievement is the creation of controlled vocabulary. Among the EHRI terms, some are identified as linguistically distinct because they are vocabulary coined by the Nazis or specifically used in reference to the concentration and extermination camps. In the continuity of the encoding work performed by the EHRI encoders, we modified the specifications for the <distinct> element. As a result, we made the @type attribute mandatory and suggested a semi-open list of values containing “camp\_language” and “nazi\_language” (Figure 7). Hence, a dialog box with the list of possible values appears every time the @type attribute from the <distinct> element is filled in when encoding a text.

```
<elementSpec ident="distinct" mode="change">
  <attList>
    <!-- @type is mandatory and its value is either camp_language or
nazi_language -->
    <attDef ident="type" mode="change" usage="req">
      <valList mode="add" type="semi">
        <valItem ident="camp_language"/>
        <valItem ident="nazi_language"/>
      </valList>
    </attDef>
  </attList>
</elementSpec>
```

Figure 7: Specifications for <distinct>

### Including Translation(s) in a Single File

The EHRI ODD is part of a broader workflow for processing Holocaust-related documents<sup>25</sup>. The last step of this workflow is the publication of the editions on a TEI Publisher<sup>26</sup> application dedicated to all the EHRI digital editions. In order to do so, we decided to include the documents in their original language as well as their translation(s) within a unique file bearing the EHRI identifier, for example "EHRI-ET-WL16560413" (Figure 1). This is done by ensuring the structuration of the <body> with first-level <div> (division) elements specified with the attributes @type (Figure 8) and @xml:lang (ISO 639 values).

```
<elementSpec ident="div" mode="change">
  <constraintSpec scheme="schematron" ident="div-1">
    <constraint>
      <s:rule context="tei:TEI/text/body/div[@type]">
        <s:assert test="@type='transcription' or
@type='translation'">Value for @type in first-level division is either
transcription or translation</s:assert>
      </s:rule>
    </constraint>
  </constraintSpec>
  <attList>
    <!-- @type is mandatory and its value should either be
transcription or translation -->
    <attDef ident="type" mode="change" usage="req">
      <valList mode="add" type="semi">
        <valItem ident="transcription"/>
        <valItem ident="translation"/>
      </valList>
    </attDef>
```

<sup>25</sup> <https://github.com/SarahBeniere/EHRI-Workflow/tree/main>. Accessed June 2024.

<sup>26</sup> <https://teipublisher.com/exist/apps/tei-publisher-home/index.html>. Accessed June 2024.

```
</attList>  
</elementSpec>
```

Figure 8: Specifications for first-level &lt;div&gt;

***Encoding Template for the <teiHeader>***

As we mentioned previously, particular attention must be given when encoding the documents' metadata. We created a template (Figure 9) to make sure that no available piece of information is missing. A good practice that needs to be implemented by the EHRI encoders is the use of the <revisionDesc> so as to follow all the modifications made within the file. The template also contains fields that are already filled in because their value is consistent for every single file: <affiliation> and <authority> will always be EHRI, and we share the documents according to the Creative Commons Attribution 4.0 International license (CC BY 4.0)<sup>27</sup>.

---

<sup>27</sup> <https://creativecommons.org/licenses/by/4.0/>. Accessed June 2024.

```

<teiHeader>
  <fileDesc>
    <titleStmt>
      <title xml:lang="en"/>
      <title xml:lang=""/>
      <principal>
        <affiliation>
          <orgName ref="https://www.ehri-project.eu">
            European Holocaust Research Infrastructure
          </orgName>
        </affiliation>
      </principal>
      <respStmt>
        <resp/>
        <persName/>
      </respStmt>
    </titleStmt>
    <publicationStmt>
      <authority>
        <ref target="https://www.ehri-project.eu">European Holocaust Research Infrastructure</ref>
      </authority>
      <availability>
        <licence target="http://creativecommons.org/licenses/by-sa/4.0">
          Attribution-ShareAlike 4.0 International
        </licence>
      </availability>
    </publicationStmt>
    <seriesStmt>
      <title ref="{link to the online edition}"/>
    </seriesStmt>
    <sourceDesc>
      <msDesc>
        <msIdentifier>
          <institution>
            <orgName/>
            <address>
              <street>
                <num/>
              </street>
              <postCode/>
              <settlement/>
              <country/>
            </address>
          </institution>
          <collection/>
          <idno/>
        </msIdentifier>
        <physDesc>
          <p/>
        </physDesc>
      </msDesc>
      <bibl>
        <textLang/>
      </bibl>
    </sourceDesc>
  </fileDesc>
  <encodingDesc>
    <projectDesc>
      <p xml:lang="en"/>
    </projectDesc>
  </encodingDesc>
  <profileDesc>
    <creation>
      <origDate when=""/>
      <origPlace ref="{GeoNames link}"/>
      <persName ref="{EHRI entity}"/>
    </creation>
    <textClass>
      <catRef target="{link to EHRI portal}"/>
      <keywords>
        <term/>
      </keywords>
    </textClass>
    <langUsage>
      <language ident=""/>
    </langUsage>
    <abstract>
      <p xml:lang="en"/>
    </abstract>
  </profileDesc>
  <revisionDesc>
    <change when="" who="{}/>
  </revisionDesc>
</teiHeader>

```

Figure 9: Template for the <teiHeader>

## 4.2.5. Discussion

We have seen in this section the TEI specifications developed in the context of the EHRI Online Editions. The implementation of the EHRI ODD is organized in two steps: the processing of editions that have already been published, and the processing of future digital editions. The texts of the previous editions must be validated against the RelaxNG schema derived from the EHRI ODD, and we have experimented with a Python script to automatically apply the new schema to the texts that were already encoded. As for future editions, the texts are to be encoded according to the TEI specifications defined in the EHRI ODD<sup>28</sup>. We present the EHRI ODD as a starting point for the standardization of encoding practices regarding Holocaust-related textual documents. Indeed, using semi-open lists for attribute values for example allows an extension to documents in more languages, and/or containing other types of specific vocabulary. As we are strong advocates of the open science approach, we make the EHRI ODD public and reusable according to the terms of the CC BY 4.0 license. Therefore, it can serve as a basis for the development of more complete encoding guidelines for Holocaust testimonies, following the “ODD chaining” tutorial by Lou Burnard (2016) for instance.

## 5. Extracting structured information

### 5.1. Integrating layout information into TEI with document layout analysis technologies

The current EHRI TEI schema relies on a basic encoding of the documents' layout. With the help of document layout analysis technologies (DLA), the representation of the layout could be more fine-grained and could be semi-automatically built in TEI. DLA aims at detecting the layout components and hierarchy of a document. State-of-the-art DLA systems are based on machine learning leveraging two approaches: visual features (Sven and Matteo 2022), or a combination of textual and visual and textual features (Huang et al. 2022).

The ALMAnaCH team sampled and annotated 200 images from the documents available in the Early Holocaust Testimonies EHRI Online Edition. This process was done manually with the Roboflow interface<sup>29</sup>. As object detection models demonstrated their reliability for segmenting the layout of historical textual documents (Clérice et al. 2024; Sven and Matteo 2022), we opted for this text-free approach and trained a YOLOv8 model. The following illustrations depict the different layout identified in this online edition. The range varies from basic to more intricate layouts.

---

<sup>28</sup> As of now, new EHRI editions have not been prepared yet.

<sup>29</sup> [Roboflow](#) is an online software platform designed to assist with computer vision tasks. It provides tools and services for creating, managing, and deploying computer vision models. Accessed June 2024.

Page number

- 3 -

Paragraphs

képviselője és Weissenmayer német követnek a helyettese. Megjelent továbbá az értekezleten a hírszót Peranczy csendőrszerepéről is, aki valószínűleg mint ismerttransportszakértő volt hivatalos.

A külügyminiszter képviselője az értekezleten hivatalosan bejelentette a német külügyminiszterium hozzájárulását a certifikáttal rendelkezők kivándorlásához. Később egyben a magyar kormány elvárta, hogy az állja le-bonyolítását / aug. 01. / a deportálódást felügyeleti és ezen idő alatt a munkatársak behívása is szabotálni fog. Ezen az értekezleten átvették a kivándorlókat befogadni a megállapításra, valamint az utazás részleteinek a megvitatására.

Ezen az értekezleten vetődött fel első ízben a "vándorhadsak" gondolata, ahová az állásidő elutazásuk előtt egybe-gyűjtendőek lettek volna.

Ezen az értekezleten megvitették a "Vadász-utca" történetét, ahová még július közepén bevonultunk, itt a "kollégák utolsó", majd az ebből folyó "Schutzpass" megszületett a itt a Schutzpassok által megszerzett 60-80 ezer emberen kívül 2800 szülő ifjúság mellett meg, köztük nagy tömegben az a cionista ifjúság, amely az ideiglenes, földalatti mentő munka szervezője és irányítója volt.

A történelmi háttérrel kapcsolatban az alkalommal azt a megítélést hoztuk, hogy a "Vadász-utca" történetét a "Vadász-utca" történetével szembe kell vetni, valamint ezek közötti különbségeket is fel kell tárni, valamint ezek közötti különbségeket is fel kell tárni.

Ezen után nem csodálkozhattunk azon, hogy a városi önkormányzatban nyilvánvalóan elleni irányulatnak a szöveg szerinti a "Vadász-utca" történetét szembe kellett vetni a "Vadász-utca" történetével szembe, amely a december 4.-i hajnali csendőrszerepéről, amelyet nagy utcai sorsra vezetett be, majd utána az utcai sorsra megjelentek a katonák. Szerepük az utcai sorsra megjelentek a katonák, amely a december 4.-i hajnali csendőrszerepéről, amelyet nagy utcai sorsra vezetett be, majd utána az utcai sorsra megjelentek a katonák.

Author

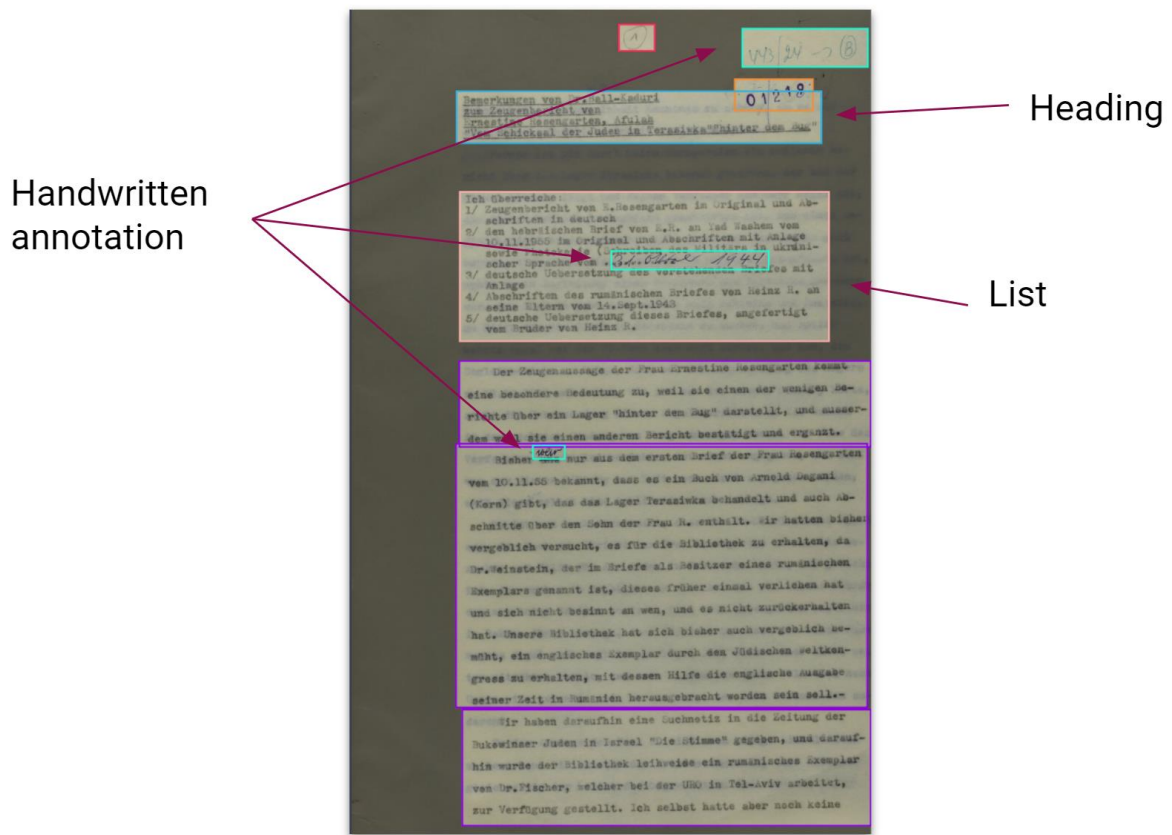
Stival Ellen

Handwritten signature

A jegyzőkönyvet felvette  
Stival Ellen

Stamp





The annotation schema is based on the SegmOnto ontology, a controlled-vocabulary based on TEI. It ensures that the annotated dataset is easily interoperable and reusable.

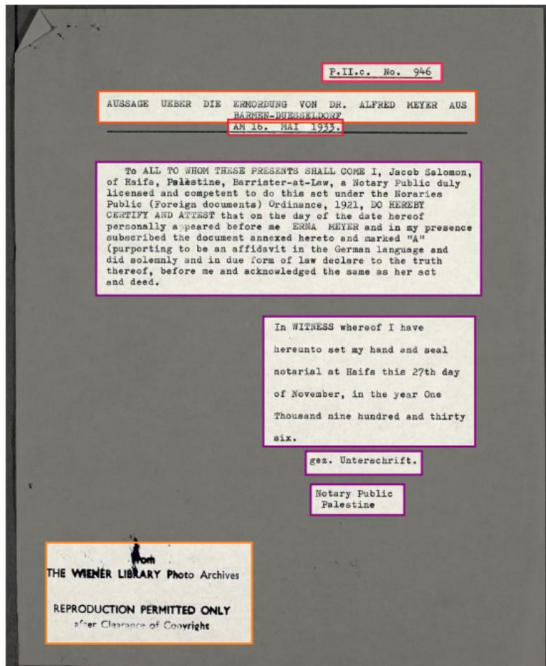
A total of 13 classes were identified:

- MainZone:Date: structured dates.
- MainZone:Form: forms that can be visually identifiable.
- MainZone:Head: any title present in a document.
- MainZone:List: bullet-point lists.
- MainZone:P: paragraphs.
- MainZone:Signature: typewritten or printed signatures.
- MarginTextZone:Signature: handwritten signatures.
- MarginTextZone:ManuscriptAddendum: manuscript annotations.
- MarginTextZone:Notes: any printed note outside the main body of text.
- NumberingZone: number of the page.
- RunningTitleZone: a title, generally at the top of the page.

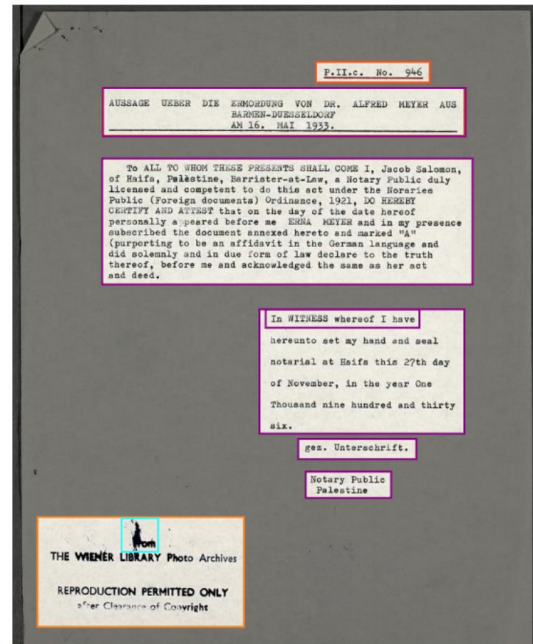


- StampZone: ink stamps.
- StampZone:Sticker: stickers.

The YOLOv8 model trained on this first dataset achieved satisfactory results with a mean Average Precision (mAP)<sup>30</sup> of 75.4%, a precision of 69.8% and a recall of 76.2%. This model, although not perfect, can be used to pre-annotate the layout of the edited textual documents. To improve the scores, more documents from other editions should be sampled and annotated. The following illustrations show the ground truth on the left, and the model's prediction on the right.

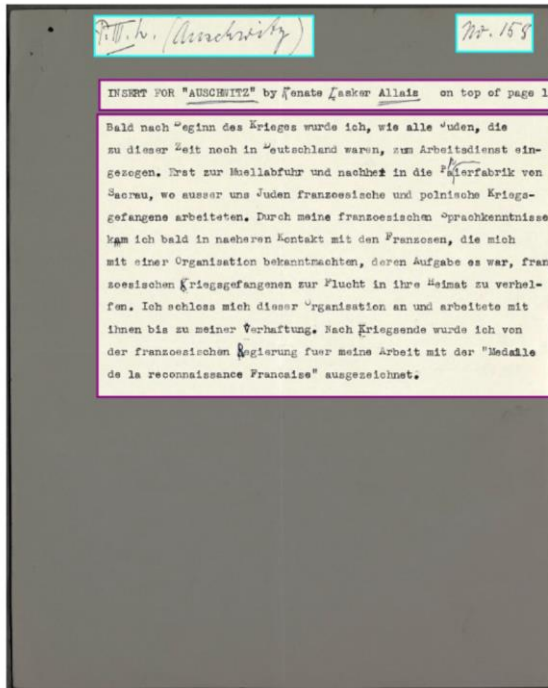


Ground truth

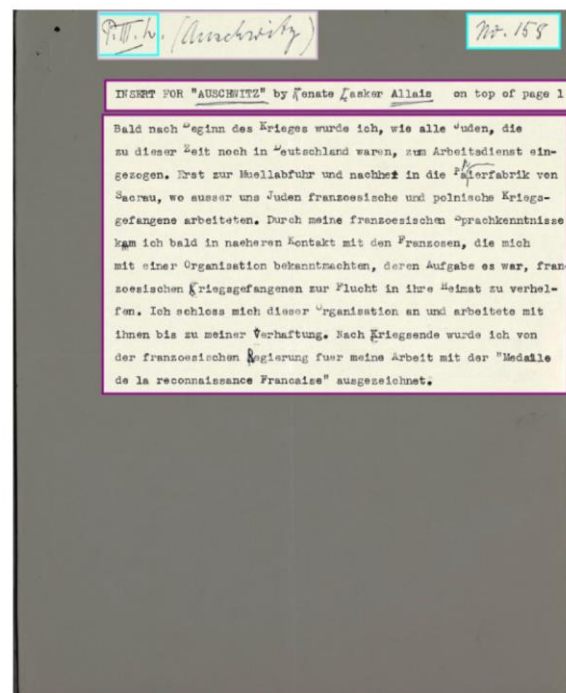


Prediction

<sup>30</sup> Mean Average Precision (mAP) is a common performance metric in information retrieval and machine learning, especially for evaluating the accuracy of object detection models. It is calculated as the mean of the average precision values across all queries or categories, where average precision considers the precision of retrievals at different recall levels.



Ground truth



Prediction

The model performs well when extracting generic layout components such as paragraphs and headers, but still has trouble on other classes less represented in the dataset. This can be overcome with a more thorough sampling of source documents.

The goal would be, ultimately, to go from the source document, apply a DLA EHRI dedicated model to annotate its layout, apply a transcription model previously mentioned, and then semi-automatically build a TEI file based on these operations, which would include the source textual material and its layout. Researchers and editors then would have to validate or correct this process, and then move on to the entity annotation step.

## 5.2. Annotating named entity

As stated during an EHRI online meeting involving several members engaged in EHRI online digital editions, held on July, 4th 2023, annotation remains the most arduous and time-consuming task, as it is done manually.<sup>31</sup> It is nonetheless a crucial task that has to be performed carefully, as EHRI Online Editions focus on linking documents to external resources. Named entities, alongside the metadata of the documents, are the basis of the linking process in TEI files.

<sup>31</sup> Persons attending the meeting: Wolfgang Schellenbacher (Vienna Wiesenthal Institute for Holocaust Studies), Aneta Plizáková (Masaryk Institute and Archives of the Czech Academy of Sciences), Michal Frankl (Masaryk Institute and Archives of the Czech Academy of Sciences), Michala Lönčíková (Masaryk Institute and Archives of the Czech Academy of Sciences), Mike Bryant (King's College London), Maria Dermentzi (King's College London), Floriane Chiffolleau (Inria, ALMANaCH), Sarah Bénière (Inria, ALMANaCH), Hugo Scheithauer (Inria, ALMANaCH)



### 5.3. Repurposing Holocaust-Related Digital Scholarly Editions to Develop Multilingual Domain-Specific Named Entity Recognition Tools

ALMAnaCH research team and King's College, London (KCL) members suggested the use of automated tools to help EHRI editors to annotate new documents for future online editions. In early 2024, Maria Dermentzi, Web Research Developer at the KCL and Hugo Scheithauer, Ph.D student in the ALMAnaCH research team at Inria collaborated on creating a named entity dataset based on the EHRI Online Editions and trained a multilingual named entity recognition model that could be used to kickstart future editions. Their research resulted in an article published in the context of the First Workshop on Holocaust Testimonies as Language Resources (HTRes) at the LREC-COLING 2024 conference, held in May 2024 in Turin, Italy (Dermentzi and Scheithauer 2024), which is the basis for this section.

#### 5.3.1. Introduction

Launched in 2010, the European Holocaust Research Infrastructure (EHRI)<sup>32</sup> aims to support Holocaust research by making information about dispersed archival material held by institutions around the world more accessible and interconnected through the EHRI Portal<sup>33</sup> (Blanke et al. 2017). While the EHRI Portal is EHRI's flagship service, the EHRI Consortium is offering a series of additional resources, tools, and services that help researchers and archivists describe, analyze, enrich, and present Holocaust-related material using innovative methods (Leeuw et al. 2018). Apart from the EHRI Portal, of particular relevance to this section are the EHRI controlled vocabularies, the EHRI authority sets, and the EHRI Online Editions<sup>34</sup>.

As an aggregator of multilingual Holocaust-related archival material from diverse institutions, the EHRI Portal is faced with a significant challenge relating to the fact that this material is often described not only in various languages but also using a variety of methodologies and in-house, language-specific controlled vocabularies that need to be normalized to a shared vocabulary to be smoothly ingested in the EHRI Portal (Erez et al. 2020). For this reason, EHRI has developed custom controlled vocabularies and authority sets mainly derived from already existing ones developed by institutions such as Yad Vashem (YV), the United States Holocaust Memorial Museum (USHMM), Arolsen Archives, etc. (Kepa J. Rodriguez et al. 2016; Erez et al. 2020), covering lists of concentration camps, ghettos, subject headings, personalities and corporate bodies<sup>35</sup>. These vocabularies are primarily used for indexing purposes in the EHRI Portal, allowing for semantic search (Colavizza, Ehrmann, and Bortoluzzi 2019) through keyword-based browsing and play a crucial role in achieving EHRI's goal of interlinking multilingual and heterogeneous Holocaust collections. They are also used to enhance the EHRI Online Editions and articles in the EHRI Document Blog<sup>36</sup> with more information and references to the EHRI Portal.

However, creating links between resources hosted across different EHRI services and the EHRI vocabularies is a resource-intensive process, usually done manually. Creating a tool

---

<sup>32</sup> EHRI project website. Accessed 2/27/2024.

<sup>33</sup> EHRI portal website. Accessed 2/27/2024.

<sup>34</sup> See EHRI controlled vocabularies, EHRI authority sets, and EHRI Online Editions. Accessed 2/27/2024.

<sup>35</sup> The aforementioned lists and sets are available online. See camps, ghettos, terms, personalities, and corporate bodies. Accessed 2/27/2024.

<sup>36</sup> See EHRI Document Blog. Accessed 2/27/2024.

capable of detecting named entities (NE) in texts such as Holocaust testimonies or the text in Holocaust-related archival descriptions would make it easier to link more material with relevant identifiers in the EHRI vocabularies, semantically enriching it and making it more discoverable in the Portal and other EHRI services. The significance that reliable Named Entity Recognition (NER) and entity linking (EL) tools may have for EHRI has been highlighted in previous work (Kepa Joseba Rodriguez et al. 2012; Leeuw et al. 2018). Having access to a good NER tool can help with building a reliable EL tool. EHRI partners have previously experimented with the development of such tools (Kepa Joseba Rodriguez et al. 2012; Leeuw et al. 2018; Nikolova and Levy 2018). However, since the publication of the most recent paper related to EHRI and NER (Leeuw et al. 2018), EHRI's growth in resources and advances in Machine Learning (ML) promise better results compared to earlier experiments. In this part, we report on recent work towards Holocaust-related NER.

Specifically, we treat the EHRI digital scholarly editions (i.e., EHRI Online Editions) as a dataset for training and evaluating ML-powered NER models. We have converted all available Extensible Markup Language (XML) files from the EHRI Online Editions into a trainable corpus in a format suitable for NER and have leveraged this dataset (See Table [tab:dataset\_stats]) to fine-tune a multilingual language model for NER. The resulting model can be used as part of a pipeline whereby, upon inputting some text into a tool that supports our models, potential named entities within the text will be automatically pre-annotated in a way that helps users detect them faster and link them to their associated controlled vocabulary entities. This has the potential to facilitate metadata enrichment of descriptions in the Portal and enhance their discoverability. It would also make it easier for EHRI to develop new Online Editions and unlock new ways for archivists and researchers within the EHRI network to organize, analyze, and present their materials and research data in ways that would otherwise require a lot of tedious work.

Our contributions are: the EHRI-NER dataset, a multilingual NER model for Holocaust-related texts, and experiments studying the multilingual learning and cross-lingual transfer capabilities of Deep Learning NER techniques. In what follows, we describe related work and provide detailed information on the source of our dataset, the EHRI Online Editions. Subsequently, we detail how we put together the dataset and how we designed and carried out our fine-tuning experiments.

### 5.3.2. Related Work

Previously, EHRI experimented with applying off-the-shelf NER tools to the Optical Character Recognition (OCR) output of type-written Holocaust survivor testimonies and newsletters for the crew of H.M.S. Kelly (Kepa Joseba Rodriguez et al. 2012). Due to the lack of an already available annotated corpus for domain-specific NER tools, Kepa Joseba Rodriguez et al. (2012) manually annotated the OCRed corpus compiled for their experiments. Given the lack of resources, their experiments remained limited and focused on comparing which of the then-existing NER tools yielded the best results. The maximum total F1 score achieved across all tools and datasets under consideration was 60% (Kepa Joseba Rodriguez et al. 2012). In 2018, Leeuw et al. (2018) published another paper detailing EHRI's efforts to offer reliable NER services for the Holocaust domain. They reiterated the lack of suitable corpora and crafted their own gold corpus by crowd-sourcing annotations on transcripts of oral testimonies provided by the USHMM (Leeuw et al. 2018; Nikolova and Levy 2018). They used this corpus to develop person and location extraction services. Their methodology included

fine-tuning and extending commercial software and they achieved an F1 score of 77% for person extraction. For location extraction, they adapted a proprietary service to tag and disambiguate locations in Holocaust testimonies. The details of these tools are not specified but the authors reported a resulting F1 score of 91% for the disambiguated place-related access points, although it is unclear how the first part of their pipeline (i.e. the tagger) performed. To our knowledge, neither the purpose-built NER datasets nor the EHRI-specific tools developed during earlier work are publicly available today or were formally deployed as EHRI services.

Apart from EHRI-related efforts, there is a broader interest in applying NER tools on Holocaust-related texts (Ezeani et al. 2023; Carter et al. 2022) as well as in developing domain-specific ones. Notable examples include Mattingly's ((2021a, 2021b)) lessons on Holocaust NER and Nanomi Arachchige et al.'s ((2023)) paper detailing their work on compiling and annotating an English corpus for Holocaust-related NER, which they used to train and evaluate rules-based and transformer-based (Vaswani et al. 2017) tools. Consistent with other publications (Luthra et al. 2023; Ehrmann et al. 2023), many of the Transformer-based models included in Nanomi Arachchige et al.'s ((2023)) experiments achieved high F1 scores across most of the entities considered, encouraging us to select a similar architecture for our experiments.

However, since the material processed by EHRI is diverse and multilingual, we wanted to work towards developing a single multilingual NER model that would leverage multilingual learning for cross-lingual transfer (Mueller, Andrews, and Dredze 2020; Ehrmann et al. 2023; Schweter et al. 2022; Wu et al. 2020). Multilingual NER in historical documents has seen a growing interest amongst the Digital Humanities (DH), Natural Language Processing (NLP), and cultural heritage communities (Ehrmann et al. 2023). In 2022, Ehrmann et al. (2022) introduced a shared task on NER and EL in multilingual historical documents, encouraging researchers to study approaches that can work well across different contexts and languages. Ehrmann et al. (2022) acknowledge that advances in AI thanks to the Transformer architecture and the increased availability of suitable resources create new opportunities for working towards such solutions. The same is true in the EHRI context, where since the work of Kepa Joseba Rodriguez et al. (2012) and Nikolova and Levy (2018), EHRI has produced a series of manually annotated digital scholarly editions. Although the original purpose of these editions was not to provide a dataset for training NER models, we argue that they nevertheless constitute a high-quality resource that is suitable to be used in this way. We therefore repurposed them to train multilingual Transformer-based NER models testing the hypothesis that we now have enough resources to develop a single domain-specific tool that can work reliably well across different languages and document types encountered in EHRI collections.

### 5.3.3. EHRI Online Editions

Since 2018, the EHRI Consortium has supported the development and publication of six Holocaust-related digital scholarly editions<sup>37</sup> (EHRI-Consortium 2021; Frankl and Schellenbacher 2018, 2023; Frankl et al. 2023, 2020; Garscha, Kuretsidis-Haider, and Schellenbacher 2022). Each edition enables digital access to facsimiles and transcripts of thematically related documents held by different EHRI partner institutions through a single web interface and unlocks new ways of presenting and browsing through historical sources using digital tools. Publishing a digital edition is a resource-intensive process. Notwithstanding the extensive archival research needed for selecting the documents, additional steps include

---

<sup>37</sup> At the time of writing: 2/26/2024.

transcribing and translating them and, most importantly, annotating words and phrases found within these texts and creating links with entities in controlled vocabularies provided by EHRI and third parties. Currently, this annotation is done manually by or under the supervision of subject matter experts, ensuring a high quality of annotations<sup>38</sup>. We repurposed these resources to convert them into a dataset suitable for training NER models, which we consider as a gold standard.

Each EHRI Online Edition consists of digitized documents originating from various archives that are selected, edited, and annotated by EHRI researchers using the Text Encoding Initiative (TEI) P5 standard (TEI Consortium, (2023)), an XML schema, which supports their online publication. Editions enhance the edited documents by contextualizing the information contained within them and linking them to EHRI vocabularies and descriptions, and by visualizing georeferenced entities through interactive maps. Thanks to their encoding in TEI, they are fully searchable and can be filtered using facets such as spatial locations, topics, persons, organizations, and institutions. All documents within an edition have a transcript, either in their original language, a translation, or both, and have access to their facsimile. EHRI Editions are published without a regular schedule and it is possible to update them with new material or improve the already published documents. The description of each EHRI Online Edition is available in the Appendix of this document.

### 5.3.4. The EHRI-NER Dataset

This section presents EHRI-NER, a multilingual NER dataset derived from the EHRI Online Editions. We fully released EHRI-NER on Hugging Face and GitHub<sup>39</sup>.

ISO code	Language	Tokens	PERS	LOC	ORG	DATE	GHETTO	CAMP
cs	Czech	106392	1415	2627	359	741	<b>212</b>	<b>502</b>
de	German	<b>218570</b>	<b>2516</b>	<b>3592</b>	<b>871</b>	<b>950</b>	202	396
en	English	58 405	363	1015	225	287	52	77
fr	French	2273	3	39	8	4	0	5
hu	Hungarian	24686	157	304	148	97	2	114
nl	Dutch	1991	17	25	33	7	0	2
pl	Polish	18385	221	328	54	126	17	51
sk	Slovak	3550	30	158	11	21	0	0
yi	Yiddish	71506	629	1311	158	4	43	82
/	All	505758	5351	9399	1867	2237	528	1229

Table 1: EHRI-NER Dataset: tokens and entity classes distribution

### Languages and Subsets

We sorted all TEI XML files available from the EHRI Online Editions by language. The resulting EHRI-NER dataset includes nine languages: Czech (cs), German (de), English (en), French (fr), Hungarian (hu), Dutch (nl), Polish (pl), Slovak (sk), and Yiddish (yi). We created a subset for each language since they are not represented in the same proportion.

<sup>38</sup> More info about this process can be found on the website of each edition.

<sup>39</sup> See the EHRI-NER organization on Hugging Face to access the model and dataset and the EHRI-NER GitHub repository to access the dataset subsets per language.

As noted earlier, the dataset includes official reports, correspondences, diplomatic notes, newspaper reports, and testimonies. The creation dates of the documents span from 1936 to 2001.

### From TEI XML to the IOB Format

To build the subsets, we created a Python script to parse the TEI XML documents and convert them to the CoNLL Inside-Outside-Beginning (IOB) format (Sang and De Meulder 2003), which is typical for NER datasets (Ehrmann, Nouvel, and Rosset 2016)<sup>40</sup>.

The BF, UH, DC, and ET editions all include translations of some of their original transcribed documents. To avoid contaminating our validation and test sets, we filtered them out. Additionally, both the BF and the UH editions contain some documents that overlap. We also filtered these out to avoid having duplicates in our dataset.

### Entity Classes

Given that the primary purpose of this work is to enhance the services and facilitate the work of EHRI stakeholders, we used a custom typology of entity classes that corresponds better to how we envision deploying this tool in the EHRI environment, extending the CoNLL typology (Sang and De Meulder 2003) to include classes such as camps and ghettos, which correspond to custom EHRI vocabularies used when annotating Holocaust materials to produce new EHRI Editions. However, our typology is coarser compared to more fine-grained typologies found in similar work (Nanomi Arachchige et al. 2023). We extracted all TEI elements `<persName>`, `<placeName>`, `<orgName>`, and `<date>` from the selected TEI XML files. The `<placeName>` element sometimes includes an attribute `@type` to indicate whether it is referencing a concentration camp or a ghetto. We distinguish between `<placeName>`, `<placeName type="camp">`, and `<placeName type="ghetto">` to include fine-grain camp and ghetto entities in addition to the coarse-grain location entity. The conversion table is presented in Table 1.

EHRI TEI XML files also contain the `<term>` element, used for annotating various subjects related to the Holocaust and for linking them with their associated entries in the EHRI vocabulary of terms<sup>41</sup>. However, we have chosen to consider these instances as non-entity tokens, as their broad coverage of themes, their variability, and lack of semantic regularity in how they are used in annotations make them unsuitable in a token classification context. Had we included them in our typology, we hypothesize that the NER models would tag a disproportionate number of tokens as terms, rendering the output noisy and confusing. Instead, EHRI is working on a different solution for extracting subject metadata, which is outside the scope of this section.

The EHRI-NER dataset includes a total of 505758 tokens, with 5351 person entities, 9399 location entities, 1867 organization entities, 2237 date entities, 528 ghetto entities, and 1229 camp entities. The distribution of tokens and entity classes is detailed in Table 1.

---

<sup>40</sup> Our script is available on GitHub.

<sup>41</sup> See the EHRI Terms database. Accessed 2/27/2024.



TEI XML Element	Entity Class
<persName>Helene Hirsch</persName>	Person
<placeName>Berlin</placeName>	Location
<orgName>Gestapo</orgName>	Organization
<date when="1937-10">Oct. 1937</date>	Date
<placeName type="camp">Auschwitz</placeName>	Camp
<placeName type="ghetto">getcie</placeName>	Ghetto

Table 2: Conversion table for TEI XML Elements and Entity Classes.

### Data Format and Preprocessing

We chose to convert TEI annotations and non-entity tokens into the CoNLL IOB format, as presented in Sang and De Meulder (2003) (see Table 2). The IOB format ensures that our dataset is interoperable with common NER tools. Each token and its annotation have been put on a separate line and there is an empty line after each sentence, as shown in the following example:

```
Von O
Gross B-CAMP
- I-CAMP
Rosen I-CAMP
Bahntransport O
nach O
Buchenwald B-CAMP
. O
```

Each language subset has been tokenized at the sentence and word levels. We used SpaCy (Honnibal et al. 2020) and its multi-language pipeline to process each subset<sup>42</sup>.

Entity	Example	Annotation
Person	Kurt Lichtenstern	B-PERS, I-PERS
Location	Moravská Ostrava	B-LOC, I-LOC
Organization	Pártfogó iroda	B-ORG, I-ORG
Date	1941 roku	B-DATE, I-DATE
Ghetto	getta łódzkiego	B-GHETTO, I-GHETTO
Camp	Auschwitz camp	B-CAMP, I-CAMP

Table 3: Entity types illustrated with examples and IOB tagging.

### 5.3.5. Experimental Setup

Two experiments were conducted to determine whether the dataset was sufficiently large for fine-tuning a reliable NER model that could be used in a real-life setting, e.g. speeding up named entity annotation when curating a new EHRI Online Edition. The multilingual aspect of our dataset was also leveraged to test XLM-RoBERTa (XLM-R) (Conneau et al. 2020) in a

<sup>42</sup> See the multi-language pipelines available on SpaCy website. Accessed 2/27/2024.

low-resource setting, as our dataset is significantly smaller than, for instance, the CONLL2003 NER dataset used for evaluating this model on a token classification task. In the following, we describe the model that we used for fine-tuning, the experiments we conducted on the dataset, and their results.

## Model

We chose to experiment with the multilingual Transformer-based masked language model XLM-RoBERTa-large (Conneau et al. 2020) as it demonstrates high efficacy in multilingual settings and strong cross-lingual transfer capabilities, especially on token classification tasks, without sacrificing per-language performance<sup>43</sup>. According to Nanomi Arachchige et al. (2023), this model outperforms the multilingual hmbERT (Schweter et al. 2022) model which was pre-trained on German, French, Swedish, Finnish, and English historical newspapers (thus not pre-trained in all of the languages present in our dataset). It is important to note that XLM-R has seen all languages represented in the EHRI-NER dataset during its pre-training.

The same fine-tuning parameters were kept for all our experiments. The learning rate is set at  $3e^{-5}$ , the number of epochs for training at 3 to avoid overfitting, the weight decay at 0.01, and the train and evaluation batch size at 16.

## Experiments

**Experiment 1:** We fine-tuned XLM-R on all subsets (cs, de, en, fr, hu, nl, pl, sk, yi) to evaluate the overall performance of the model on a multilingual level. Instead of relying on a simple shuffle, and to ensure that all languages are represented in the train, validation, and test set, we first split each subset into train (80%), validation (10%), and test (10%) sets, using a seed of 42 for result reproducibility<sup>44</sup>. Each split subset is then concatenated and the final dataset is used for fine-tuning. Our objective was to acquire a single fine-tuned model with reliably good performance across most if not all languages, suitable primarily as part of an editorial pipeline that streamlines the creation of new digital scholarly editions related to the Holocaust.

**Experiment 2:** To assess the cross-lingual capabilities of XLM-R in a low-resource setting, we fine-tuned it two more times—each time leaving out one language subset which was reserved for testing. Our chosen target languages were nl (experiment 2.1) and yi (experiment 2.2) as they represent some of the smallest subsets, while still containing enough examples for meaningful evaluation. For each target language, we fine-tuned XLM-R on every other subset split into train and validation (80% / 20%) and used the entire subset of the target language as the test set. This experiment sought to simulate a scenario where we would need to use the fine-tuned model to pre-annotate documents from a Holocaust domain but in a language not seen by our model during fine-tuning. The fine-tuning processes were repeated three times for each experiment. Followed by the computation of the average of each of the three runs to obtain a reliable evaluation.

## Evaluation

**Experiment 1** yielded a consistent and satisfying overall performance across the validation and the test sets, with an overall F1 score of 81.3% for the former and 81.5% for the

---

<sup>43</sup> See XLM-RoBERTa-large model card on Hugging Face website. Accessed 02/27/2024.

<sup>44</sup> The mentioned seed used for splitting the subsets was used for all the experiments.

latter (Table [\[tab:score\\_all\]](#)), achieving higher scores compared to earlier EHRI NER work, surpassing Rodriguez et al.'s (2012) maximum total F1 score of 60% while additionally tagging domain-specific entities and exceeding the F1 score of 77% reported for the person tagger (Leeuw et al. 2018). Domain-specific entities (Camp, Ghetto) are also consistently classified by the model. Only the Organization entity demonstrated poor F1 scores probably caused by the relatively low number of examples (1867 in total). This behavior has been previously observed by Kepa Joseba Rodriguez et al. (2012). The overall evaluations for cs, de, en, hu, pl, and yi test sets showed that the performance of the fine-tuned model is corollary to the number of examples in the training set. However, even though we see a decrease in the F1 scores depending on the size of the subset (minimum 73.2% overall F1 score for the hu test set), we still consider the performance of the fine-tuned model strong considering the relatively small size of some of the subsets.

The confusion matrix for predicted classes in the test set (Fig. 10) shows instances where the fine-tuned model occasionally misclassifies entities as non-entity tokens, I-GHETTO being the most confused entity. The fine-tuned model occasionally encounters challenges in extracting multi-tokens entities, such as I-CAMP, I-LOC, and I-ORG, which are sometimes confused with the beginning of an entity. Moreover, it tends to misclassify B-GHETTO and B-CAMP as B-LOC, which is not surprising given that they are semantically close and there are cases where even an expert would hesitate to pick a single label. Indeed, sometimes an entity such as the camp/ghetto "Theresienstadt" could be assigned any of these classes without introducing errors<sup>45</sup>.

Overall, these scores are high enough to at least pre-annotate Holocaust-related textual documents when developing a new EHRI Online Edition or when wanting to enrich an archival description with access points that an archivist can verify. Additionally, as long as the new unseen texts to be fed into the model belong to a similar domain and period, it can be assumed that the scores will remain relatively consistent across all nine languages used for fine-tuning. The fine-tuned XLM-R model was released on Hugging Face<sup>46</sup>.

---

<sup>45</sup> See more about the function of Theresienstadt here. Accessed 2/27/2024.

<sup>46</sup> See the EHRI-NER fine-tuned XLM-R model on Hugging Face.



Entity	Validation Set				Test Set			
	Acc. (%)	Prec. (%)	Rec. (%)	F1 (%)	Acc. (%)	Prec. (%)	Rec. (%)	F1 (%)
Person	/	<b>85</b>	90.3	87.5	/	83.8	<b>88.7</b>	<b>86.2</b>
Location	/	78.1	86	81.8	/	78.1	87.3	82.5
Organization	/	62.3	56.8	59.4	/	61.9	60.7	61.3
Date	/	81.5	<b>92.9</b>	<b>86.8</b>	/	81.1	90.3	85.4
Camp	/	76.4	68.7	72.3	/	73	72.7	72.8
Ghetto	/	75.2	75.2	75.2	/	<b>87.1</b>	80.7	83.7
<b>Overall</b>	98	78.9	83.9	81.3	98	78.6	84.7	81.5
<b>Overall - CS test set</b>	/	/	/	/	98.3	82.5	87.1	84.7
<b>Overall - DE test set</b>	/	/	/	/	98.6	78	86.6	82.1
<b>Overall - EN test set</b>	/	/	/	/	98	75.4	84.4	79.6
<b>Overall - HU test set</b>	/	/	/	/	98.5	71.9	74.6	73.2
<b>Overall - PL test set</b>	/	/	/	/	97.2	73.3	77.7	75.5
<b>Overall - YI test set</b>	/	/	/	/	98.5	75.6	78.8	77.2

Table 4: Evaluation of fine-tuned XLM-R on EHRI-NER on all languages, by entity type (experiment 1), and specific overall evaluation on cs, de, en, hu, pl, and yi test sets. fr, nl, and sk test sets were omitted because of a lack of examples.

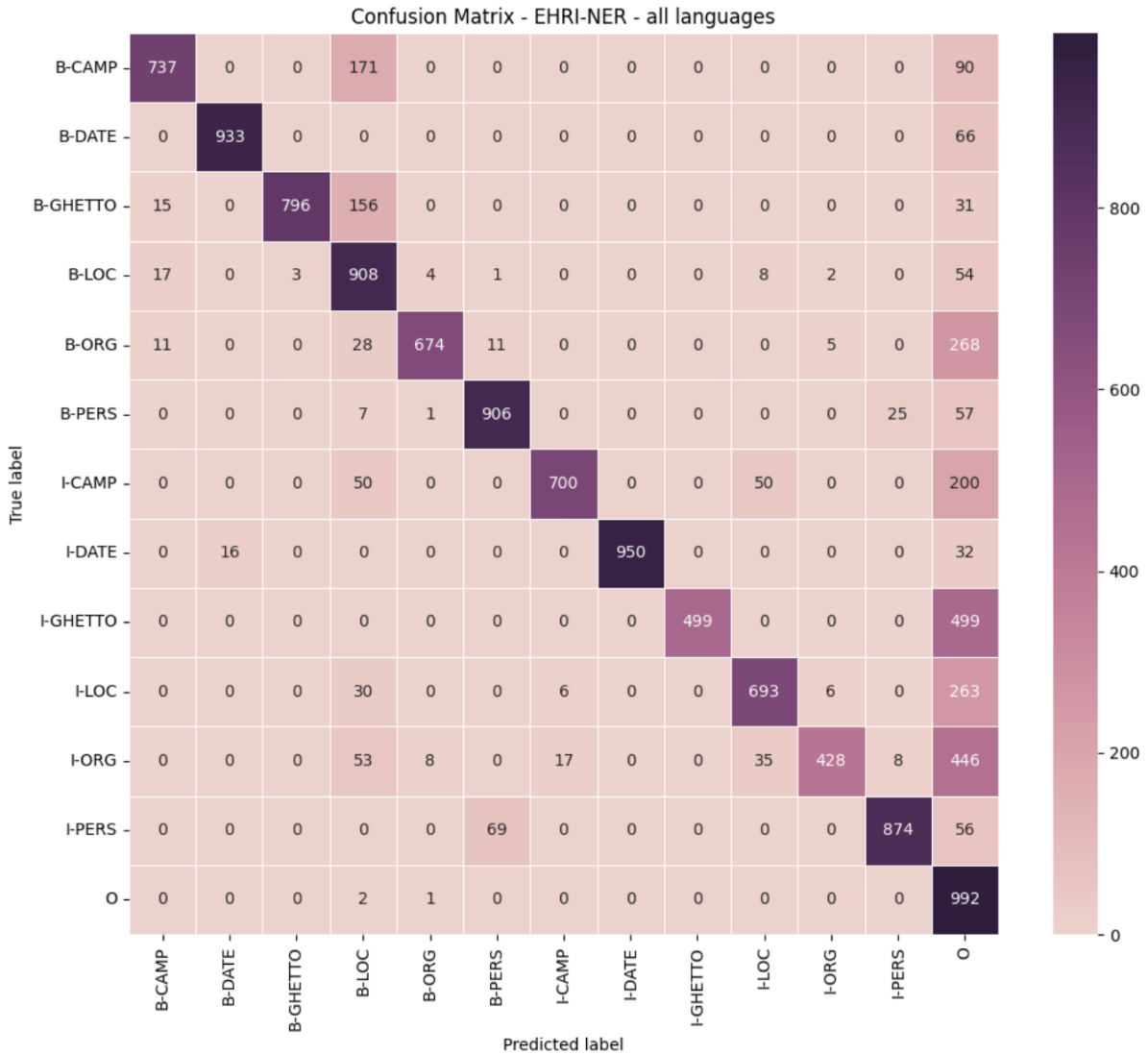


Figure 10: Matrix confusion for predicted classes in the test set, when fine-tuning XLM-R on all languages (experiment 1). The confusion matrix was normalized using a scaling factor of 1000.

**Experiment 2** revealed that we can leverage the cross-lingual capabilities of XLM-R depending to some extent on how much data it has seen about a specific language during its pre-training and on how many examples the training dataset has.

**Experiment 2.1** showed unexpectedly high performance, about 94% overall F1 score, in one of the runs on the Dutch subset. However, it decreased in the second run to around 80% F1 score. After the third run, the overall F1 score of 84.6% proved that the fine-tuned model achieved satisfying performance, except for the classification of Organization entities (see Table 3), and despite not being evaluated on the Ghetto entity due to lack of examples. The confusion matrix shows that the fine-tuned model has trouble extracting multi-token entities, as noted in experiment 1 (Fig. 11).

Entity	Acc. (%)	Prec. (%)	Rec. (%)	F1 (%)
Person	/	100	96	97.9
Location	/	83.2	96	89
Organization	/	76.1	61.6	67.5
Date	/	100	100	100
Camp	/	100	100	100
Ghetto	/	/	/	/
<b>Overall</b>	<b>98.7</b>	<b>86.4</b>	<b>82.9</b>	<b>84.6</b>

Table 5: Evaluation of XLM-R on the nl subset, when fine-tuned on all languages except nl, by entity type (experiment 2.1).



Figure 11: Evaluation of XLM-R on the nl subset, when fine-tuned on all languages except nl, by entity type (experiment 2.1).

**Experiment 2.2** on Yiddish yielded poor performance, with an overall F1 score of 46.5% (Table 4). Only Person and Location entities showed an F1 score of above or equal to 50%. The Organization entity and the domain-specific entities Date, Camp, and Ghettos are all under 10% F1 score, the latter having an F1 score of 0. As depicted in Fig. 12, the model mainly misclassified entities as non-entity tokens, which is a common problem in NER (Luthra et al. 2023).

The fluctuations in performance are probably related to the pre-training of XLM-R. As reported in Conneau et al. (2020), the model was pre-trained on 5025M tokens for Dutch, but merely 34M tokens for Yiddish. Therefore, we can hypothesize that the performance of XLM-R on the Yiddish subset is likely due to the limitations in its representation of this language after pre-training. This may have impacted the fine-tuning of the model and its cross-lingual capabilities for a token classification task on a small subset, such as the Yiddish subset, whereas the fine-tuning on the Dutch subset, despite being smaller, achieved a good performance. Other work on the zero-shot language transfer capabilities of multilingual Transformer models supports this hypothesis (Lauscher et al. 2020). Since the authors do not understand Yiddish, a comprehensive error analysis was not possible. However, it is worth noting that the challenges observed, as shown in experiment 1, can be mitigated when fine-tuning XLM-R on all subsets.

This experiment also confirms the hypothesis we made when discussing the lack of examples for the Organization entity and its consequence on the results in experiment 1.

<b>Entity</b>	<b>Acc.</b> (%)	<b>Prec.</b> (%)	<b>Rec.</b> (%)	<b>F1</b> (%)
Person	/	68.9	53.1	59.9
Location	/	48.4	52.2	50
Organization	/	21.3	04.8	07.6
Date	/	00.7	41.6	01.3
Camp	/	28.4	02	03.7
Ghetto	/	0	0	0
<b>Overall</b>	<b>96.6</b>	<b>47.2</b>	<b>46.2</b>	<b>46.5</b>

Table 6: Evaluation of XLM-R on the yi subset, when fine-tuned on all languages except yi, by entity type (experiment 2.2).

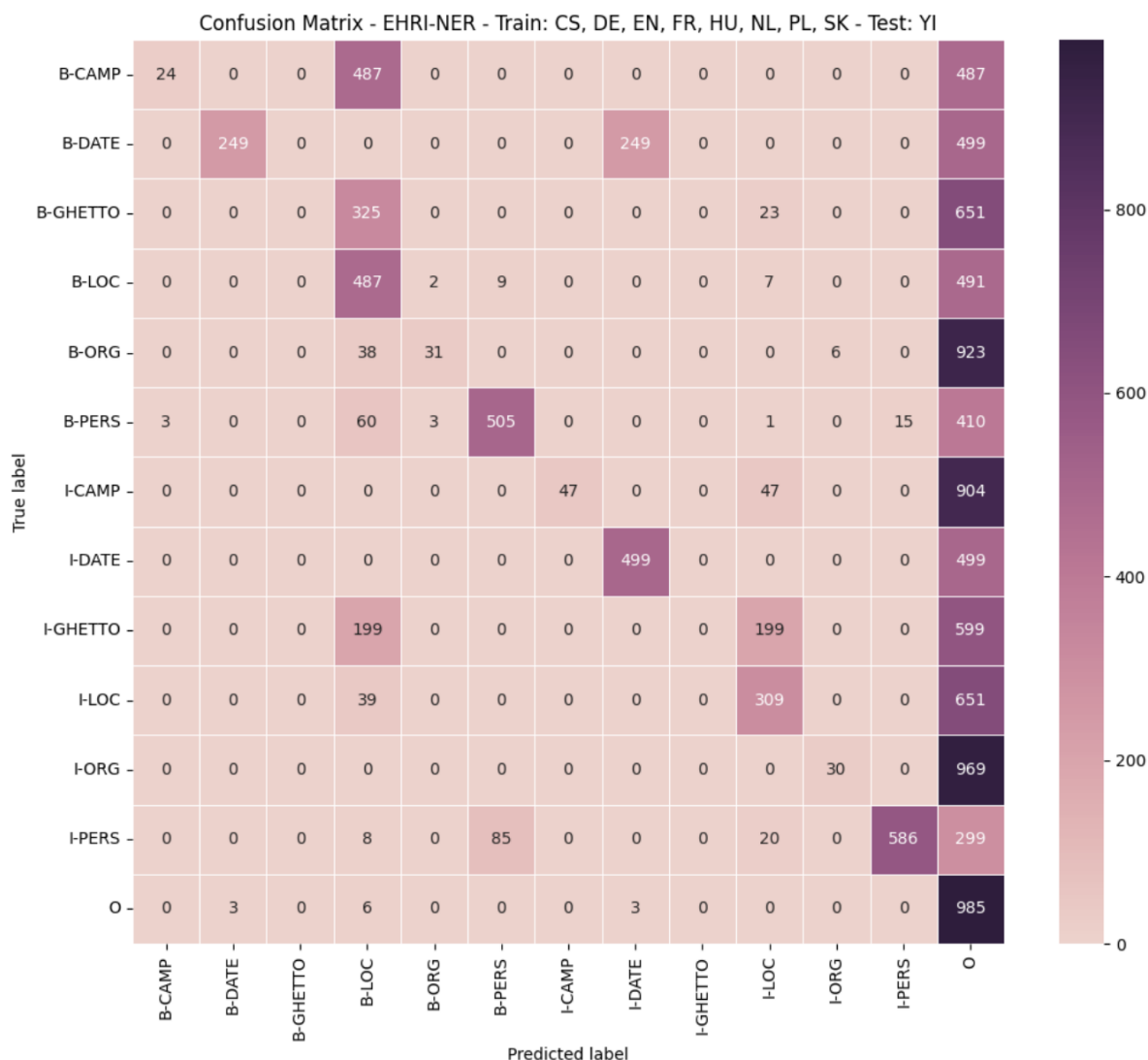


Figure 12: Matrix confusion for predicted classes in the yi subset, when fine-tuning XLM-R on all languages except yi (experiment 2.2). The confusion matrix was normalized using a scaling factor of 1000.

### 5.3.6. Conclusion

We released EHRI-NER, a multilingual dataset for NER in Holocaust-related textual documents, built from the numerous TEI XML files made available across all EHRI Online Editions. We also evaluated the multilingual and cross-lingual capabilities of XLM-R by fine-tuning it on our dataset and proved that it can perform well when using relatively small domain-specific datasets. We also provided a baseline for future evaluations of NER systems on the dataset. Our future objective for the dataset is to include the translations mentioned previously while filtering them out from the train set. They indeed represent a sizable portion of data that would increase the number of examples in our dataset and could potentially lead to an increase in the fine-tuned model’s performance.

The next steps should be to experiment on multilingual named entity disambiguation, which would allow us to automatically link recognized entities with IDs in the EHRI vocabularies mentioned in the introduction (1). Another idea for future work could be to source similar annotated datasets and merge them with EHRI-NER. Besides, it is encouraged that EHRI

partners evaluate the model qualitatively as part of their work and provide feedback. Based on that feedback, it will be possible to improve the model and deploy it as part of EHRI's cataloging and editorial pipelines. Another interesting course for future work would be to create a stable annotation typology for Holocaust documents with the help of experts. Finally, it should be possible in the near future to provide a more complete baseline by experimenting with more multilingual Large Language Models (LLMs), including state-of-the-art LLMs for zero-shot and few-shot NER.

## 6. Publishing encoded sources

### 6.1. The EHRI Omeka plugin

EHRI employs Omeka to publish its edited TEI files, utilizing the Neatline mapping plugin to enhance the presentation of geographic data. However, as Omeka was not originally designed to display TEI files or extract structured information from XML sources, the EHRI team developed a custom Omeka plugin to fulfill these needs.

According to the EHRI documentation, this plugin<sup>47</sup> enhances the headers of TEI documents with metadata from the EHRI Portal and Geonames. Additionally, it enables the creation of Omeka items from uploaded TEI files, with the metadata elements in Omeka being populated through customizable XPath mappings. This plugin also supports the association of images and other files with these items, and facilitates the creation of Neatline exhibits using location data and other metadata from the TEI headers.

The plugin's functionalities extend to the ingestion, updating, and association of tertiary files with TEI-based Omeka items. It also allows for the exportation of TEI data and associated files, and the configuration of XPath-to-Omeka field mappings. Despite these advanced features, a notable observation is that EHRI's online editions currently lack the ability to download all TEI XML files from a given edition, which could be a valuable addition for users needing comprehensive access to the data.

### 6.2. TEI Publisher: toward a new publishing interface for EHRI Online Editions?

We suggest the creation of a dedicated TEI Publisher instance where all existing EHRI digital editions and those that will be created in the future will be published and centralized.

TEI Publisher is an open source and full stack application, fully customizable, that can be deployed online with a dedicated server. It is based on an eXist-db XML database, and was designed for publishing TEI XML files<sup>48</sup>.

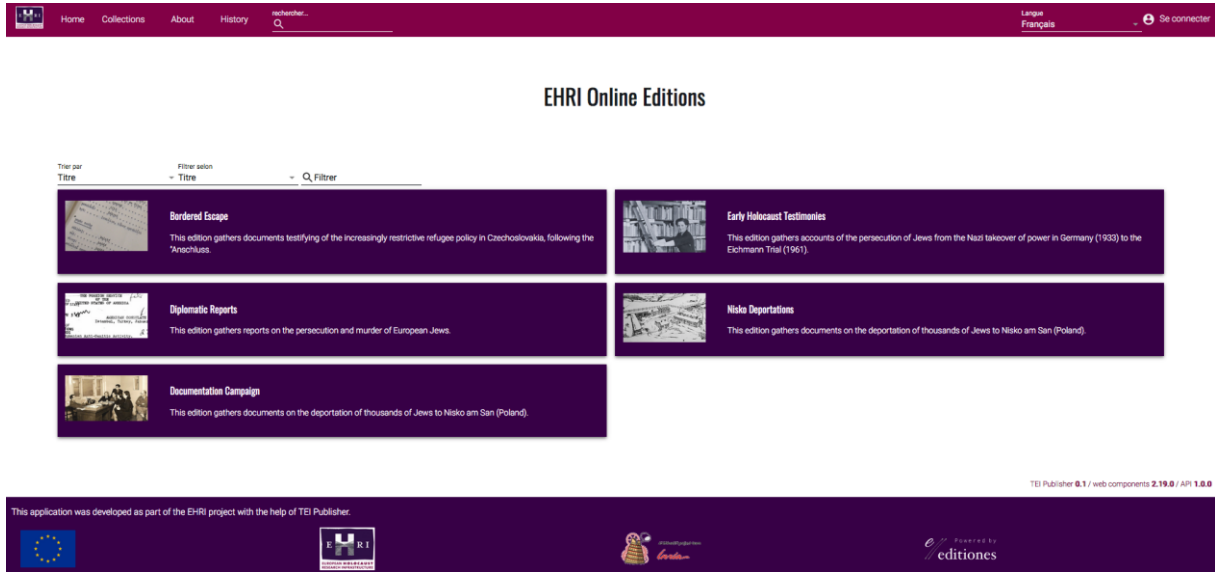
TEI Publisher would improve EHRI Online Editions design by offering an horizontal, panel-based display, enhancing the presentation of maps, indexes, and reading aids compared to Omeka's vertical display. It supports complex layouts and contextual information. The XML backend allows quick bug fixes and new feature development, though a thorough comparison between the existing Omeka solution and the XML database should be analyzed.

---

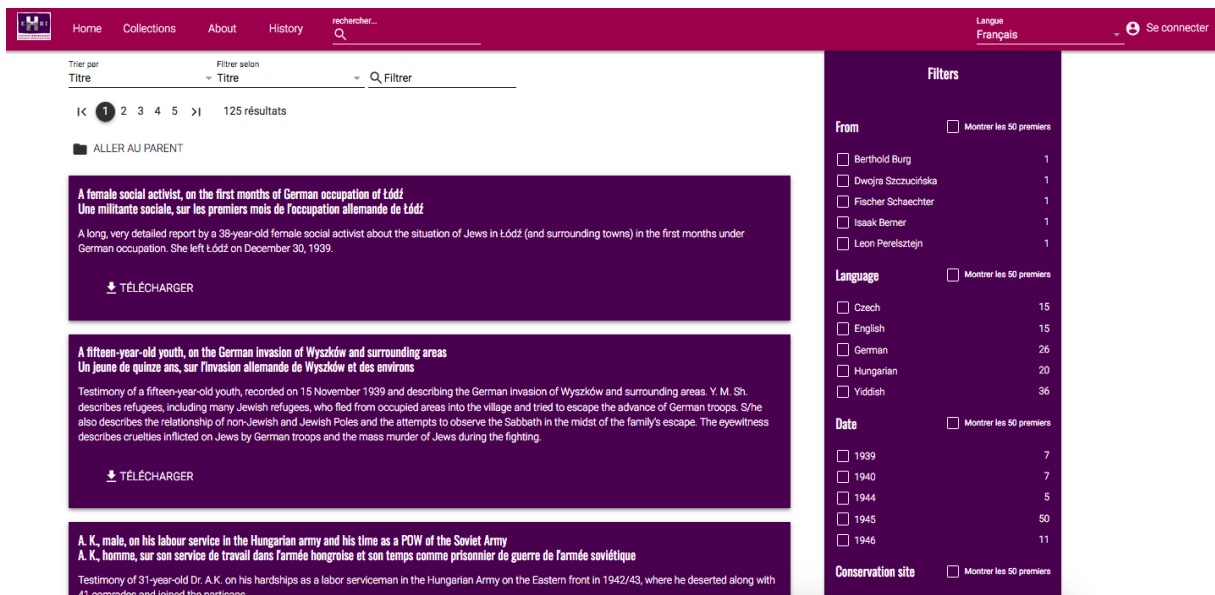
<sup>47</sup> The plugin is available in the [EHRI documentation](#) and on [GitHub](#). Accessed June 2024.

<sup>48</sup> See TEI Publisher [website](#). Accessed June 2024.

The ALMAnaCH team is working on an EHRI TEI Publisher proof-of-concept. While the proof-of-concept still has issues with the functioning implementation of new features, the current version was positively received at the EHRI online meeting of July 2023 on online editions, and at the EHRI academic conference of June 2024 in Warsaw, Poland. Concerns remain about integrating current Omeka plugins and whether TEI Publisher can meet all existing needs. User feedback through polls could help evaluate the new design's usefulness.



Homepage of the TEI Publisher EHRI application



Early Holocaust Testimonies collection



Home Collections About History

Langue Français Se connecter

Trier par Titre Filtrer selon Titre

1 3 résultats

ALLER AU PARENT

**A fifteen-year-old youth, on the German invasion of Wyszków and surrounding areas**  
**Un jeune de quinze ans, sur l'invasion allemande de Wyszków et des environs**

Testimony of a fifteen-year-old youth, recorded on 15 November 1939 and describing the German invasion of Wyszków and surrounding areas. Y. M. Sh. describes refugees, including many Jewish refugees, who fled from occupied areas into the village and tried to escape the advance of German troops. S/he also describes the relationship of non-Jewish and Jewish Poles and the attempts to observe the Sabbath in the midst of the family's escape. The eyewitness describes cruelties inflicted on Jews by German troops and the mass murder of Jews during the fighting.

TÉLÉCHARGER

**Leyb Blumberg, in hiding in Warsaw then escape to Vilnius in October 1939**  
**Leyb Blumberg, caché à Varsovie puis évadé à Vilnius en octobre 1939**

Brief testimony of Leyb Blumberg, recorded on 16 November 1939 and describing his flight from Warsaw to Vilnius in October 1939, during which he and his fellow travelers were not bothered by German troops. They reached Vilnius without difficulty.

TÉLÉCHARGER

**Shlome Perkal, on bombings of Międzyrzec Podlaski, and movement of Jewish refugees toward Chełm and Piaski**  
**Shlome Perkal, sur les bombardements de Międzyrzec Podlaski et le mouvement de réfugiés juifs vers Chełm et Piaski**

Testimony of Shlome Perkal, recorded on 12 November 1939 and describing his experiences in Międzyrzec Podlaski at the time of the German invasion of

**Filters**

**Language**

Yiddish 3

**Date**

1939 3

11 3

12 1

15 1

16 1

**Conservation site**

The Wiener Library for the Study of the Holocaust and Genocide 3

**Place**

Białystok 1

Międzyrzec 1

Przasnysz 1

Faceted search options

Home Collections About History

Langue Français Se connecter

TABLE OF CONTENTS METADATA DISPLAY CITATION LICENCE NEXT PAGE

View Original version

קאמיטע צו זאמלען מאטעריאלן

וועגן יידישן חורבן אין פוילן 1939

י.מ.ש.

ווירשאָו

יאר אָלט, ביי די עלטערן 18

פראָטאָקאל נומער 3

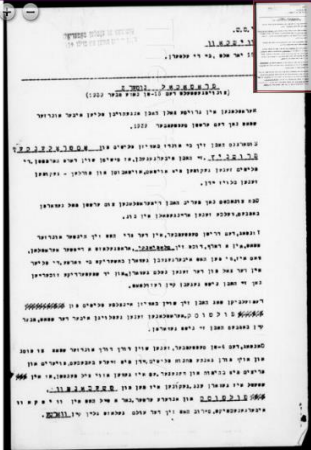
(צונויפגעשטעלט דעם 15טן נאוועמבער 1939)

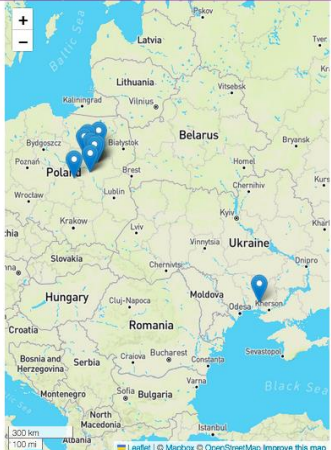
אָטאָפּאָלענע אין גרויסע צאָלן האָבן אָנגעהויבן פֿליען איבער אונדזער שטאַט נאָך דעם ערשטן סעפטעמבער 1939.

צומאָרגעס האָבן זיי ביי אונדז באַוויזן פֿלייסן פֿון אָטאָפּאָלענע, ווירשניץ, זיי האָבן איבערגעגעבן, אָן שײַסן שוין דאָרט האַראַמאָטן. די פֿלייסן זענען געקומען מיט אויטאָס, אויטאָבוסן און פֿורלעך – געקומען זענען בלויז יידן.

שבת צו נאכטס נאָך מעריב האָבן די אָטאָפּאָלענע צום ערשטן מאל געוואָרפֿן באַמבעס, וועלכע זענען אָריינגעפֿאלן אין בוג.

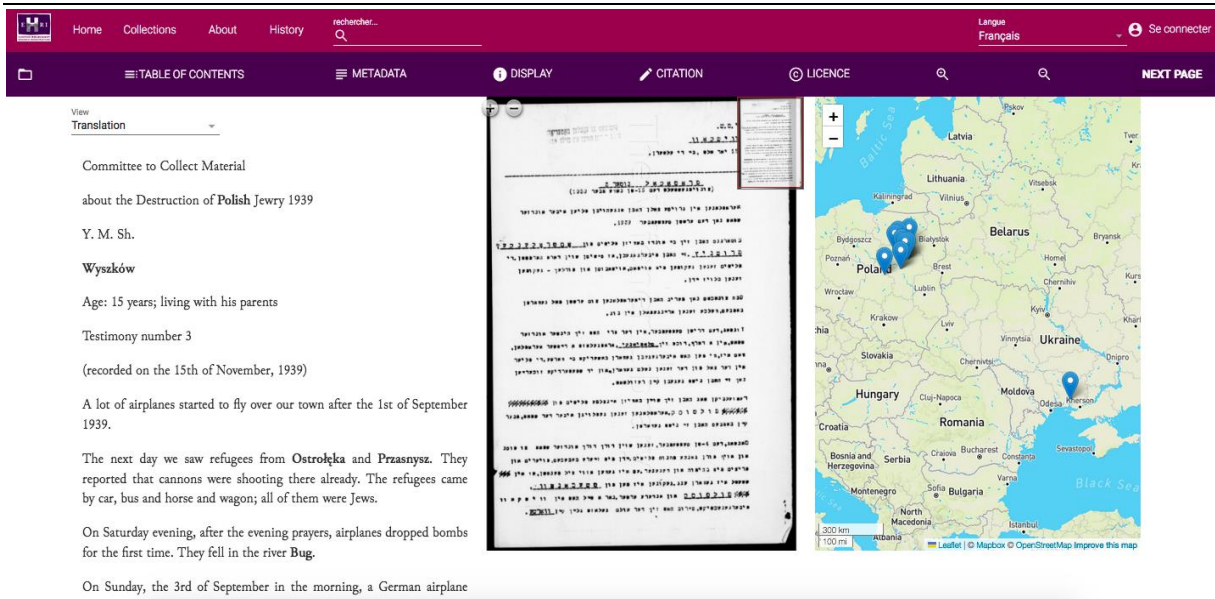
זונאָך, דעם דריטן סעפטעמבער, אין דער פֿרי האָט זיך הינטער אונדזער שטאַט, אין





Original text with the facsimile and the map





Home Collections About History recherche... Langues Français Se connecter

TABLE OF CONTENTS METADATA DISPLAY CITATION LICENCE NEXT PAGE

View Translation

Committee to Collect Material

about the Destruction of Polish Jewry 1939

Y. M. Sh.

Wyszów

Age: 15 years; living with his parents

Testimony number 3

(recorded on the 15th of November, 1939)

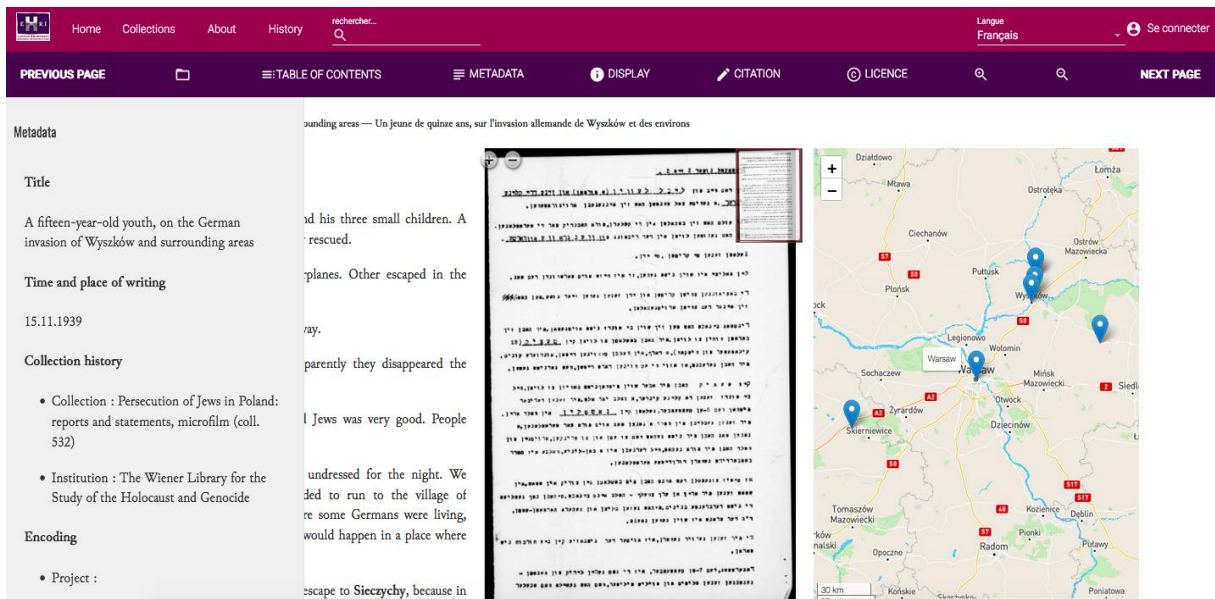
A lot of airplanes started to fly over our town after the 1st of September 1939.

The next day we saw refugees from Ostrołęka and Przasnysz. They reported that cannons were shooting there already. The refugees came by car, bus and horse and wagon; all of them were Jews.

On Saturday evening, after the evening prayers, airplanes dropped bombs for the first time. They fell in the river Bug.

On Sunday, the 3rd of September in the morning, a German airplane

Translated text with the facsimile and the map



Home Collections About History recherche... Langues Français Se connecter

PREVIOUS PAGE TABLE OF CONTENTS METADATA DISPLAY CITATION LICENCE NEXT PAGE

Metadata

Title

A fifteen-year-old youth, on the German invasion of Wyszów and surrounding areas

Time and place of writing

15.11.1939

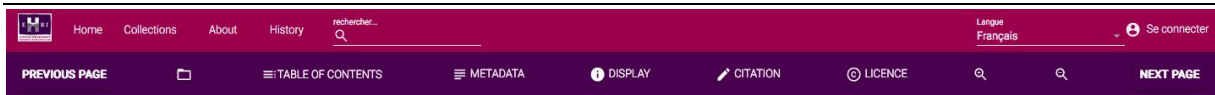
Collection history

- Collection : Persecution of Jews in Poland: reports and statements, microfilm (coll. 532)
- Institution : The Wiener Library for the Study of the Holocaust and Genocide

Encoding

- Project :

Display of the metadata on the TEI Publisher application



A fifteen-year-old youth, on the German invasion of Wyszków and surrounding areas — Un jeune de quinze ans, sur l'invasion allemande de Wyszków et des environs

View  
Translation

the wife of **Leybl Levin** (a coachman) and his three small children. A certain **Levin, Leybl** successfully rescued.

People **Coachman from Wyszków** of the airplanes. Other escaped in the directions of **węgrów and warsaw**.

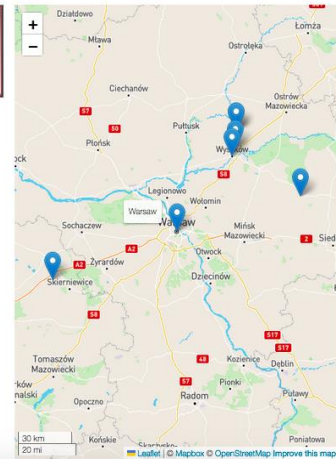
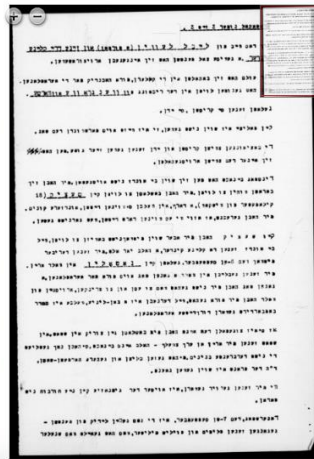
Both Christians and Jews were running away.

There were no policemen any more, apparently they disappeared the same day.

The relationship between Christians and Jews was very good. People were helping each other.

Tuesday in the evening we did not get undressed for the night. We discussed to where to escape. We decided to run to the village of **Sieczychy** (18 km from **Wyszków**), where some Germans were living, our customers. We believed that nothing would happen in a place where Germans were living.

But on Wednesday we did not manage to escape to **Sieczychy**, because in our family there are small children, six months old. So on Wednesday,



Display of a named entity on the TEI Publisher application



History of archival collections of Early Holocaust Testimonies

- The Ball-Kaduri Testimony Collection (Yad Vashem)
- Documentation Campaign in Prague
- Testimonies Collection in the Jewish Historical Institute in Warsaw
- The Konluchowsky Testimony Collection (Yad Vashem)
- National Committee for Attending Deportees (DEGOB, Hungary)
- The Wiener Library and Its Eyewitness Accounts
- The Central Historical Commission of the Central Committee of Liberated Jews in the US Zone in Munich – Testimonies Collected in DP Camps (Yad Vashem)

History of archival collections of Early Holocaust Testimonies

The Ball-Kaduri Testimony Collection (Yad Vashem)

Dr. Kurt Ball-Kaduri was born in Berlin in 1891. A lawyer and legal adviser to the Prussian government, he was also active in Jewish affairs. He made Aliyah to Eretz Israel in December 1938. Ball-Kaduri, who was active in collecting material and writing about German Jewry, became aware that much material that reached the archives regarding Jewish life in Germany from 1933 to 1945 was incomplete and that there were large information gaps.

He decided to gather testimonies of people involved in Jewish life and the activities of Jewish organizations. In 1943, Ball-Kaduri began to collect the information and actually established his [collection](#). He contacted various people in Eretz Israel whom he knew, asking them to write their recollections and interviewing some of them himself. In 1955 he handed the [collection](#) over to [Yad Vashem](#) while continuing to collect related documentation for [Yad Vashem](#) until 1960.

The [collection](#) includes testimonies of Jewish leaders in various areas of Jewish life in Germany. Although it includes significant documentation regarding the fate of individual victims of the Holocaust, the main emphasis of the [Record Group](#) is on the different Jewish organizations.

There are over 300 files in the [record group](#). Most of the collection is written in German and about half of the testimonies have been translated into Hebrew.

Documentation Campaign in Prague

[The Jewish Museum in Prague](#), whose history is intrinsically intertwined with the persecution of Bohemian and Moravian Jews, has been collecting archival sources relating to the persecution and genocide of Jews in the Czech lands since the end of WWII. It holds various types of materials, including interviews with and witness accounts of Shoah survivors.

The testimonies presented within the EHRI online edition were gathered mainly in the framework of the so-called 'documentation campaign' (Dokumentáčnická akce). This was one of the earliest postwar projects to document the events of the Shoah, collecting evidence, documents, and witness testimonies. The founder and a driving force behind the campaign was Zeev Scheck, a prewar Zionist and survivor of Theresienstadt and Auschwitz, who emigrated Palestine in 1946 to. He later worked as an Israeli diplomat and was an initiator of the Association of Theresienstadt Prisoners which built the [Beit Theresienstadt archive and museum](#).

Taking inspiration from his wartime clandestine documentation in Theresienstadt and from a visit to Budapest after liberation, Scheck and a few of his former fellow prisoners initiated a Czechoslovak Jewish documentation effort. Scheck was thereby continuing the clandestine collection of documents in the Theresienstadt ghetto in which he and a group of Zionist youth activists had been involved. After liberation, Scheck's partner and future wife transferred his Theresienstadt collection to Prague, later moving it to Palestine. Today it forms the basis of the Theresienstadt documentation in the [Yad Vashem Archives](#).

HTML page of the “History of archival collections of Early Holocaust Testimonies”

## Conclusion

EHRI connects archival institutions with a two-fold solution: The EHRI portal, with EAD finding aids and EHRI Online Editions. They act as bridges between them and researchers, and further the history of the Holocaust.

Standards are crucial at each step of the creation of digital resources ensuring that they remain FAIR. We highly recommend the following standards:

- XML EAD for finding aids,
- ALTO XML or PAGE XML for documents' transcriptions,

- XML TEI for documents' editions and an ODD for ensuring consistency across all online editions,
- A controlled named entity schema for semantically markup entities in TEI files, and the creation of IOB NER training datasets for sharing training data to the research community.

## Acknowledgements

The work described in this report has benefited from the support of Mike Bryant, Lydia Nishimwe, and Armel Randy Zebaze, Maria Dermentzi, Michal Frankl, Aneta PlzÁková, Wolfgang Schellenbacher, and Magdalena Sedlická. We thank them for their guidance and helpful discussions. The work described herein was made possible thanks to the previous work of the editors and contributors of the EHRI Online Editions, including the annotators, the people who produced digital facsimiles of the original archival material, and those who created the transcripts and translations.

## Bibliographical References

Alexiev, Vladimir, Ivelina Nikolova, and Neli Hateva. 2019. "Semantic Archive Integration for Holocaust Research." *Umanistica Digitale* 1 (4): 131–75. <https://doi.org/10.6092/issn.2532-8816/9049>.

Bauman, Syd. 2011. "Interchange vs. Interoperability." In *Proceedings of Balisage: The Markup Conference 2011*. <https://doi.org/10.4242/BalisageVol7.Bauman01>.

Blanke, Tobias, Michael Bryant, Michal Frankl, Conny Kristel, Reto Speck, Veerle Vanden Daelen, and René Van Horik. 2017. "The European Holocaust Research Infrastructure Portal." *Journal on Computing and Cultural Heritage* 10 (1): 1–18. <https://doi.org/10.1145/3004457>.

Burnard, Lou. 2014. *What is the Text Encoding Initiative?: How to Add Intelligent Markup to Digital Resources*. OpenEdition Press. <https://doi.org/10.4000/books.oep.426>.

Burnard, Lou. 2016. *ODD Chaining for Beginners*. GitHub. <http://teic.github.io/PDF/howtoChain.pdf>.

Burnard, Lou. 2018. "What is TEI Conformance, and Why Should You Care?" *Journal of the Text Encoding Initiative* 1 (12). <https://doi.org/10.4000/jtei.177>.

Carter, Kirsten Strigel, Abby Gondek, William Underwood, Teddy Randby, and Richard Marciano. 2022. "Using AI and ML to Optimize Information Discovery in Under-Utilized, Holocaust-Related Records." *AI & SOCIETY*, May. <https://doi.org/10.1007/s00146-021-01368-w>.

Chagué, Alix, and Thibault Clérice. 2022. "Sharing HTR Datasets with Standardized Metadata: The HTR-United Initiative." <https://inria.hal.science/hal-0370398>.

Chagué, Alix, Thibault Clérice, Jade Norindr, Maxime Humeau, Baudoin Davoury, Elsa Van Kote, Anaïs Mazoue, Margaux Faure, and Soline Doat. 2023. "Manu McFrench, from Zero to Hero: Impact of Using a Generic Handwriting Recognition Model for Smaller Datasets." <https://inria.hal.science/hal-04094241>.

Clérice, Thibault, Juliette Janes, Hugo Scheithauer, Sarah Bénéière, Laurent Romary, and Benoît Sagot. 2024. “Layout Analysis Dataset with SegmOnto.” <https://inria.hal.science/hal-04513725>.

Colavizza, Giovanni, Maud Ehrmann, and Fabio Bortoluzzi. 2019. “Index-Driven Digitization and Indexation of Historical Archives.” *Frontiers in Digital Humanities* 6. <https://www.frontiersin.org/articles/10.3389/fgdigh.2019.00004>.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. “Unsupervised Cross-Lingual Representation Learning at Scale.” arXiv. <https://doi.org/10.48550/arXiv.1911.02116>.

EHRI-Consortium. 2021. *Diplomatic Reports - Digital Edition*. EHRI Online Editions. European Holocaust Research Infrastructure project (EHRI). <https://diplomatic-reports.ehri-project.eu/>.

Ehrmann, Maud, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2023. “Named Entity Recognition and Classification in Historical Documents: A Survey.” *ACM Computing Surveys* 56 (2): 27:1–47. <https://doi.org/10.1145/3604931>.

Ehrmann, Maud, Damien Nouvel, and Sophie Rosset. 2016. “Named Entity Resources - Overview and Outlook.” In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, edited by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, et al., 3349–56. Portorož, Slovenia: European Language Resources Association (ELRA). <https://aclanthology.org/L16-1534>.

Ehrmann, Maud, Matteo Romanello, Antoine Doucet, and Simon Clematide. 2022. “Introducing the HIPE 2022 Shared Task: Named Entity Recognition and Linking in Multilingual Historical Documents.” In *Advances in Information Retrieval*, edited by Matthias Hagen, Suzan Verberne, Craig Macdonald, Christin Seifert, Krisztian Balog, Kjetil Nørvåg, and Vinay Setty, 347–54. Lecture Notes in Computer Science. Cham: Springer International Publishing. [https://doi.org/10.1007/978-3-030-99739-7\\_44](https://doi.org/10.1007/978-3-030-99739-7_44).

Erez, Sigal Arie, Tobias Blanke, Mike Bryant, Kepa Rodriguez, Reto Speck, and Veerle Vanden Daelen. 2020. “Record Linking in the EHRI Portal.” *Records Management Journal* 30 (3): 363–78. <https://doi.org/10.1108/RMJ-08-2019-0045>.

Ezeani, Ignatius, Paul Rayson, Ian Gregory, Erum Haris, Anthony Cohn, John Stell, Tim Cole, et al. 2023. “Towards an Extensible Framework for Understanding Spatial Narratives.” In *Proceedings of the 7th ACM SIGSPATIAL International Workshop on Geospatial Humanities*, 1–10. Hamburg Germany: ACM. <https://doi.org/10.1145/3615887.3627761>.

Frankl, Michal, Michael Bryant, Jessica Green, Wolfgang Schellenbacher, and Magdalena Sedlická. 2018. “Edition of Documents.” 654164 (H2020-INFRAIA-2014-2015). European Holocaust Research Infrastructure. <https://www.ehri-project.eu/sites/default/files/downloads/Deliverables/D12%20%20Thematic%20approach%20%20Edition%20of%20documents.pdf>.



Frankl, Michal, and Wolfgang Schellenbacher, eds. 2018. *BeGrenzte Flucht: Die Österreichischen Flüchtlinge an Der Grenze Zur Tschechoslowakei Im Krisenjahr 1938 - Digital Edition*. Translated by Annette Kraus. EHRI Online Editions. European Holocaust Research Infrastructure project (EHRI). <https://begrenzte-flucht.ehri-project.eu/>.

Michal Frankl and Wolfgang Schellenbacher, editors. 2023. *Uzavřít hranice! - Digital Edition*. EHRI Online Editions. European Holocaust Research Infrastructure project (EHRI).

Frankl, Michal, Magdalena Sedlická, Hana Dauš, and Wolfgang Schellenbacher, eds. 2023. *Documentation Campaign - Digital Edition*. Translated by Caroline Kovtun and Molly Roza. EHRI Online Editions. European Holocaust Research Infrastructure project (EHRI). <https://documentation-campaign.ehri-project.eu/>.

Frankl, Michal, Magdalena Sedlická, Wolfgang Schellenbacher, Daniela Bartáková, Michal Czajka, Jessica Green, Kat Hubschmann, et al., eds. 2020. *Early Holocaust Testimony - Digital Edition*. Translated by Yochanan Amichai, Jennifer Carvill Schellenbacher, and Caroline Kovtun. EHRI Online Editions. 2020: European Holocaust Research Infrastructure project (EHRI). <https://early-testimony.ehri-project.eu/>.

Garscha, Winfried, Claudia Kuretsidis-Haider, and Wolfgang Schellenbacher, eds. 2022. *VON WIEN INS NIRGENDWO: DIE NISKO-DEPORTATIONEN 1939*. EHRI Online Editions. <https://nisko-transport.ehri-project.eu/>.

Holmes, Martin. 2016. "Whatever happened to interchange?" *Digital Scholarship in the Humanities* 32: i63–68. <https://doi.org/10.1093/lc/fqw048>.

Honnibal, Matthew, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. "spaCy: Industrial-Strength Natural Language Processing in Python." <https://doi.org/10.5281/zenodo.1212303>.

Huang, Yupan, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. "LayoutLMv3: Pre-Training for Document AI with Unified Text and Image Masking." arXiv. <https://doi.org/10.48550/arXiv.2204.08387>.

Lauscher, Anne, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. "From Zero to Hero: On the Limitations of Zero-Shot Language Transfer with Multilingual Transformers." In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, edited by Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, 4483–99. Online: Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.emnlp-main.363>.

Leeuw, Daan de, Mike Bryant, Michal Frankl, Ivelina Nikolova, and Vladimir Alexiev. 2018. "Digital Methods in Holocaust Studies: The European Holocaust Research Infrastructure." In *2018 IEEE 14th International Conference on e-Science (e-Science)*, 58–66. <https://doi.org/10.1109/eScience.2018.00021>.

Levy, Michael. 2019. "Some Perspectives on the Practice of Sharing Collection Data." *Umanistica Digitale* 1 (4): 21–32. <https://doi.org/10.6092/issn.2532-8816/9039>.

Luthra, Mrinalini, Konstantin Todorov, Charles Jeurgens, and Giovanni Colavizza. 2023.

“Unsilencing Colonial Archives via Automated Entity Recognition.” *Journal of Documentation* ahead-of-print (ahead-of-print). <https://doi.org/10.1108/JD-02-2022-0038>.

Mattingly, W. J. B. 2021a. “Holocaust Named Entity Recognition.” <https://www.youtube.com/playlist?list=PL2VXyKi-KpYsyzKhcr2zHHP76mdZ5Dfa>.

Mattingly, W.J.B. 2021b. [wjbmattngly/holocaust\\_ner\\_lessons](https://www.youtube.com/watch?v=wjbmattngly/holocaust_ner_lessons). Original-date: 2021-01-04T18:25:11Z.

Mueller, David, Nicholas Andrews, and Mark Dredze. 2020. “Sources of Transfer in Multilingual Named Entity Recognition.” In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, edited by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, 8093–8104. Online: Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.720>.

Muñoz, Trevor, and Raffaele Vigiante. 2015. “Texts and Documents: New Challenges for TEI Interchange and Lessons from the Shelley-Godwin Archive.” *Journal of the Text Encoding Initiative* 1 (8). <https://doi.org/10.4000/jtei.1270>.

Nanomi Arachchige, Isuri Anuradha, Le Ha, Ruslan Mitkov, and Johannes-Dieter Steinert. 2023. “Enhancing Named Entity Recognition for Holocaust Testimonies Through Pseudo Labelling and Transformer-Based Models.” In *Proceedings of the 7th International Workshop on Historical Document Imaging and Processing*, 85–90. San Jose CA USA: ACM. <https://doi.org/10.1145/3604951.3605514>.

Nikolova, Ivelina, and Michael Levy. 2018. “Using Named Entity Recognition to Enhance Access to a Museum Catalog – Document Blog.” <https://blog.ehri-project.eu/2018/08/27/named-entity-recognition/>.

Pinche, Ariane, and Jean-Baptiste Camps. 2022. “CremmaLab Project: Transcription Guidelines and HTR Models for French Medieval Manuscripts.” In *Colloque “Documents Anciens et Reconnaissance Automatique Des Écritures Manuscrites”*. Paris, France. <https://hal.science/hal-03716526>.

Rodriguez, Kepa J, Vladimir Alexiev, Laura Brazzo, Charles Riondet, Yael Gherman, and Reto Speck. 2016. “EHRI-2 - D.11.2 Road Map Domain Vocabularies.” Deliverable GA no. 654164. [https://www.ehri-project.eu/sites/default/files/downloads/ehri\\_downloads/D%2011.2%20Road%20map%20domain%20vocabularies.pdf](https://www.ehri-project.eu/sites/default/files/downloads/ehri_downloads/D%2011.2%20Road%20map%20domain%20vocabularies.pdf).

Rodriguez, Kepa Joseba, Mike Bryant, Tobias Blanke, and Magdalena Luszczynska. 2012. “Comparison of Named Entity Recognition Tools for Raw OCR Text.” In, 410–14. [https://www.oegai.at/konvens2012/proceedings/60\\_rodriguez12w/](https://www.oegai.at/konvens2012/proceedings/60_rodriguez12w/).

Romary, Laurent, and Charles Riondet. 2019. “Towards Multiscale Archival Digital Data.” *Umanistica Digitale* 1 (4): 89–99. <https://doi.org/10.6092/issn.2532-8816/9046>.

Sang, Erik F. Tjong Kim, and Fien De Meulder. 2003. “Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition.” arXiv.



<https://doi.org/10.48550/arXiv.cs/0306050>.

Schmidt, Desmond. 2014. "Towards an Interoperable Digital Scholarly Edition." *Journal of the Text Encoding Initiative* 1 (7). <https://doi.org/10.4000/jtei.979>.

Schweter, Stefan, Luisa März, Katharina Schmid, and Erion Çano. 2022. "hmBERT: Historical Multilingual Language Models for Named Entity Recognition." In *CEUR Workshop Proceedings*. Bologna, Italy: arXiv. <https://doi.org/10.48550/arXiv.2205.15575>.

Stokes, Peter A., Benjamin Kiessling, Daniel Stökl Ben Ezra, Robin Tissot, and El Hassane Gargem. 2021. "The eScriptorium VRE for Manuscript Cultures, in Ancient Manuscripts and Virtual Research Environments." *Classic @ Journal*, no. 18. <https://classics-at.chs.harvard.edu/classics18-stokes-kiessling-stokl-ben-ezra-tissot-gargem/>.

Stutzmann, Dominique. 2011. "Paléographie statistique pour décrire, identifier, dater... Normaliser pour coopérer et aller plus loin?" In , 247. BoD. <https://shs.hal.science/halshs-00596970>

Sven, Najem-Meyer, and Romanello Matteo. 2022. "Page Layout Analysis of Text-Heavy Historical Documents: A Comparison of Textual and Visual Approaches." arXiv. <https://doi.org/10.48550/arXiv.2212.13924>.

TEI Consortium, eds. TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 4.7.0. Last modified November 16, 2023. TEI Consortium. <http://www.tei-c.org/Guidelines/P5/> (Accessed July 5, 2024).

Unsworth, John. 2011. "Computational Work with Very Large Text Collections." *Journal of the Text Encoding Initiative* 1 (1). <https://doi.org/10.4000/jtei.215>.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. "Attention Is All You Need." In *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc. [https://proceedings.neurips.cc/paper\\_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html](https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html).

Wu, Qianhui, Zijia Lin, Guoxin Wang, Hui Chen, Börje F. Karlsson, Biqing Huang, and Chin-Yew Lin. 2020. "Enhanced Meta-Learning for Cross-Lingual Named Entity Recognition with Minimal Resources." *Proceedings of the AAAI Conference on Artificial Intelligence* 34 (05): 9274–81. <https://doi.org/10.1609/aaai.v34i05.646>.

## Appendix

### **BeGrenzte Flucht Edition**

The BeGrenzte Flucht (BF) edition (Frankl and Schellenbacher 2018) gathers documents kept in various Czech and Austrian archives relating to Austrian refugees on the border to Czechoslovakia in the crisis year 1938, including official reports, correspondence, diplomatic notes, newspaper reports, and documents from Jewish aid organizations. The BF edition is in German and the vast majority of documents, if not originally in German, have been translated into German. Transcripts in the original languages of the documents, including Czech, Slovak, and English are also included.

### **Early Holocaust Testimonies Edition**

The Early Holocaust Testimony (ET) edition (Frankl et al. 2020) contains selected and edited testimonies and reports on the persecution of the Jews in Nazi Germany written by foreign diplomats stationed in Nazi Germany to their respective Ministry of Foreign Affairs, which are kept in five different archives: the Wiener Holocaust Library in London, Yad Vashem in Jerusalem, the Jewish Historical Institute in Warsaw, the Hungarian Jewish Archives in Budapest, and the Jewish Museum in Prague. The testimonies span from 1933, when Adolf Hitler was appointed Chancellor, to the trial of Adolf Eichmann in 1961. All of the documents have an English translation but transcripts of the original documents in Czech, German, Hungarian, Polish, Dutch, and Yiddish are provided.

### **Diplomatic Reports Edition**

The Diplomatic Reports (DR) edition (EHRI-Consortium 2021) gathers documents created by the diplomatic staff of allied countries, opponents, and neutral countries. They all report on the German occupation. They include reports from the diplomatic staff of Denmark, Italy, Japan, Hungary, Slovakia, and the US. All of the documents have been translated into English, regardless of their original language.

### **Von Wien ins Nirgendwo: Die Nisko-Deportationen 1939 Edition**

The Von Wien ins Nirgendwo: Die Nisko-Deportationen 1939 (ND) edition (Garscha, Kuretsidis-Haider, and Schellenbacher 2022) was created in cooperation with the Documentation Archive of the Austrian Resistance. It is a collection of testimonies and letters documenting the Nisko Plan, which aimed at creating a Jewish reservation, built by the Jews themselves, in Nisko and Lublin (Poland). The edition focuses on the deportation of approximately 1,600 Jewish men from Vienna to Nisko on the 20<sup>th</sup> and 26<sup>th</sup> October 1939 and what became of them. The source documents are from various archival institutions in different countries and are provided in German.

### **Documentation Campaign Edition**

The Documentation Campaign (DC) edition (Frankl et al. 2023) gathers documents held by the Jewish Museum in Prague and by Yad Vashem consisting of Holocaust survivor testimonies and photographs collected within the framework of the so-called “Documentation Campaign” in Prague, one of the earliest postwar projects to document the events of the

Shoah, collecting evidence, documents, and witness testimonies. All of the documents have been translated into English but transcripts of the original documents in Czech and German are provided.

### **Uzavřít Hranice Edition**

Similar to the BF edition, the Uzavřít Hranice (UH) edition (Frankl and Schellenbacher 2023) gathers documents kept in various Czech, Austrian, and other archives relating to Austrian refugees on the border to Czechoslovakia in the crisis year 1938. The UH edition is in Czech and the vast majority of documents, if not originally in Czech, have been translated into Czech. Transcripts in the original languages of the documents, including Czech, Slovak, and English are also included. Since the EHRI Online Editions cover a variety of languages, document types, periods, and thematic and spatial areas of focus, training NER models on this dataset may lead to tools that can generalize better on different types of Holocaust-related documents, compared to training them only on testimony-based corpora like in previous work. This will hopefully make our models more robust and interoperable across different EHRI services.