# TEI Specifications for a Sustainable Management of Digitized Holocaust Testimonies

**Sarah Bénière[1], Floriane Chiffoleau[1,2], Laurent Romary[1,3]**
[1]ALMAnaCH, Inria, Paris
[2]Le Mans Université, Le Mans
[3]Directorate for Scientific Information and Culture, Inria
{firstname.surname}@inria.fr

**Abstract**

Data modeling and standardization are central issues in the field of Digital Humanities, and all the more so when dealing with Holocaust testimonies, where stable preservation and long-term accessibility are key. The EHRI Online Editions are composed of documents of diverse nature (testimonies, letters, diplomatic reports, etc.), held by EHRI's partnering institutions, and selected, gathered thematically and encoded according to the TEI Guidelines by the editors within the EHRI Consortium. Standardization is essential in order to make sure that the editions are consistent with one another. The issue of consistency also encourages a broader reflection on the usage of standards when processing data, and on the standardization of digital scholarly editions of textual documents in general. In this paper, we present the normalization work we carried out on the EHRI Online Editions. It includes a customization of the TEI adapted to Holocaust-related documents, and a focus on the implementation of controlled vocabulary. We recommend the use of these encoding specifications as a tool for researchers and/or non-TEI experts to ensure their encoding is valid and consistent across editions, but also as a mechanism for integrating the edition work smoothly within a wider workflow leading from image digitization to publication.

**Keywords:** Standardization, TEI-XML, Digital Humanities, European Holocaust Research Infrastructure

## 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[1] (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 in Section 3.

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 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 Viglianti, 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.

This paper presents 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. Section 2 presents the EHRI Online Editions, Section 3 deals with data structuration in TEI, and Section 4 focuses on the EHRI TEI customization[2]. Finally, Section 5 discusses the extension of the EHRI specifications to all encoding projects dealing with Holocaust-related documents.

[1]https://www.ehri-project.eu/

[2]https://github.com/SarahBeniere/EHRI-Workflow/blob/main/ENCODING/Guidelines/ODD_EHRI.xml

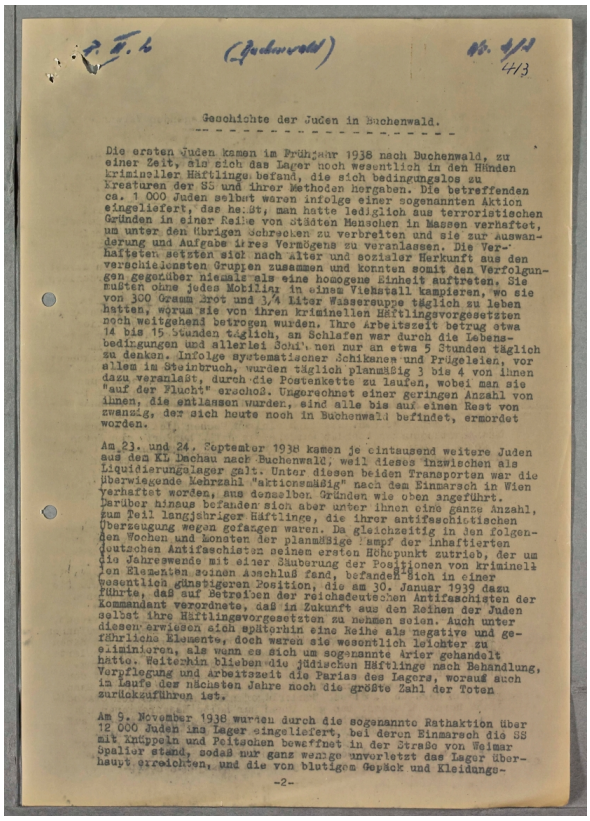


*The First Workshop on Holocaust Testimonies as Language Resources @LREC-COLING-2024*, pages 10–17
20-25 May, 2024. 

## 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 WP12 "New Approaches to Holocaust Research and Archiving".

Within the framework of WP12, EHRI has already published five online editions[3]. 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[4].



Geschichte der Juden in Buchenwald.

Die ersten Juden kamen im Frühjahr 1938 nach Buchenwald, zu einer Zeit, als sich das Lager noch wesentlich in den Händen krimineller Häftlinge befand, die sich bedingungslos zu Kreaturen der SS und ihrer Methoden hergaben. Die betreffenden ca. 1 000 Juden selbst waren infolge einer sogenannten Aktion eingeliefert, das heißt, man hatte lediglich aus terroristischen Gründen in einer Reihe von Städten Menschen in Massen verhaftet, um unter den übrigen Schrecken zu verbreiten und sie zur Auswanderung und Aufgabe ihres Vermögens zu veranlassen. Die Verhafteten setzten sich nach alter und sozialer Herkunft aus den verschiedensten Gruppen zusammen und konnten somit den Verfolgungen gegenüber niemals als eine homogene Einheit auftreten. Sie mußten ohne jedes Mobiliar in einem Viehstall kampieren, wo sie von 300 Gramm Brot und 3/4 Liter Wassersuppe täglich zu leben hatten, wozu sie von ihren kriminellen Häftlingsvorgesetzten noch weitgehend betrogen wurden. Ihre Arbeitszeit betrug etwa 14 bis 15 Stunden täglich, an Schlafen war durch die Lebensbedingungen und allerlei Schikanen nur an etwa 5 Stunden täglich zu denken. Infolge systematischer Schikanen und Prügeleien, vor allem im Steinbruch, wurden täglich planmäßig 3 bis 4 von ihnen dazu veranlaßt, durch die Postenkette zu laufen, wobei man sie "auf der Flucht" erschoß. Ungerechnet einer geringen Anzahl von ihnen, die entlassen wurden, sind alle bis auf einen Rest von zwanzig, der sich heute noch in Buchenwald befindet, ermordet worden.

Am 23. und 24. September 1938 kamen je eintausend weitere Juden aus den KL Dachau nach Buchenwald, weil dieses inzwischen als Liquidierungslager galt. Unter diesen beiden Transporten war die überwiegende Mehrzahl "aktionsmäßig" nach dem Einmarsch in Wien verhaftet worden, aus denselben Gründen wie oben angeführt. Darüber hinaus befanden sich aber unter ihnen eine ganze Anzahl, zum Teil langjähriger Häftlinge, die ihrer antifaschistischen Überzeugung wegen gefangen waren. Da gleichzeitig in den folgenden Wochen und Monaten der planmäßige Kampf der inhaftierten deutschen Antifaschisten seinen ersten Höhepunkt zutrieb, der um die Jahreswende mit einer Säuberung der Positionen von kriminellen Elementen seinen Abschluß fand, befanden sich in einer wesentlich günstigeren Position, die am 30. Januar 1939 dazu führte, daß auf Betreiben der reichsdeutschen Antifaschisten der Kommandant verordnete, daß in Zukunft aus den Reihen der Juden selbst ihre Häftlingsvorgesetzten zu nehmen seien. Auch unter diesen erwiesen sich späterhin eine Reihe als negative und gefährliche Elemente, doch waren sie wesentlich leichter zu eliminieren, als wenn es sich um sogenannte Arier gehandelt hätte. Weiterhin blieben die jüdischen Häftlinge nach Behandlung, Verpflegung und Arbeitszeit die Parias des Lagers, woraus auch im Laufe des nächsten Jahre noch die größte Zahl der Toten zurückzuführen ist.

Am 9. November 1938 wurden durch die sogenannte Rathaktion über 12 000 Juden ins Lager eingeliefert, bei deren Einmarsch die SS mit Knüppeln und Peitschen bewaffnet in der Straße von Weimar Spalier stand, sodaß nur ganz wenige unverletzt das Lager überhaupt erreichten, und die von blutigem Gepäck und Kleidungs-

-2-

Figure 1: Example of testimony

**Bordered Escape** contains testimonies on the forced emigration of the Jewish population of Austria after its annexation in March 1938. It focuses on the situation at the Czechoslovakian border, especially as Czechoslovakia's immigration policy became more and more restricted.

**Early Holocaust Testimony** is composed of written or transcribed oral testimonies on the persecution of the Jews in Nazi Germany. The testimonies span from 1933, when Adolf Hitler was appointed Chancellor, to the trial of Adolf Eichmann in 1961.

**Diplomatic Reports** gathers reports written by foreign diplomats stationed in Nazi Germany to their respective Ministry of Foreign Affairs.

**From Vienna to Nowhere: The 1939 Nisko Deportations** 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 20th and 26th October 1939 and what became of them.

**Documentation Campaign** is composed of testimonies of survivors collected in 1945 and 1946 during the "Documentation Campaign" in Prague (Czechoslovakia), which is one of the first postwar initiatives to document the events of the Holocaust.

## 3. Structuring Data in TEI

### 3.1. 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

[3] https://www.ehri-project.eu/ehri-online-editions

[4] When unavailable on EHRI's website, the translations of the titles of the editions in English are our own.

particular on EAD-XML (Alexiev et al., 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.

### 3.2. 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[5]. 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[6] codes for the representation of languages;
2. ISO 3166[7] codes for the representation of names of countries;
3. ISO 8601[8] standard for dates (`YYYY-MM-DD`).

## 4. TEI Customization for Holocaust Testimonies

### 4.1. 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="subeject"` 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 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.

### 4.2. 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 in Section 4.1. 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 the ISO 639 norm (Figure 6), available

[5] https://github.com/EHRI/ehri-online-editions
[6] https://www.iso.org/iso-639-language-code
[7] https://www.iso.org/iso-3166-country-codes.html
[8] https://www.iso.org/iso-8601-date-and-time-format.html

through the Iana Language Subtag Registry[9].

```
<valList mode="add" type="semi">
    <valItem ident="cs">
        <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[10] presents itself as one of the main resources about the Holocaust for it gathers information on archival sources from across the world. One of their primary achievements 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[11]. The last step of this workflow is the publication of the editions on a TEI Publisher[12] 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).

**Encoding Template for the `<teiHeader>`** As we mentioned previously, particular attention must be given when encoding the documents' metadata. We created a template (Appendix A) 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)[13].

[9] https://www.iana.org/assignments/language-subtag-registry/language-subtag-registry

[10] https://portal.ehri-project.eu/

[11] https://github.com/SarahBeniere/EHRI-Workflow/tree/main

[12] https://teipublisher.com/exist/apps/tei-publisher-home/index.html

[13] https://creativecommons.org/licenses/by/4.0/

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

Figure 8: Specifications for first-level `<div>`

## 5. Discussion and Conclusion

This paper presents 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[14]. 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.

## 6. Acknowledgements

This work has been carried out in the context of the EHRI-3 project funded by the European Commission under the call H2020-INFRAIA-2018-2020, with grant agreement ID 871111 and DOI 10.3030/871111. We would like to warmly thank our colleagues at EHRI for their precious work on the digital editions, mainly Michael Bryant, Maria Dermentzi, Michal Frankl, Aneta Plzáková, Wolfgang Schellenbacher, and Magdalena Sedlická.

We would also like to thank our colleagues Lydia Nishimwe and Hugo Scheithauer for their thorough proofreading of the paper.

[14] As of now, new EHRI editions have not been prepared yet.

## 7. Bibliographical References

Vladimir Alexiev, Ivelina Nikolova, and Neli Hateva. 2019. Semantic Archive Integration for Holocaust Research. *Umanistica Digitale*, 1(4):131–175.

Syd Bauman. 2011. Interchange vs. Interoperability. In *Proceedings of Balisage: The Markup Conference 2011*.

Lou Burnard. 2014. *What is the Text Encoding Initiative?* OpenEdition Press.

Lou Burnard. 2016. *ODD chaining for Beginners*. GitHub.

Lou Burnard. 2018. What is TEI Conformance, and Why Should You Care? *Journal of the Text Encoding Initiative*, 1(12).

Michal Frankl, Michael Bryant, Jessica Green, Wolfgang Schellenbacher, and Magdalena Sedlická. 2018. Edition of Documents. Technical Report 654164 (H2020-INFRAIA-2014-2015), European Holocaust Research Infrastructure.

Martin Holmes. 2016. Whatever happened to interchange? *Digital Scholarship in the Humanities*, 32:i63–i68.

Michael Levy. 2019. Some Perspectives on the Practice of Sharing Collection Data. *Umanistica Digitale*, 1(4):21–32.

Trevor Muñoz and Raffaele Viglianti. 2015. Texts and Documents: New Challenges for TEI Interchange and Lessons from the Shelley-Godwin Archive. *Journal of the Text Encoding Initiative*, 1(8).

Laurent Romary and Charles Riondet. 2019. Towards Multiscale Archival Digital Data. *Umanistica Digitale*, 1(4):89–99.

Desmond Schmidt. 2014. Towards an Interoperable Digital Scholarly Edition. *Journal of the Text Encoding Initiative*, 1(7).

TEI Consortium, editor. 2023. *TEI P5: Guidelines for Electronic Text Encoding and Interchange (ver. 4.7.0)*. TEI Consortium.

John Unsworth. 2011. Computational Work with Very Large Text Collections. *Journal of the Text Encoding Initiative*, 1(1).

# A. Template for the <teiHeader>

```
<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>
```