



**HAL**  
open science

# Streamlining the Creation of Holocaust-related Digital Editions with Automatic Tools

Sarah Bénéière, Floriane Chiffolleau, Hugo Scheithauer

► **To cite this version:**

Sarah Bénéière, Floriane Chiffolleau, Hugo Scheithauer. Streamlining the Creation of Holocaust-related Digital Editions with Automatic Tools. EHRI Academic Conference - Researching the Holocaust in the Digital Age, EHRI-3, Jun 2024, Varsovie, Poland. hal-04594190v2

**HAL Id: hal-04594190**

**<https://inria.hal.science/hal-04594190v2>**

Submitted on 24 Jun 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

# Streamlining the Creation of Holocaust-related Digital Editions with Automatic Tools



*ALMAnaCH project-team*

*Inria*

**Sarah Bénéière, Floriane Chiffolleau, Hugo Scheithauer**  
EHRI Academic Conference, Warsaw, Poland  
June 18th, 2024

# Outline of the presentation

1. The **project** and us
2. The **sources**
3. **Outline** of the semi-automated workflow
4. **Acquiring the content** of the source faster
5. **Obtaining the structure** of the document
6. **Enriching the transcriptions** semi-automatically
7. **Proposing a centralized interface** for future holocaust-related editions
8. **Conclusion**
9. **Resources**

# THE PROJECT AND US



# A bit of background

## Who are we?

### Floriane Chiffoleau

PhD Candidate in Digital Humanities (Inria, Le Mans Université).

Thesis topic: Automatic Text Recognition and Ground Truth

### Hugo Scheithauer

PhD Candidate in Digital Humanities (Inria, Ecole Pratique des Hautes Etudes, Paris).

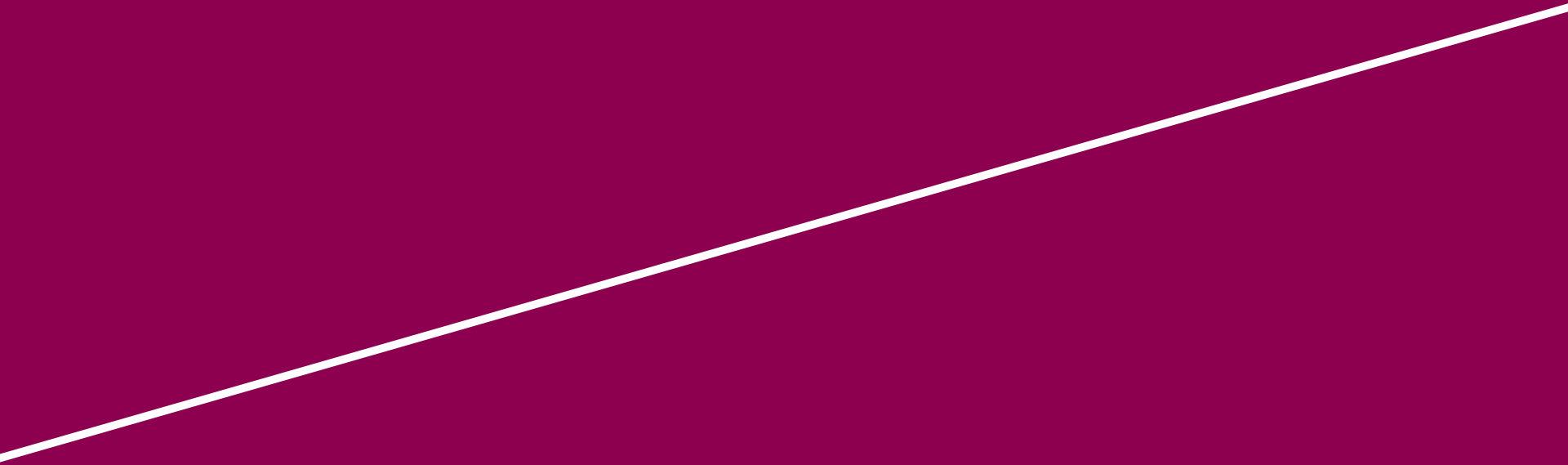
Thesis topic: Document Layout Analysis and Document Understanding

### Sarah Benière

R&D Engineer in Digital Humanities (Inria).

Master's Internship on the project

# THE SOURCES



# What are we working with?

Six Holocaust-related online editions developed and published by the EHRI consortium:

- ❑ **Begrenzte Flucht Edition** (2018) & **Uzavřít Hranice Edition** (2023): Documents related to the crisis year 1938 at the Czechoslovakia border and Austrian refugees.
- ❑ **Early Holocaust Testimonies Edition** (2020): Testimonies in Czech, German, Hungarian, Polish, Dutch, and Yiddish.
- ❑ **Diplomatic Reports Edition** (2021): Reports from the diplomatic staff of Denmark, Italy, Japan, Hungary, Slovakia, and the United States.
- ❑ **Von Wien ins Nirgendwo - Die Nisko-Deportationen 1939 Edition** (2023): Documents on the history of the Viennese Jewish deportees to Nisko
- ❑ **Documentation Campaign Edition** (2023): Holocaust survivor testimonies

# OUTLINE OF THE SEMI-AUTOMATED WORKFLOW





# Semi-automated workflow

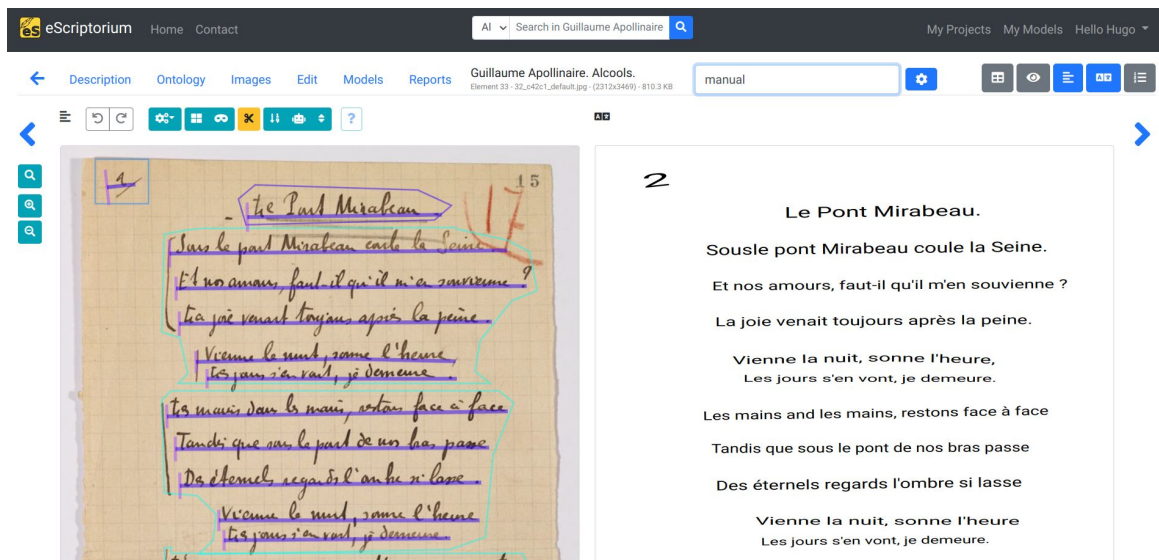


**ACQUIRING THE CONTENT OF THE  
SOURCE FASTER**



# A tool for ATR: eScriptorium

- ❑ GUI for transcribing textual documents, creating training sets, training segmentation and transcription models, etc.
- ❑ Project management tool
- ❑ eScriptorium uses the language-agnostic transcription engine Kraken (Benjamin Kiessling, PSL)
- ❑ Inria hosts an [eScriptorium instance](#). You can ask for an account with Cremma Call form (in French for now, sorry!).



# Creation of a multilingual recognition model

- ❑ Two **objectives**:
  - ❑ Producing an **efficient multilingual recognition model** for typescript documents
  - ❑ Contributing to a **PhD experiment**
- ❑ Creation of **training data** for the model
  - ❑ **252 typescript documents, 7 languages**
  - ❑ Documents **segmented and transcribed** (copy/paste or manual)
  - ❑ Images/Texts/XML: <https://github.com/FloChiff/ehri-dataset>
  - ❑ Presentation of the dataset: <https://flochiff.github.io/phd/dataset/ehri/dataset.html>

# Creation of a multilingual recognition model

- ❑ Model with **97.20% of accuracy**
- ❑ **Efficiency** of the model:
  - ❑ Working pretty well with the **languages it was trained on**
  - ❑ Recognition abilities for the **languages it was not trained on** (if same Latin script)
  - ❑ Trouble with the **diacritics and uppercase**
- ❑ **Other models available**: production of single-language models for each language of the training data

# EHRI automatic text recognition dataset and training results

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%

# OBTAINING THE STRUCTURE OF THE DOCUMENT



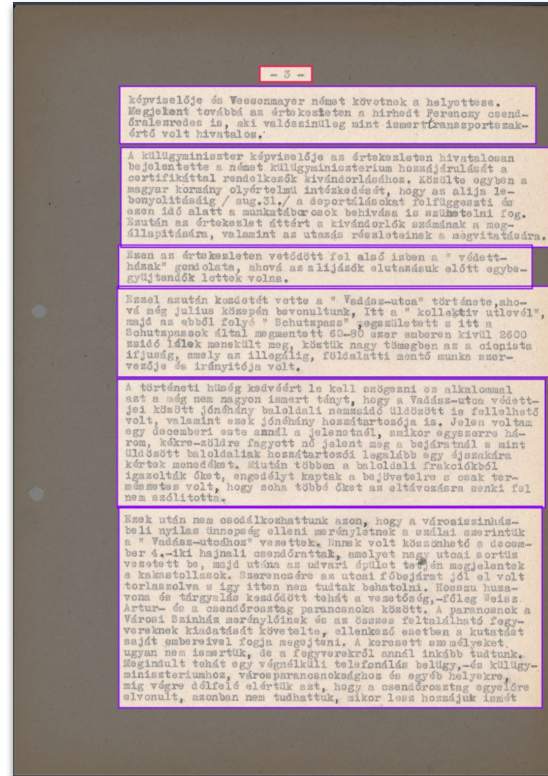
# Semi-automatic encoding with layout analysis

## Document Layout Analysis (DLA)

### Detection of the **layout components** and their **hierarchy**

Two main approaches:

- **Visual features** (YOLOv8)
- Combination of textual and visual features (LayoutLM)



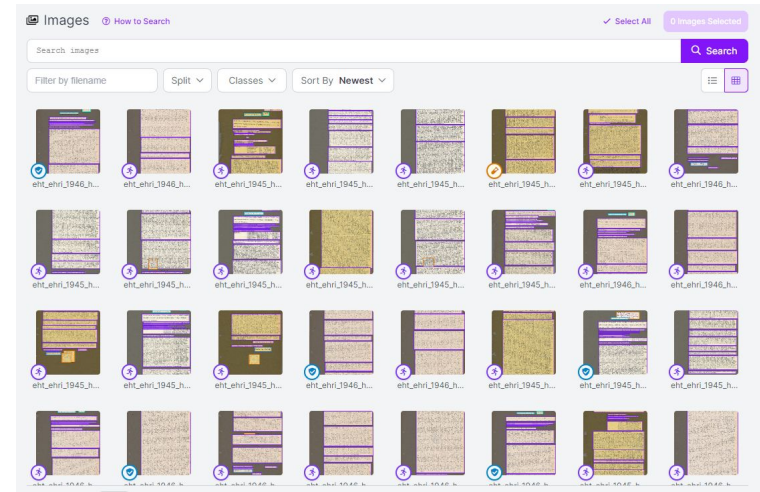
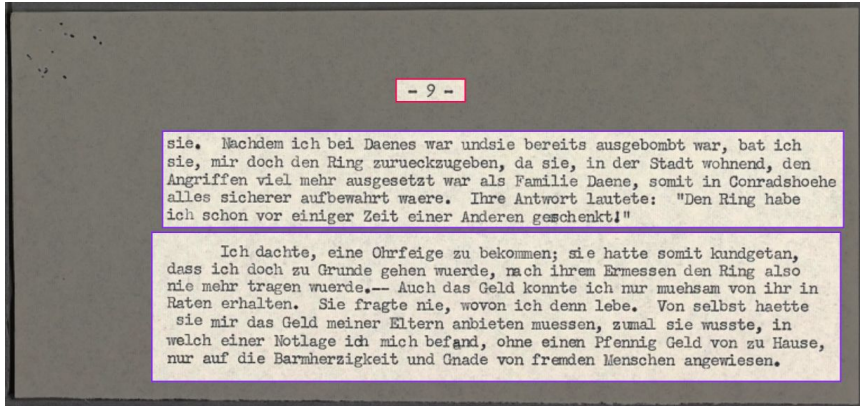


# Semi-automatic encoding with layout analysis

200 images from the Early Testimonies edition

Annotated with **Roboflow** → dataset

**YOLOv8 model** (object detection, pixel based)



Clérice, T. (2023). You Actually Look Twice At it (YALTAi): using an object detection approach instead of region segmentation within the Kraken engine. *Journal of Data Mining & Digital Humanities*, 2023. <https://doi.org/10.46298/jdmhdh.9806>.

Clérice, T., Janès, J., Scheithauer, H., Bénérie, S., Romary, L., & Sagot, B. (2024). *Layout Analysis Dataset with SegmOnto*. DH2024 - Annual conference of the Alliance of Digital Humanities Organizations, Washington, D.C., United States. ([hal-04513725](https://hal-04513725))

# Semi-automatic encoding with layout analysis

Page number

- 3 -

Paragraphs

Képviselője és Wesemannyer német követnek a helyettese.  
Mégjelent továbbá az értekezleten a hírhedt Parancsnok csend-  
őrszolgálatos is, aki valószínűleg mint amerikai sportszak-  
értő volt hivatalos.

A külügyminiszter képviselője az értekezleten hivatalosan  
bejelentette a német külügyminisztérium hozzájárulását a  
certifikáttal rendelkezők kivándorlásához. Később egyben a  
magyar kormány eljuttatási iránti kérelmét, hogy az alább le-  
bonyolításig / aug.31./ a deportálódókat felügyeleti és  
ezen idő alatt a munkatáborok behívása is szabotálni fog.  
Számtalan értekezlet történt a kivándorlók számára a meg-  
állapítására, valamint az utazás visszatérési a megvitására.

Éppen az értekezleten vetődött fel első ízben a "vándor-  
hírek" gondolatja, ahová az eljártak elutazásuk előtt egybe-  
gyűjtendőek lettek volna.

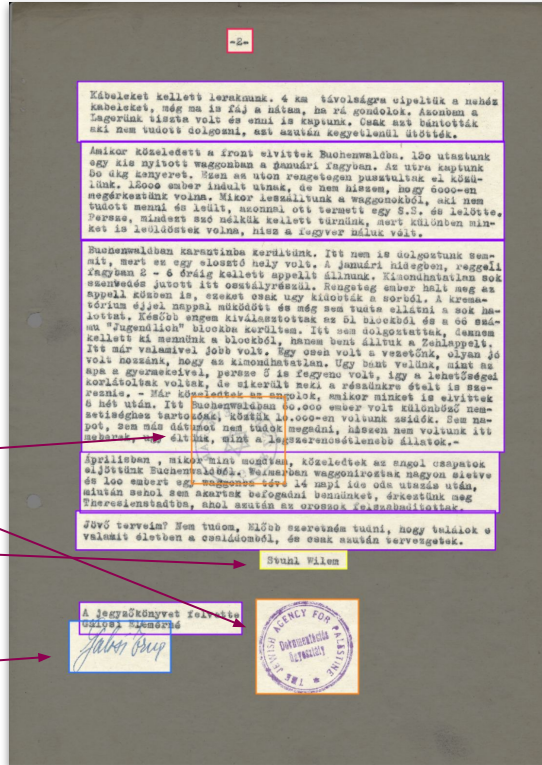
Ezzel együtt kiderült, hogy a "Vadász-utazás" története, ah-  
ová még július közepén bevonultunk, itt a "kollektív utazás",  
majd az ebből folyó "Schutzpass" megszületett a itt a  
Schutzpassok által megnevezett 60-80 ember számára kívül 2500  
szűk lélek menekült meg, köztük nagy tömegben az a csoport  
ifjúság, amely az illegálisan, földalatti mentés munka esze-  
resztése és irányítása volt.

A történeti háttér kedvéért le kell szögezni az alkalommal  
arról a még nem nagyon ismert tényről, hogy a Vadász-utazás vándor-  
hírek kiderült jogszabályi belvárosi nemcsak utazásról is fellelhető  
volt, valamint ezek jogszabályi háttérrel is. Jelen voltam  
egy decemberi este emeli a jelentést, amikor egyszerre há-  
rom, három-négyre fejtett német jelentést meg a bejárattal a mint  
először beloldaliak hozzátartozói legalább egy éjszakára  
kértek menedéket. Mivel többet a beloldali frakciókbeli  
igazságtörvények, engedélyt kaptak a bejövatalra a csak ter-  
vezéses volt, hogy csak többé őket az utazásra senki fel  
nem engedte.

Ezek után nem eszélkőzhettünk azon, hogy a városi színház-  
beli nyilas ünnepség elleni merényletnek a színi színház  
a "Vadász-utazás" vezetők, ennek volt köszönhető a decem-  
ber 4-iki hajnali események, amelyet nagy utcai soros  
vezetett be, majd utána az utcai épület területén megjelentek  
a katonaságok. Ezerenként az utcai főbejárat jól al volt  
szóval az a egy itten nem voltak bejutni. Hosszú husa-  
vona és tárgyalás kezdődött tehát a vezetőség, főleg Weisz  
Artúr és a csendőrszolgálat parancsnoka között. A parancsnok a  
Városi színház színházának és az összes feltalálták fagy-  
vezetők kiderült követelt, ellenkező esetben a kiderült  
majd sereggel fogja segíteni. A kérését azonnal  
ugyan nem ismertük, de a fegyverekről annál inkább tudunk.  
Segítségül tehát az a végül is telefonos beszélgetés, de külgy-  
minisztériumhoz, városparancsnokhoz és egyéb helyekre,  
sőt végül a főfelől elvették azt, hogy a csendőrszolgálat egyelőre  
elvonult, azonban nem tudhattuk, siker lesz hozzájuk innét

# Semi-automatic encoding with layout analysis

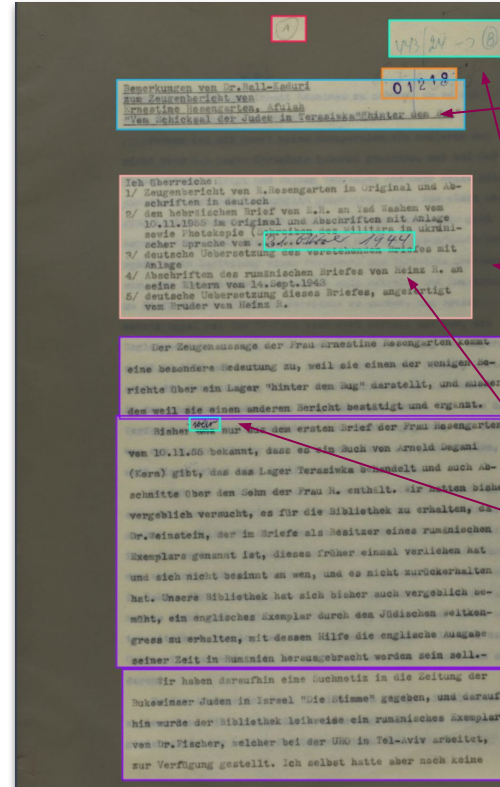
The range varies from very simple layout to more intricate layouts.



Stamp

Author

Handwritten signature



Heading

List

Handwritten annotation

# Semi-automatic encoding with layout analysis



Gabay, S., Pinche, A., Christensen, K., & Camps, J.-B. (2023). SegmOnto: A Controlled Vocabulary to Describe and Process Digital Facsimiles. ([hal-04343404](https://hal.archives-ouvertes.fr/hal-04343404)).

**13 classes** in the EHRI dataset  
(for now)

Annotations derived from the  
SegmOnto **controlled vocabulary**

- ↳ *Interoperable*
- ↳ *Reusable*

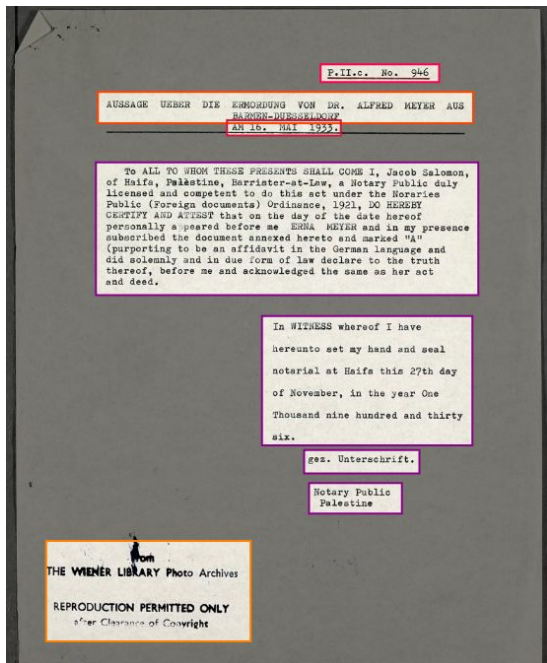
SegmOnto is **highly compatible with**  
**TEI-XML** encoding

↘

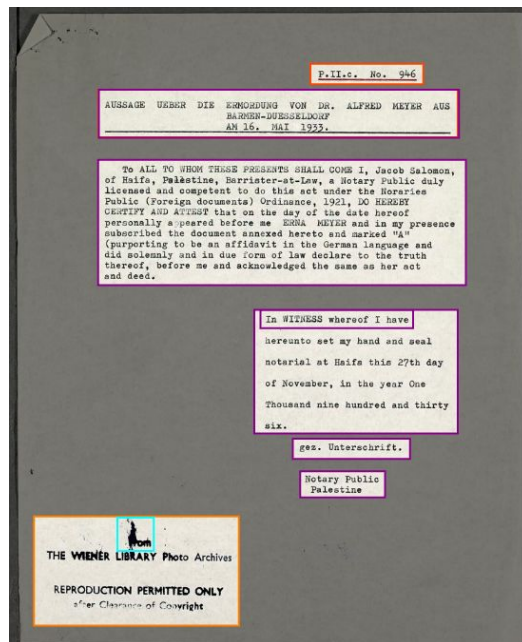
- MainZone:P
- NumberingZone
- MarginTextZone:ManuscriptAddendum
- MainZone:Head
- ManuscriptAddendum:Signature
- StampZone
- MainZone:Date
- MainZone:Signature
- StampZone:Sticker
- RunningTitleZone
- MainZone:List
- MainZone:Form
- MarginTextZone:Notes



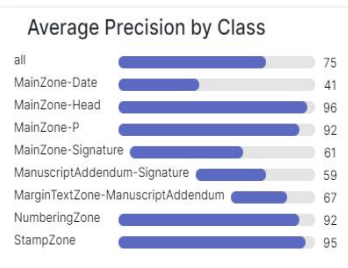
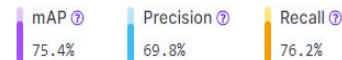
# Semi-automatic encoding with layout analysis



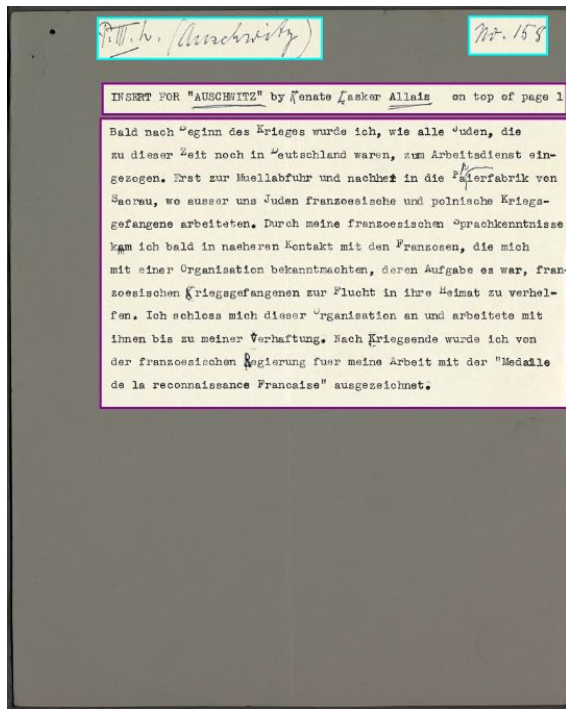
Ground truth



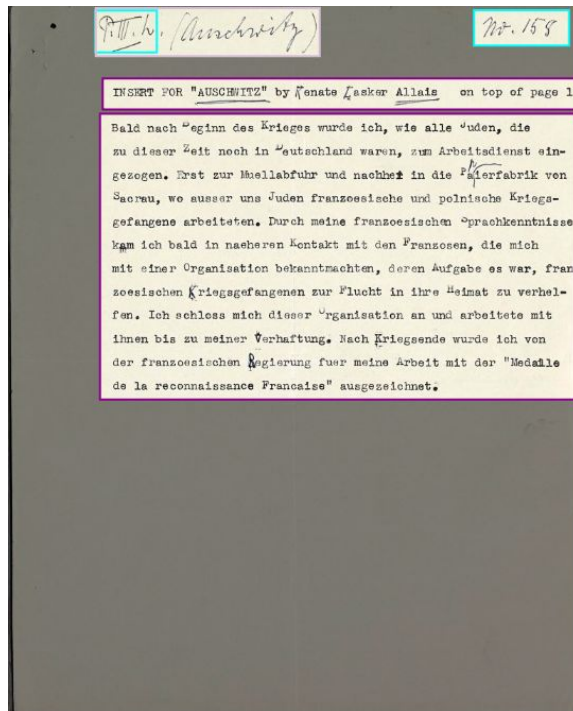
Prediction



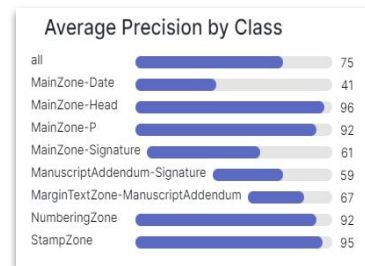
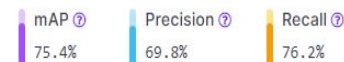
# Semi-automatic encoding with layout analysis



Ground truth



Prediction



# Semi-automatic encoding with layout analysis

- 3 -

Wien, März 1939

Liebe Tante,

Ich sende Dir einen Brief von meiner Tante. Hansi ist jetzt immer bei Frau F., bis die Tante nach Hause kommt.

Viele Russert von Leni.

Aus den Gefängnis, März 1939

Mein liebes gutes Tootchen,

Gestern war mein Leuchen bei mir und hat mir eingestanden, dass sie Dir die ganze Wahrheit geschrieben hat. Liebe Tante, wirst Du jetzt schlecht von mir denken? Ich weiß nicht, ob Leni Dir alles genau gesagt haben hat, ich kann hier nicht soviel mit ihr sprechen. Man hat uns alles weggenommen, wir sind nur mit dem was wir an Liebe gehabt haben, da gestanden, ich habe in der Aufregung einige Worte gesagt, die ich nicht hätte sagen sollen, und deshalb bin ich schon 3 Monate hier. Demke deswegen nicht schlecht von mir. Schon Liebes Tantele, Dein Bruder war doch auch voriges Jahr 13 Wochen hier, wo ich bin, das weist Du doch und deswegen war er auch nicht schlecht. Ich danke Dir viel tausendmal für die schönen Sachen, wenn ich sie nicht hätte, so müsste ich immer in einem Kleid herumgehen. Sie passen mir sehr gut, nur die Schuhe sind mir zu gross aber ich kann sie auch tragen. Du kannst Dir nicht vorstellen, wie ich mich gefreut habe mit den Sachen und hauptsächlich, weil sie von Dir sind.

Ich habe immer an Dich gedacht. Sei mir nicht böse, dass Leni Dir geschrieben hat, dass ich im Spital bin, ich habe mich geschämt, Dir die Wahrheit zu schreiben. Bitte schreibe der Tante Berta nichts davon. Wenn ich auch hier bin, so versage ich trotzdem nicht und ich hoffe, dass ich in kurzer Zeit bei meinem geliebten Baberl sein werde. Leni ist ein gutes Kind, wenn ich sie nicht hätte, so wäre ich ganz verlassen. Ich werde sie deswegen immer bei mir halten. Ich würd'lich freuen, wenn Leni mir ein Mitwoch von Dir einen Brief bringen würde. Sie kommt jeden Mitwoch zu mir. Tausend Grüsse und Kusse

Deine Marthe.

Grüsse Deinen Mann und Tochter und Tante Anna herzlich.

Ich sende Dir eine Aufnahme von mir, sie ist von hier und ist nicht gut. Den Brief gebe ich Leni mit. Hoffentlich kommt er gut an. Bitte nochmals, sage der Tante Berta nichts davon.

Wien, April 1939

Liebes gutes Tantele,

Habe Deinen Brief erhalten. Das Packet habe ich noch nicht, aber das dauert immer länger. Ich werde Dir gleich schreiben, wenn es kommt. Dieses Mal war es nicht möglich, einen Brief von Tante Martha bei zu lassen. Der "Onkel" hat zu gut aufgepasst. Sie darf weder schreiben noch Briefe empfangen. Liebes Tantele, Du schreibst, wir sollen Gottvertrauen haben. Ich glaube nicht an Gott. Warum hat er uns so gestrafft? Ich kann Dir nicht schildern, was wir mitgemacht haben. Mir hat Gott meine lieben

EHRI-ET-WL1375B310

- 3 -

Wien, März 1939

Liebe Tante,

Ich sende Dir einen Brief von meiner Tante. Hansi ist jetzt immer bei Frau F., bis die Tante nach Hause kommt.

Viele Russert von Leni.

Aus den Gefängnis, März 1939

Mein liebes gutes Tootchen,

Gestern war mein Leuchen bei mir und hat mir eingestanden, dass sie Dir die ganze Wahrheit geschrieben hat. Liebe Tante, wirst Du jetzt schlecht von mir denken? Ich weiß nicht, ob Leni Dir alles genau gesagt haben hat, ich kann hier nicht soviel mit ihr sprechen. Man hat uns alles weggenommen, wir sind nur mit dem was wir an Liebe gehabt haben, da gestanden, ich habe in der Aufregung einige Worte gesagt, die ich nicht hätte sagen sollen, und deshalb bin ich schon 3 Monate hier. Demke deswegen nicht schlecht von mir. Schon Liebes Tantele, Dein Bruder war doch auch voriges Jahr 13 Wochen hier, wo ich bin, das weist Du doch und deswegen war er auch nicht schlecht. Ich danke Dir viel tausendmal für die schönen Sachen, wenn ich sie nicht hätte, so müsste ich immer in einem Kleid herumgehen. Sie passen mir sehr gut, nur die Schuhe sind mir zu gross aber ich kann sie auch tragen. Du kannst Dir nicht vorstellen, wie ich mich gefreut habe mit den Sachen und hauptsächlich, weil sie von Dir sind.

Ich habe immer an Dich gedacht. Sei mir nicht böse, dass Leni Dir geschrieben hat, dass ich im Spital bin, ich habe mich geschämt, Dir die Wahrheit zu schreiben. Bitte schreibe der Tante Berta nichts davon. Wenn ich auch hier bin, so versage ich trotzdem nicht und ich hoffe, dass ich in kurzer Zeit bei meinem geliebten Baberl sein werde. Leni ist ein gutes Kind, wenn ich sie nicht hätte, so wäre ich ganz verlassen. Ich werde sie deswegen immer bei mir halten. Ich würd'lich freuen, wenn Leni mir ein Mitwoch von Dir einen Brief bringen würde. Sie kommt jeden Mitwoch zu mir. Tausend Grüsse und Kusse

Deine Marthe.

Grüsse Deinen Mann und Tochter und Tante Anna herzlich.

Ich sende Dir eine Aufnahme von mir, sie ist von hier und ist nicht gut. Den Brief gebe ich Leni mit. Hoffentlich kommt er gut an. Bitte nochmals, sage der Tante Berta nichts davon.

Wien, April 1939

Liebes gutes Tantele,

Habe Deinen Brief erhalten. Das Packet habe ich noch nicht, aber das dauert immer länger. Ich werde Dir gleich schreiben, wenn es kommt. Dieses Mal war es nicht möglich, einen Brief von Tante Martha bei zu lassen. Der "Onkel" hat zu gut aufgepasst. Sie darf weder schreiben noch Briefe empfangen. Liebes Tantele, Du schreibst, wir sollen Gottvertrauen haben. Ich glaube nicht an Gott. Warum hat er uns so gestrafft? Ich kann Dir nicht schildern, was wir mitgemacht haben. Mir hat Gott meine lieben

DLA

Automatic text recognition on segmented components

Transformation of combined DLA and text recognition outputs into <TEI>

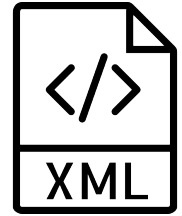
# ENRICHING THE TRANSCRIPTIONS SEMI-AUTOMATICALLY





# TEI-XML

- ❑ **XML standard** for encoding textual documents
  - ❑ Sustained by the **TEI Consortium** and improved by the involvement of its community
- ❑ **Semantic encoding** of the text (e.g. useful for **disambiguation tasks**)



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



Text Encoding Initiative

- ❑ TEI Guidelines = **recommendations**
- ❑ EHRI ODD = **customization of the TEI for Holocaust testimonies**
  - ❑ Contains both **specifications** and **documentation**
  - ❑ **Framework** for the encoders

# Homogenizing the encoding

- ❑ **English** = main language for **metadata** encoding
- ❑ ISO norms for **data reusability**
  - ❑ **ISO 639** for **language codes** ⇨ [iana Language Subtag Registry](#)
  - ❑ **ISO 3166** for representing the **names of countries** ⇨ `<country key="DE">Germany</country>`
  - ❑ **ISO 8601** for **dates** ⇨ `<date when-iso="YYYY-MM-DD" />`
- ❑ Integration of the **EHRI controlled vocabulary** on the [EHRI Portal](#)
- ❑ **Avoid encoding mistakes** generated by manual encoding
  - ❑ `@type="subeject"`
  - ❑ `@type="subjekt"`

# Automating parts of the encoding

**Ready-to-use templates** for an encoding by hand

- ❑ Use of the **recurrent metadata** present in the EHRI Online Editions
- ❑ **Templates** for the files and the indexes

**Python scripts** for a semi-automated encoding

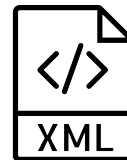
- ❑ Created along the model of **this script** and **this script**
- ❑ Use of **punctuations, lines and recurrent elements** to encode part of the document such as the structure, the keywords, etc.
- ❑ **Script** already done to **update** the online editions already created

**PROPOSING A CENTRALIZED  
INTERFACE FOR FUTURE  
HOLOCAUST-RELATED EDITIONS**



# TEI Publisher is ...

- ❑ a platform based on eXist-db, a database management system based on **XML technology**
- ❑ an **easy-to-use tool** to publish TEI XML files or other types of XML
- ❑ a ***prêt-à-porter application*** customizable with few tweaks



# The EHRI TEI Publisher application

- ❑ Generation of a **TEI Publisher application** dedicated to the EHRI editions
- ❑ Retrieval of the **concept and themes** of the websites already existing
- ❑ Addition of the **homogenized XML files**
- ❑ Creation of **XML indexes** containing the information of all **named entities from the editions**

# The EHRI TEI Publisher application

## Perks of the EHRI TEI Publisher application

- ❑ **Centralization** of information
  - ❑ All the collections are in the **same place**
  - ❑ Display of the text, image, map, and even named entities information at the **same level**
- ❑ Easier **accessibility** of the whole content
- ❑ Several **filter options** available to look through a collection

## EHRI Online Editions

Trier par

Titre

Filtrer selon

Titre

Q Filtrer



### Bordered Escape

This edition gathers documents testifying of the increasingly restrictive refugee policy in Czechoslovakia, following the 'Anschluss.



### Early Holocaust Testimonies

This edition gathers accounts of the persecution of Jews from the Nazi takeover of power in Germany (1933) to the Eichmann Trial (1961).



### Diplomatic Reports

This edition gathers reports on the persecution and murder of European Jews.



### Nisko Deportations

This edition gathers documents on the deportation of thousands of Jews to Nisko am San (Poland).



### Documentation Campaign

This edition gathers documents on the deportation of thousands of Jews to Nisko am San (Poland).

TEI Publisher 0.1 / web components 2.19.0 / API 1.0.0

This application was developed as part of the EHRI project with the help of TEI Publisher.

Powered by  
e/editions



Trier par

Titre

Filtrer selon

Titre

Filtrer

[|<](#) **1** [2](#) [3](#) [4](#) [5](#) [>|](#) 125 résultats

ALLER AU PARENT

**A female social activist, on the first months of German occupation of Łódź**  
**Une militante sociale, sur les premiers mois de l'occupation allemande de Łódź**

A long, very detailed report by a 38-year-old female social activist about the situation of Jews in Łódź (and surrounding towns) in the first months under German occupation. She left Łódź on December 30, 1939.

TÉLÉCHARGER

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

Testimony of a fifteen-year-old youth, recorded on 15 November 1939 and describing the German invasion of Wyszaków 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

**A. K., male, on his labour service in the Hungarian army and his time as a POW of the Soviet Army**  
**A. K., homme, sur son service de travail dans l'armée hongroise et son temps comme prisonnier de guerre de l'armée soviétique**

Testimony of 31-year-old Dr. A.K. on his hardships as a labor serviceman in the Hungarian Army on the Eastern front in 1942/43, where he deserted along with 41 comrades and joined the partisans.

## Filters

## From

 Montrer les 50 premiers

- Berthold Burg 1
- Dwojra Szczucińska 1
- Fischer Schaechter 1
- Isaak Berner 1
- Leon Perelsztejn 1

## Language

 Montrer les 50 premiers

- Czech 15
- English 15
- German 26
- Hungarian 20
- Yiddish 36

## Date

 Montrer les 50 premiers

- 1939 7
- 1940 7
- 1944 5
- 1945 50
- 1946 11

## Conservation site

 Montrer les 50 premiers

Trier par Filtrer selon  
 Titre Titre Filtrer

&lt; 1 &gt; 3 résultats

ALLER AU PARENT

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

Testimony of a fifteen-year-old youth, recorded on 15 November 1939 and describing the German invasion of Wyszaków 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

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

Testimony of Shloyme 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

View Original version

קאמיטעט צו זאמלען מאטעריאלן

וועגן יידישן חורבן אין פוילן 1939

י.מ.ש.

ווישקאָוו

יאָר אַלט, ביי די עלטערן 18

פּראָטאָקאָל נומער 3

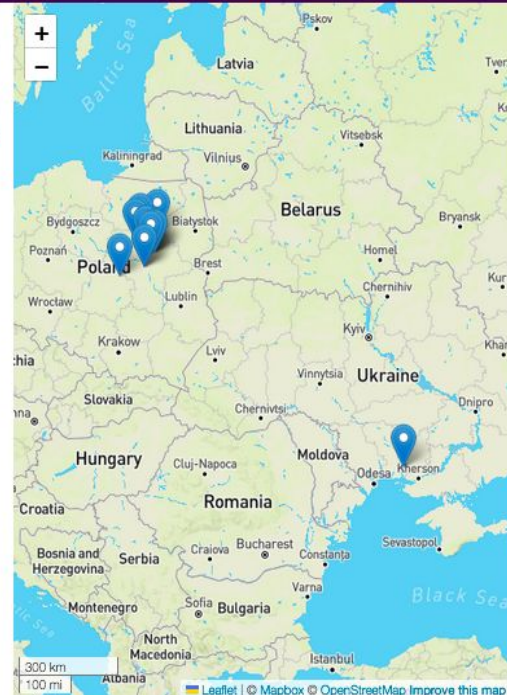
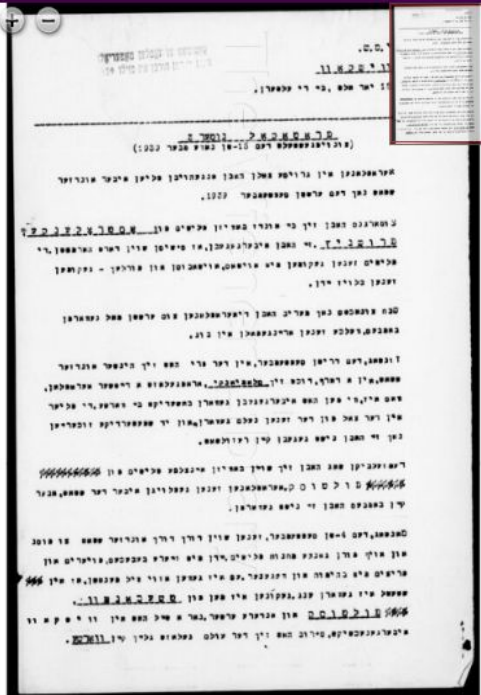
(צונויפגעשטעלט דעם 15טן נאָוועמבער 1939)

אַעראָפּלאַנען אין גרויסע צאָלן האָבן אָנגעהויבן פֿליען איבער אונדזער שטאָט נאָך דעם ערשטן סעפטעמבער 1939.

צומאָרגנס האָבן זיך ביי אונדז באַוויזן פּליטים פֿון אַסטראַלענקע, פּרוּשניץ. זיי האָבן איבערגעגעבן, אַז ס'שיסן שוין דאָרט האַרמאַטן. די פּליטים זענען געקומען מיט אויטאָס, אויטאָבוסן און פּוּרלעך - געקומען זענען בלויז יידן.

שבת צו נאַכטס נאָך מעריב האָבן די אַעראָפּלאַנען צום ערשטן מאל געוואָרפֿן באַמבעס, וועלכע זענען אַרטינגעפּלאַן אין בוג.

זונטאָג, דעם דריטן סעפטעמבער, אין דער פֿרי האָט זיך הינטער אונדזער שטאָט, אין





View Translation

Committee to Collect Material  
about the Destruction of Polish Jewry 1939

Y. M. Sh.

Wyszaków

Age: 15 years; living with his parents

Testimony number 3

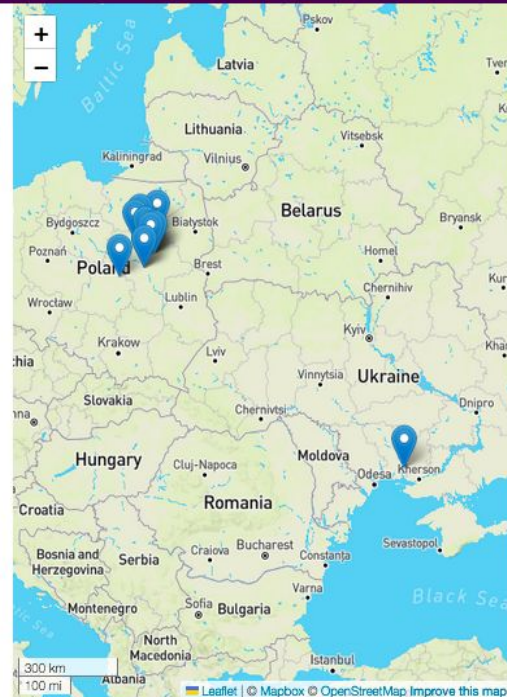
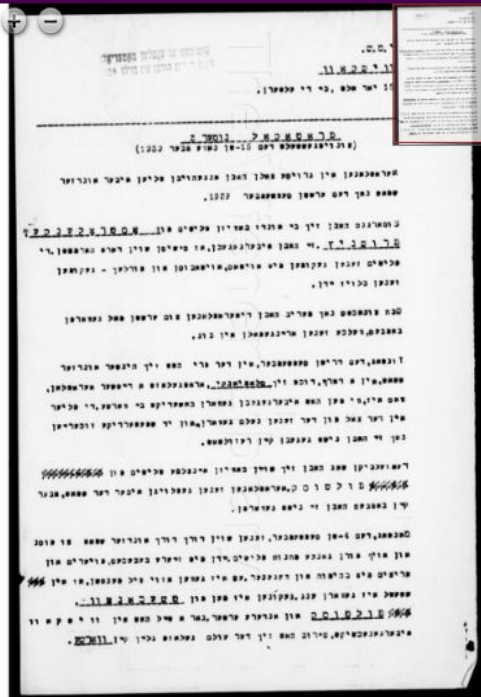
(recorded on the 15th of November, 1939)

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



Metadata

Title

A fifteen-year-old youth, on the German invasion of Wyszków 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 :

...ounding areas — Un jeune de quinze ans, sur l'invasion allemande de Wyszków et des environs

...nd his three small children. A ...rescued.

...planes. Other escaped in the

...way.

...parently they disappeared the

...l Jews was very good. People

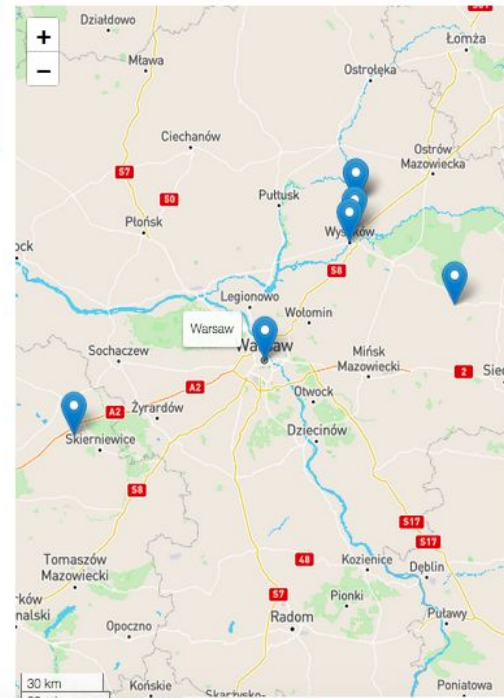
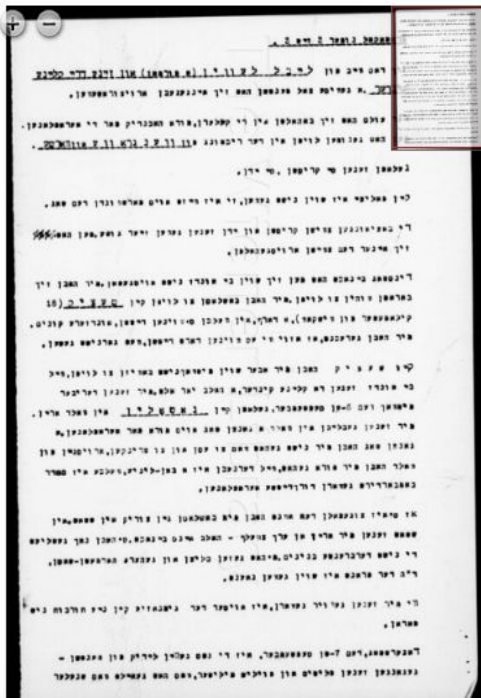
...undressed for the night. We

...ded to run to the village of

...re some Germans were living,

...would happen in a place where

...escape to Siczczychy, because in



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**. r of the airplanes. Other escaped in the directions of **węgrów** and **warsaw**.

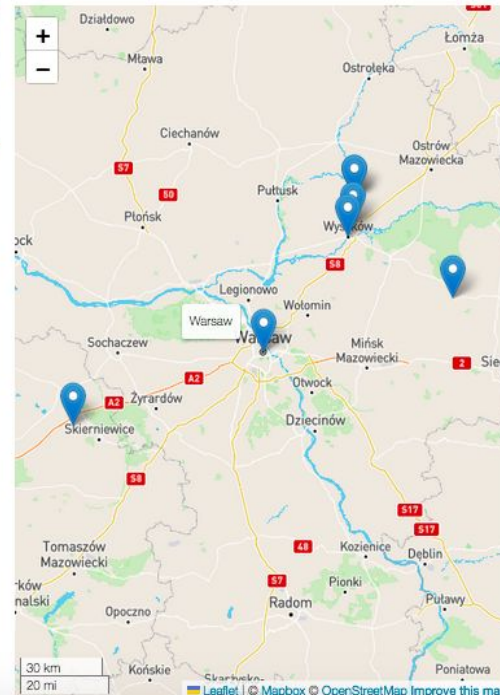
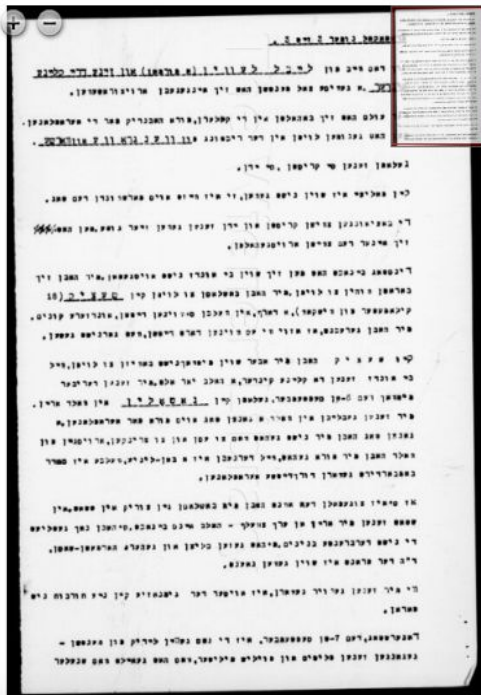
Both Christians and Jews were running away.

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,





## 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 Koniuchowsky 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" (Dokumentační 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](#).

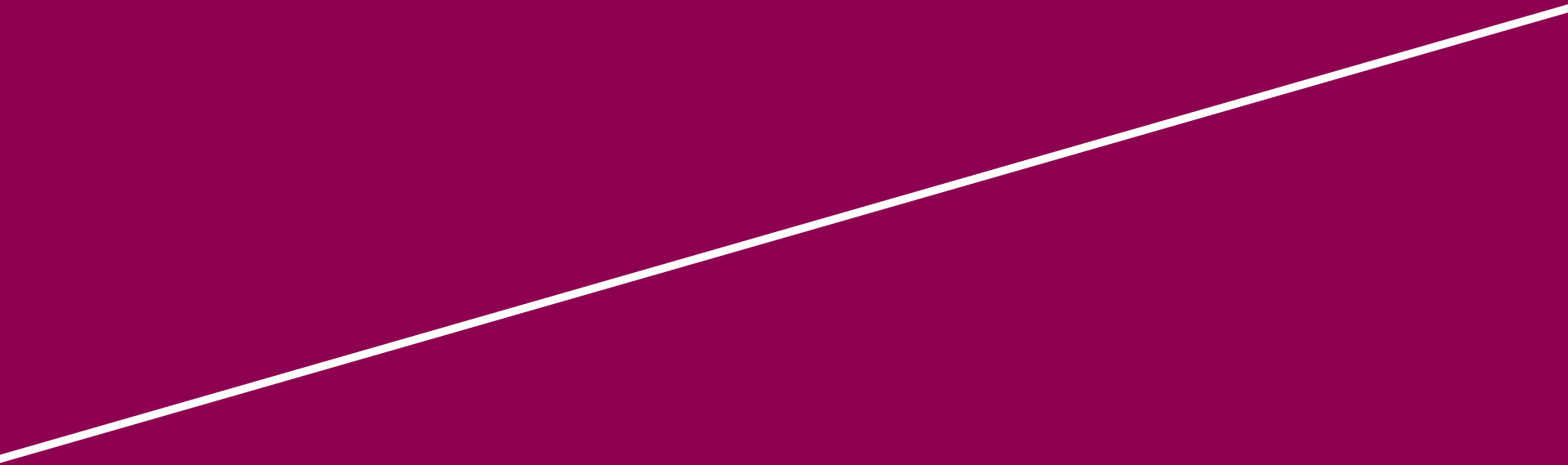
# CONCLUSION



# Conclusion

- ❑ **Feedback** for the use of one or several of our solution
- ❑ Layout analysis
  - ❑ **Sample** other EHRI Online Editions
  - ❑ **Annotate** more images
  - ❑ Have a **more balanced class representation**
  - ❑ **Apply DLA model** on new Holocaust-related documents to test the model's robustness on unseen data
- ❑ Finishing the semi-automated encoding **scripts**
- ❑ Possible **online launch** of the application

# RESOURCES



# Web resources

- ❑ EHRI dataset for ATR: <https://github.com/FloChiff/ehri-dataset/tree/main>
  - ❑ Presentation of the EHRI ATR dataset: <https://flochiff.github.io/phd/dataset/ehri/dataset.html>
- ❑ EHRI dataset for layout segmentation: <https://universe.roboflow.com/ehri/ehri-ladas>
- ❑ ODD EHRI: <https://github.com/SarahBeniere/EHRI-Workflow/tree/main/ENCODING/Guidelines>
- ❑ TEI Publisher: <https://teipublisher.com/index.html>
- ❑ eScriptorium: <https://escriptorium.inria.fr/>
- ❑ EHRI Online Editions: <https://github.com/EHRI/ehri-online-editions>

# Bibliography

- Bénière, S., Chiffolleau, F., & Romary, L. (2024). TEI Specifications for a Sustainable Management of Digitized Holocaust Testimonies. *Proceedings of the First Workshop on Holocaust Testimonies as Language Resources (HTRes) @ LREC-COLING 2024*, pages 10–17, Torino, Italia. ELRA and ICCL. <https://aclanthology.org/2024.htres-1.2/>.
- Chagué, A., Scheithauer, H., Terriel, L., Chiffolleau, F. & Tadjou-Takianpi, Y. (2022). Take a sip of TEI and relax: a proposition for an end-to-end workflow to enrich and publish data created with automatic text recognition. *Digital Humanities 2022 : Responding to Asian Diversity*, ADHO; University of Tokyo, Tokyo, Japan. [\(hal-03739767\)](#)
- Chiffolleau, F. & Scheithauer, H. Leveraging (2024). EHRI Online Editions for training automated edition tools. *Workshop Natural Language Processing Meets Holocaust Archives*, EHRI-3, Prague, Czech Republic. [\(hal-04594084\)](#)
- Chiffolleau, F. & Scheithauer, H. (2022). From a collection of documents to a published edition : how to use an end-to-end publication pipeline. *TEI 2022 - Text Encoding Initiative 2022 Conference*, Newcastle, United Kingdom. [\(hal-03780316\)](#)
- Clérice, T. (2023). You Actually Look Twice At it (YALTAi): using an object detection approach instead of region segmentation within the Kraken engine. *Journal of Data Mining & Digital Humanities*, 2023. <https://doi.org/10.46298/jdmdh.9806>.
- Clérice, T., Janès, J., Scheithauer, H., Bénière, S., Romary, L., & Sagot, B. (2024). *Layout Analysis Dataset with SegmOnto*. DH2024 - Annual conference of the Alliance of Digital Humanities Organizations, Washington, D.C., United States. [\(hal-04513725\)](#)
- Gabay, S., Pinche, A., Christensen, K., & Camps, J.-B. (2023). SegmOnto: A Controlled Vocabulary to Describe and Process Digital Facsimiles. [\(hal-04343404\)](#)

# Thank you for your attention

*Any questions ?*

Contact :  
*floriane.chiffoleau[at]inria.fr*  
*sarah.beniere[at]inria.fr*