



HAL
open science

Leveraging EHRI Online Editions for training automated edition tools

Floriane Chiffoleau, Hugo Scheithauer

► To cite this version:

Floriane Chiffoleau, Hugo Scheithauer. Leveraging EHRI Online Editions for training automated edition tools. EHRI Workshop Natural Language Processing Meets Holocaust Archives, EHRI-3, Mar 2024, Prague, Czech Republic. hal-04594084

HAL Id: hal-04594084

<https://inria.hal.science/hal-04594084v1>

Submitted on 30 May 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

Leveraging EHRI Online Editions for training automated edition tools

Paper session



ALMAnaCH project-team

Inria

Floriane Chiffoleau (PhD) and Hugo Scheithauer (PhD), Inria
Workshop *Natural Language Processing Meets Holocaust Archives*
Thursday, 28 March 2024

Summary of the presentation

1. **Introduction**
2. **Layout analysis** on the editions
3. **Automating** the transcription
4. **Homogenizing** the encoding
5. Publishing the editions with **TEI Publisher**
6. **Conclusion**
7. **Resources**

INTRODUCTION

A little 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

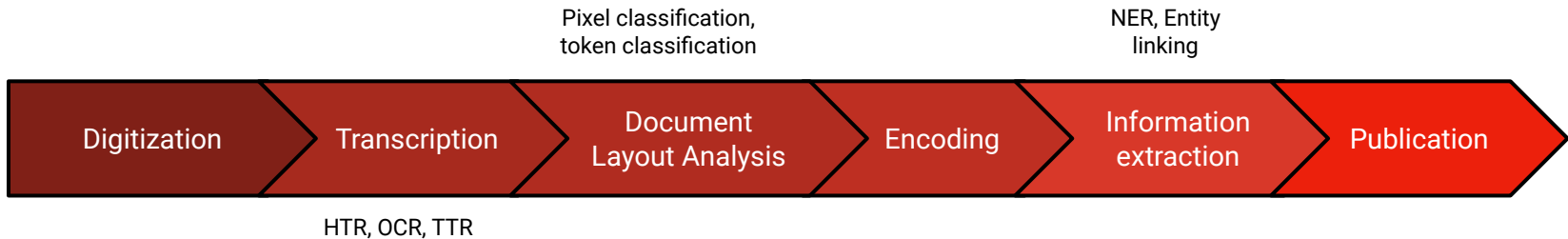
Hugo Scheithauer



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

Thesis topic: Document Layout Analysis and Document Understanding

What are we doing ?

- At ALMAAnCH research team, Inria Paris, DH engineers and researchers mainly work with **text-related objects** (digitized documents, PDFs or raw texts) and put the emphasis on **programming, data creation,** and **machine learning engineering**. We also focus on processing textual data with NLP solutions.



- Open source, open data, open science  
- We also aim at building software and machine learning solutions that can easily be used by institutional partners and the DH community in general.

What are we doing ?

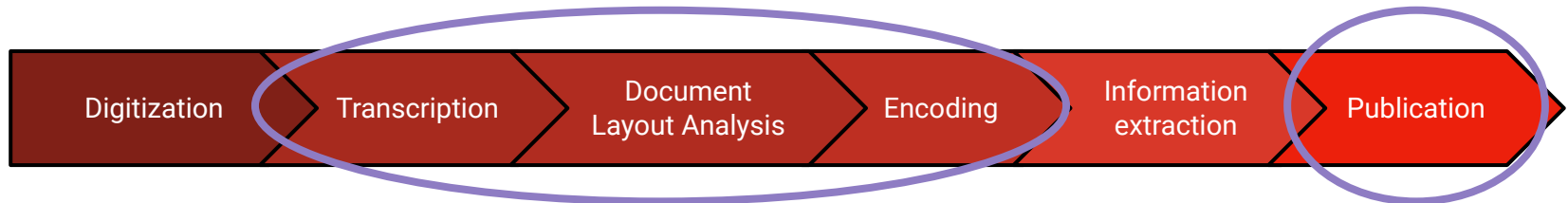
The EHRI consortium has supported the development and publication of six Holocaust-related online editions.

- **Early Holocaust Testimonies Edition:** Testimonies kept in the Wiener Holocaust Library, London, Yad Vashem, Jerusalem, the Jewish Historical Institute in Warsaw, the Hungarian Jewish Archives in Budapest, and the Jewish Museum in Prague. Originals are in Czech, German, Hungarian, Polish, Dutch, and Yiddish.
- **Documentation Campaign Edition:** Holocaust survivor testimonies held by the Jewish Museum in Prague and by Yad Vashem, Jerusalem.
- **Diplomatic Reports Edition:** Reports from the diplomatic staff of Denmark, Italy, Japan, Hungary, Slovakia, and the United States.
- **Von Wien ins Nirgendwo - Die Nisko-Deportationen 1939 Edition:** Documents on the history of the Viennese Jewish deportees to Nisko, held in various institutions.
- **Begrenzte Flucht Edition & Uzavřít Hranice Edition:** Documents kept in Czech, Austrian and other archives and related to the crisis year 1938 at the Czechoslovakia border and Austrian refugees.

→ **So a multilingual and multi-domains setting.**

What are we doing ?

For today's presentation, we will talk about:

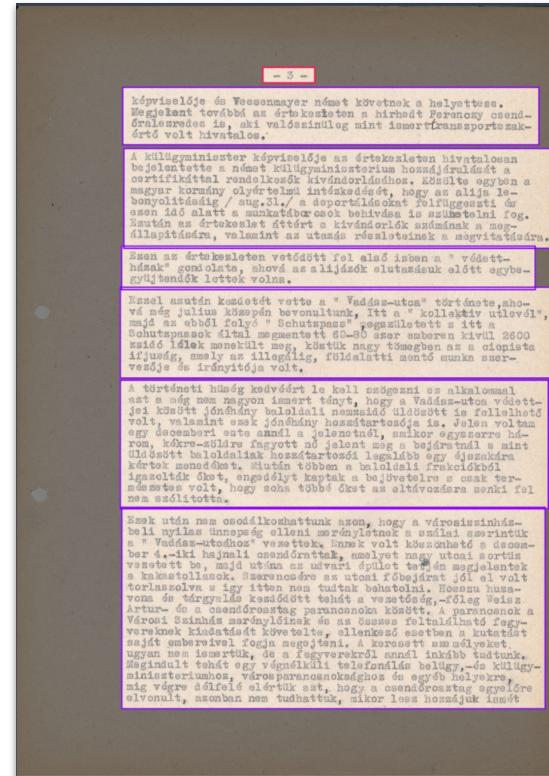


SEMI-AUTOMATIC ENCODING WITH LAYOUT ANALYSIS



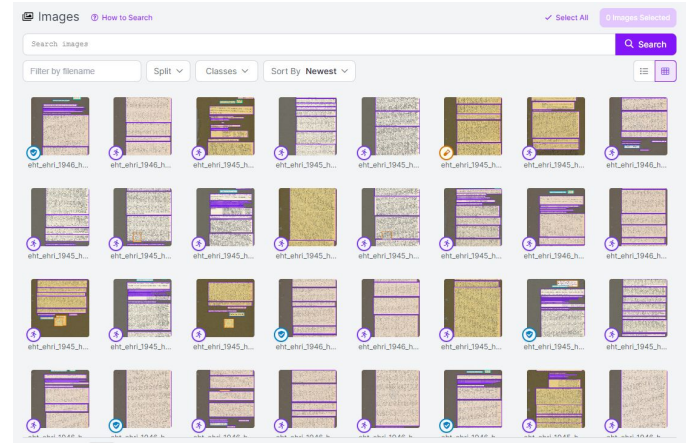
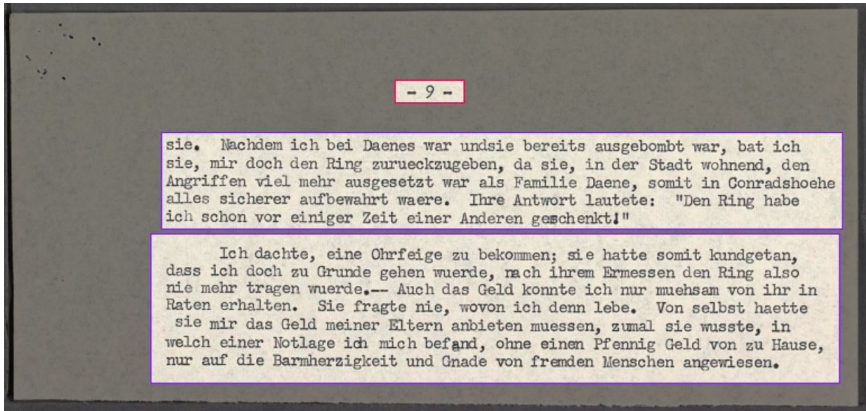
Semi-automatic encoding with layout analysis

- Document Layout Analysis (DLA): a task that aims at detecting the layout components and hierarchy of a document.
- SOTA DLA systems are based on two approaches: **visual features** (YOLOv8), and a combination of **textual and visual features** (LayoutLM).



Semi-automatic encoding with layout analysis

We sampled and annotated **200 images** from the documents available in the **Early Testimonies EHRI Online Edition** with the **Roboflow** (<https://roboflow.com/>) interface in order to train a YOLOv8 model (object detection, pixel based).



Thibault Clérice - 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, 26 décembre 2023, Documents historiques et reconnaissance automatique de texte - <https://doi.org/10.46298/jdmhdh.9806>

Thibault Clérice, Juliette Janes, Hugo Scheithauer, Sarah Bénére, Laurent Romary, et al.. Layout Analysis Dataset with SegmOnto. DH2024 - Annual conference of the Alliance of Digital Humanities Organizations, ADHO, Aug 2024, Washington DC, United States. (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
nagyvezetési eljárástól írták ki, hogy az alábbi be-
bonyolításig / aug.31./ a deportálásokat felülvizsgálati és
ezen idő alatt a munkatársakok behívása is szabotálni fog.
Számtalan értekezlet történt a kivándorlók számára a meg-
állapítására, valamint az utazás visszatérési a megvitására.

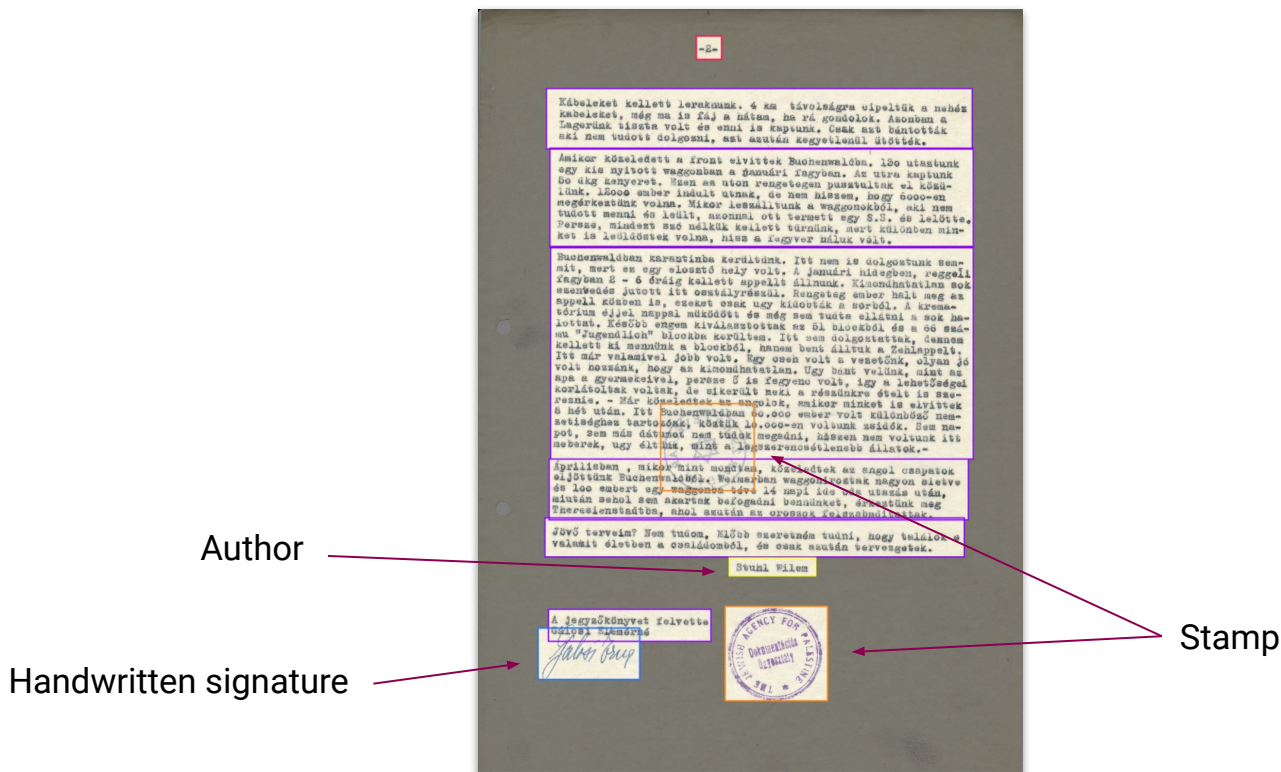
Ezen az értekezleten vetődött fel első ízben a "vándor-
hírek" gondolatja, ahová az eljövők elutazásuk előtt egybe-
gyűjtendőek lettek volna.

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 mentő munka esze-
resztése és irányítása volt.

A történeti háttér kedvéért le kell szögezni az alkalommal
arról, hogy nem nagyon ismert tény, hogy a Vadász-utazás vándor-
hírek közötti időközben beloldali nemcsak ülésszék is folyhatott
volt, valamint ezek időközben hosszátartóssága is. Jelen voltam
egy decemberi este emiatt a jelentésről, amikor egyszerre há-
rom, három-ötletre írtam meg a jelentés meg a bejárattal a mint
ülésszék beloldali hosszátartóssági legalább egy éjszakára
kértem megmondást. Mivel többem a beloldali frakciókbeli
igazságtörés és, engedélyt kaptak a bejárattal a csak ter-
vezéses volt, hogy soha többé őket az átjárásra senki fel
nem szállította.

Ezek után nem oszálkozhattunk azon, hogy a városi lakó-
béli nyilas ünnepéig elleni szerénylőnek a szílei szerintük
a "Vadász-utazás" vezetők, ennek volt köszönhető a decem-
ber 4-iki hajnali ülésszékünk, amelyet nagy utcai soros
vezetett be, majd utána az utcai épület területén megjelentek
a katonatisztek. Szerencsére az utcai főbejárat jól el volt
szervezve a egy itten nem voltak behatolni. Hosszú husa-
vona és tárgyalás kezdődött tehát a vezetőség, főleg Weisz
Artúr és a csendőrszolgálat parancsnoka között. A parancsnok a
Városi lakóház szerénylőnek és az utasok feloldható fegy-
vereknek kiderítését követően, ellenkező esetben a katonák
majd embereivel fogja segíteni. A kérését amúgyis
ugyan nem ismertük, de a fegyverekről annál inkább tudunk.
Segítség volt az a végül eljött telefonos beszélgetés, és külgy-
minisztériumhoz, városparancsnokokhoz és egyéb helyekre,
sőt végül előfordult az, hogy a csendőrszolgálat egyelőre
elvonult, azonban nem tudtuk, akkor lesz hozzájuk hozzánk

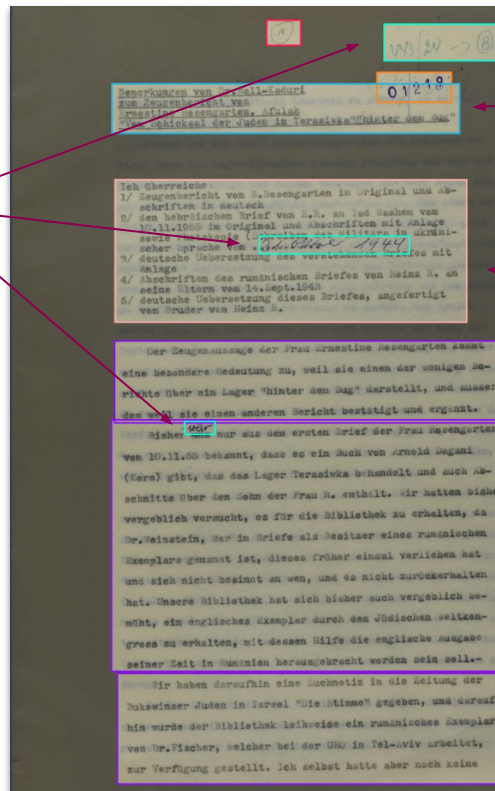
Semi-automatic encoding with layout analysis



Semi-automatic encoding with layout analysis

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

Handwritten annotation



Heading

List

Semi-automatic encoding with layout analysis

As of March 2024, we identified 13 classes that are relevant for early holocaust testimonies:

- Date
- Form
- Head
- List
- Paragraph
- Signature
- Handwritten signature
- Handwritten annotation
- Notes
- Page number
- Running title
- Stamp
- Sticker

This list is open as new use cases can arise.

Semi-automatic encoding with layout analysis



We use the SegmOnto controlled vocabulary, a TEI-based ontology, to annotate our sample dataset.

By using SegmOnto, the annotated dataset is easily interoperable and reusable.

See SegmOnto's website: <https://segmonto.github.io/> and Simon Gabay, Ariane Pinche, Kelly Christensen, Jean-Baptiste Camps. SegmOnto: A Controlled Vocabulary to Describe and Process Digital Facsimiles. 2023. [hal-04343404](#).

Semi-automatic encoding with layout analysis



- Date > MainZone:Date
- Form > MainZone:Form
- Head > MainZone:Head
- List > MainZone:List
- Paragraph > MainZone:P
- Signature > MainZone:Signature
- Handwritten signature > MarginTextZone:Signature
- Handwritten annotation > MarginTextZone:ManuscriptAddendum
- Notes > MarginTextZone:Notes
- Page number > NumberingZone
- Running title > RunningTitleZone
- Stamp > StampZone
- Sticker > StampZone:Sticker

Semi-automatic encoding with layout analysis

838
- 6 -

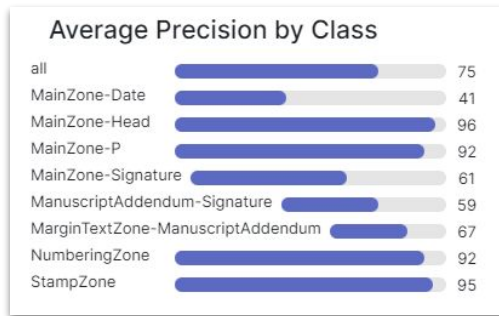
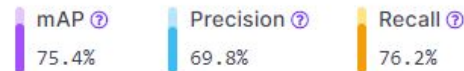
gingen jüdisch
und das Schwarze Meer ging, zahlenmassig aber nicht sehr ins

Ende März, oder Anfang April 1940 - das Land war noch
inner neutral und die Verbindungen mit dem Westen daher nicht
voellig abgerissen - kam ein Vertreter der Pariser Zentrale der
GTF - Organisation nach Budapest und verhandelte mit den dortigen
Gemeindevertretern, aber einem Ausbau des Berufsamtchungs-
Arbeit. Als Ergebnis seiner Verhandlungen wurde eine Budapest-
Ortsgruppe des GTF gebildet, die das Umschichtungsgewerk in ihre
eigene Leitung uebernahm und auch die bereits vorhandenen Lehrg-
gange weiterleitete. Von der Budapest-Ortsgruppe wurde das
natuerlich mit Fremden begrenzt, da es fuer sie eine finanzielle
Entlastung bedeutete. Die GTF-Gruppe unterhielt ein eigenes Buero
in der Ringstrasse, doch geriet ihre Arbeit spaeater, nach
in den Krieg voellig ins Stocken.

Neben der Arbeitsvermittlung und der Berufsamtchungs-
galt mein Interesse vor allem einem anderen Arbeitsgebiet, das
gerade zu jener Zeit, Anfang 1940 eine wachsende Bedeutung
bekam, naemlich der Fluechtlingsbewerger. Im vergangenen Jahre
waren aus Polen viele Juden vor der deutschen Invasion ueber
die Karpathen nach Ungarn gefluchtet, aber auch aus staemigen
Oesterreich und anderen mit-ueberwunden Gebieten kamen viele
mit oder ohne Pass nach Ungarn. Die ungarische Regierung behan-
delte sie sunstrecht human und machte voererst keinen Unterschied
zwischen juedischen und nichtjuedischen Fluechtlingen aus Polen.
Es muss bemerkt werden, dass beim Zusammenbruch der polnischen
Armee im September 1939 viele Tausende von Polen aller Katego-
rien, teils von den von Westen vordringenden Heere teils von
den von Osten anrückenden Bolschewisten ueber die Karpathen-
pässe fluechteten und in Ungarn und Rumänien Zuflucht suchten.
Sogar kleinere Militaerseinheiten bis zu Bataillonsstaerke unter-
schritten, geschloesen die Grenze und liessen sich hier entwaffnen.
Die ungarischen Behoerden internierten sie zuerst, liessen ihnen
dann aber grosse Bewegungsfreiheit und drueckten sie gegen zu,
wenn sie auf juedisches Gebiet weitergingen, um von dort zur
der Westmaechte zu stossen.

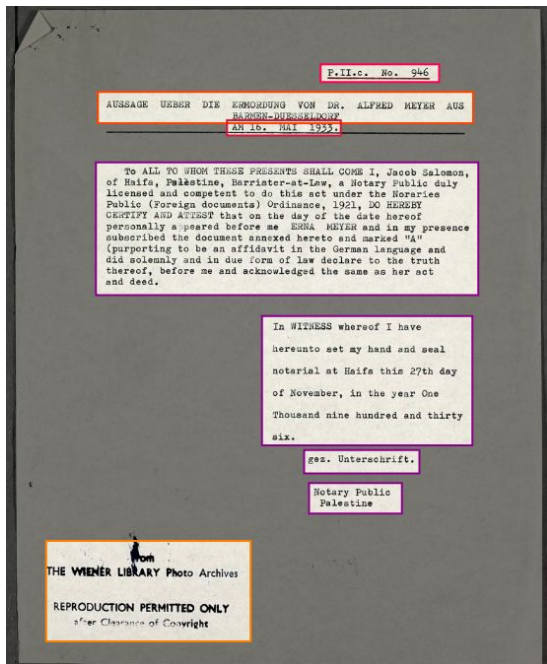
Anfang 1940 begannen jedoch die Behoerden ihre Haltung
gegenueber den Fluechtlingen zu sondern. Es bildete sich die
Frage heraus, dass die im Lande verbliebenen Fluechtlinge,
soweit es Nichtjuden waren, unter Polizeiaufsicht gestellt
aber auf freien Pass belassen, soweit es aber Juden waren, in
Internierungslager gesperrt wurden. Im Rahmen des Innenministe-
riums wurde eine besondere Behoerde, ENKKE genannt, geschaffen
und diese sollte allmaechlich die Torgew, alle nach Kriegsge-
suehr ins Land gekommenen Juden nichtungarischer
Staatsangehoerigkeit, d. h. staatenloses Juden und solche mit
zweifelhafter Staatsangehoerigkeit zusammenfuehren und zu inter-
nieren. Die Verschaeerung dieses Kurzes und die willkuerliche
Auslegung der Bestimmungen sollte spaeater auch mir und meiner
Familie zum Verhaengnis werden, doch ahnte ich in diesen Zeit-

5 objects detected

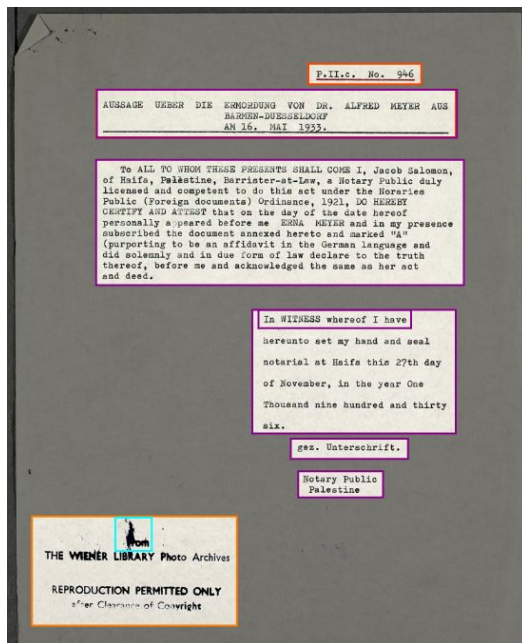


Preliminary experiment DLA model is available at:
<https://app.roboflow.com/ehri/ehri-ladas/visualize/3>

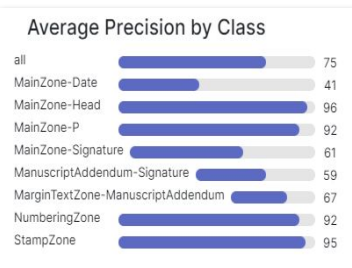
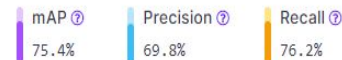
Semi-automatic encoding with layout analysis



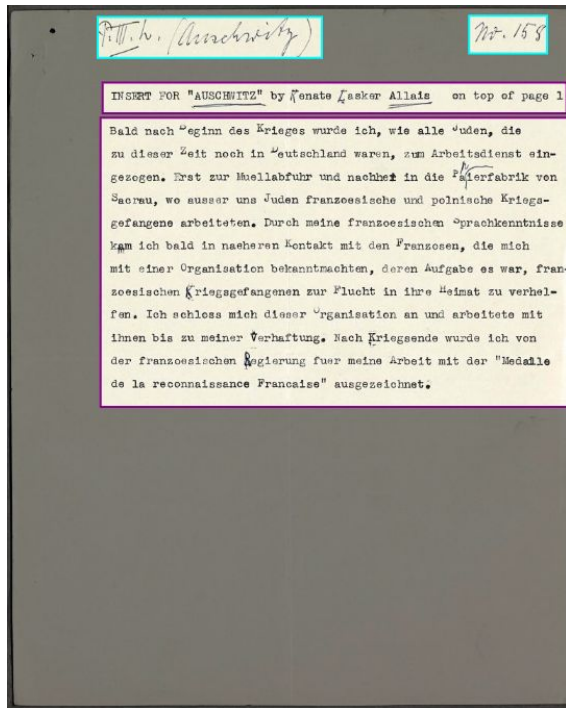
Ground truth



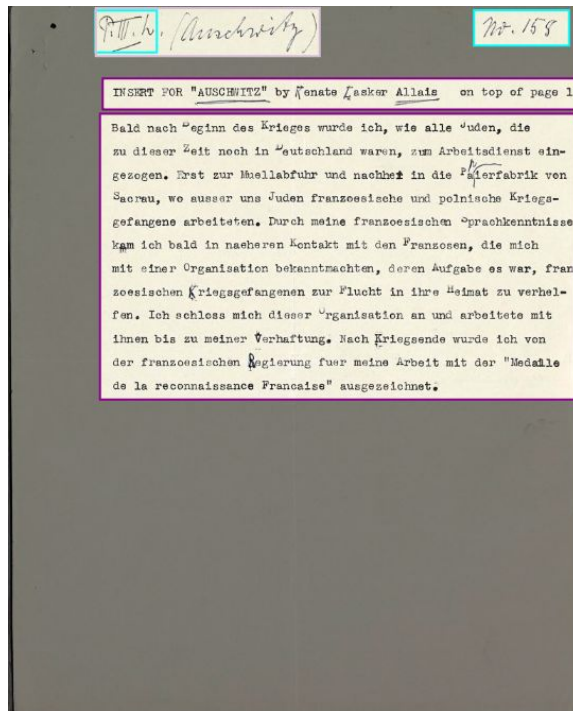
Prediction



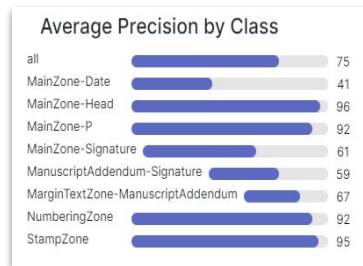
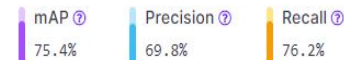
Semi-automatic encoding with layout analysis



Ground truth

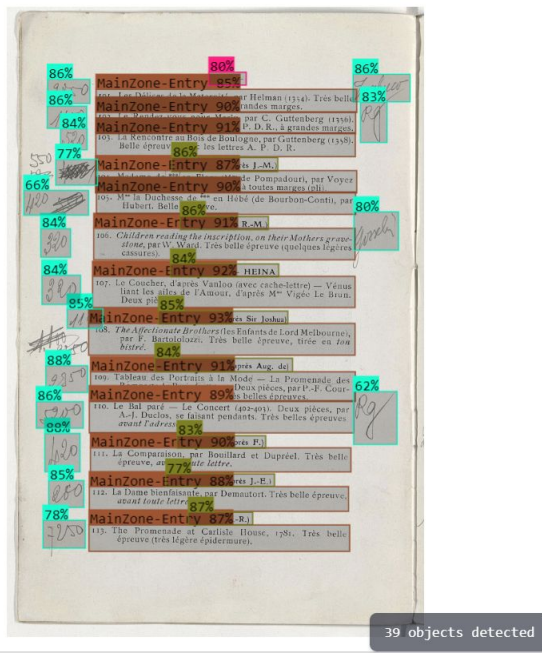
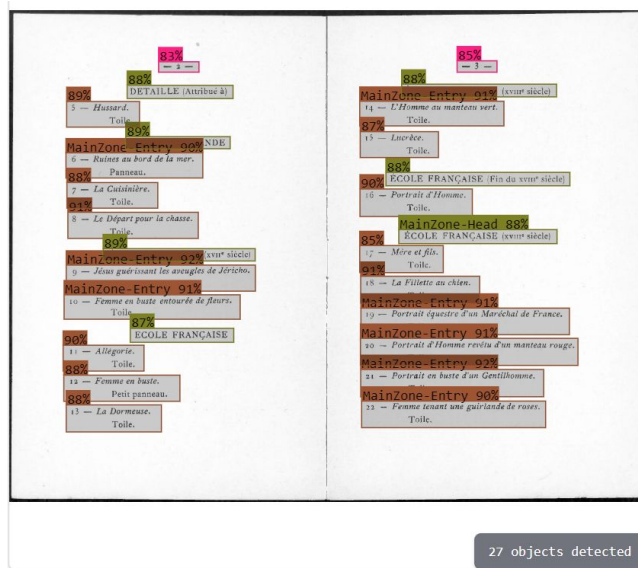


Prediction



Semi-automatic encoding with layout analysis

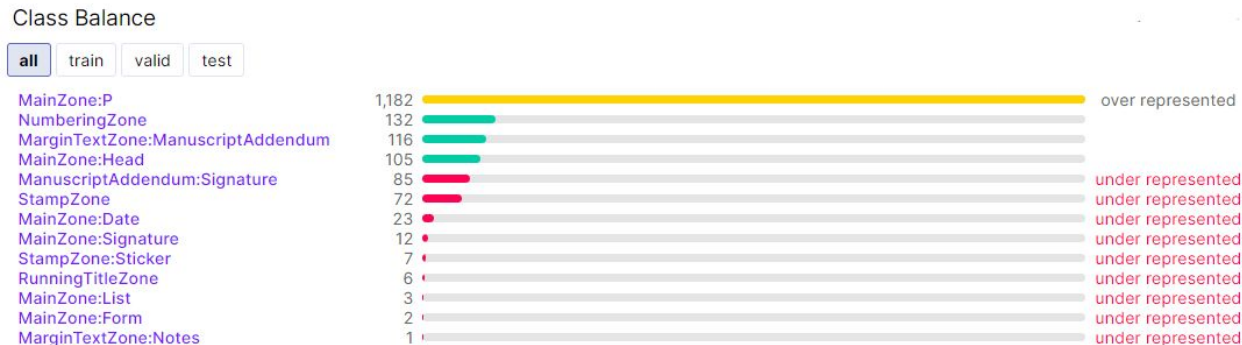
As a comparison, we annotated around 1400 images for the DataCatalogue project (Inria, French National Library, French National Institute for Art History), and the model achieves very good scores on complex and different layouts.



Semi-automatic encoding with layout analysis

Next step:

- 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



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

DLA

- 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

Automatic text recognition on segmented components

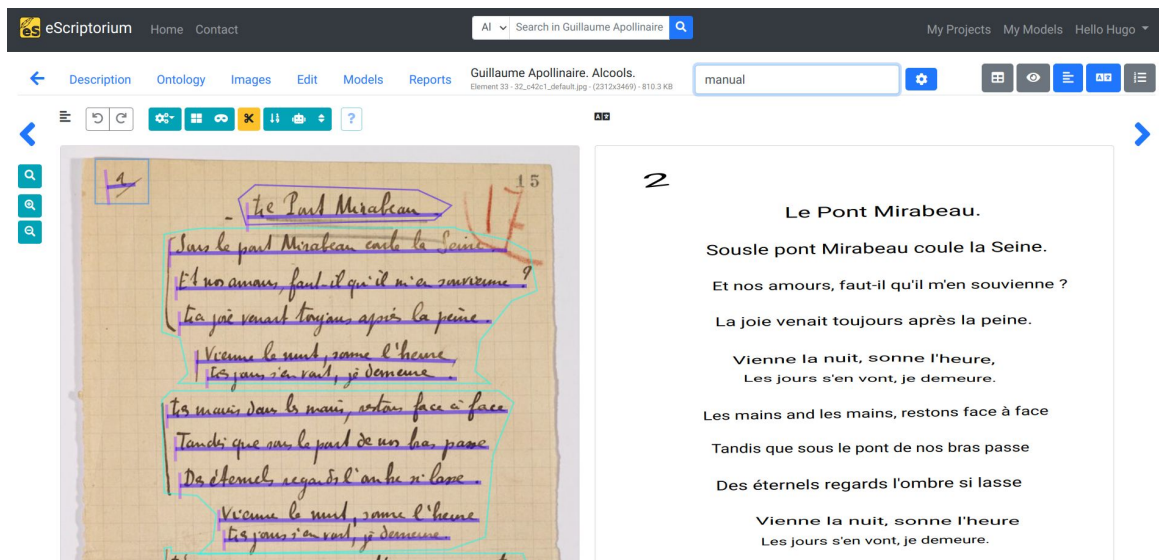
Transformation of combined DLA and text recognition outputs into <TEI>

AUTOMATING THE TRANSCRIPTION



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!).



The screenshot displays the eScriptorium web application interface. At the top, there is a navigation bar with the eScriptorium logo, 'Home', and 'Contact' links. A search bar contains the text 'Search in Guillaume Apollinaire'. On the right side of the navigation bar, there are links for 'My Projects', 'My Models', and 'Hello Hugo'. Below the navigation bar, a secondary menu includes 'Description', 'Ontology', 'Images', 'Edit', 'Models', and 'Reports'. The main content area is titled 'Guillaume Apollinaire, Alcools.' and shows a document with the title 'Le Pont Mirabeau'. The document image on the left shows handwritten French text with colored bounding boxes around the lines. On the right, the transcribed text is displayed in a clean, digital font. The transcribed text is as follows:

2

Le Pont Mirabeau.

Sous le pont Mirabeau coule la Seine.
Et nos amours, faut-il qu'il m'en souvienne ?
La joie venait toujours après la peine.
Viens la nuit, sonne l'heure,
Les jours s'en vont, je demeure.

Les mains and les mains, restons face à face
Tandis que sous le pont de nos bras passe
Des éternels regards l'ombre si lasse
Viens la nuit, sonne l'heure
Les jours s'en vont, je demeure.

A solution for dataset visibility: HTR-United

- ❑ Project created by Alix Chagué & Thibault Clérice (Inria)
- ❑ A catalogue of metadata describing free (and open) datasets of ground truth for HTR
- ❑ A standardization proposal for describing ground truth datasets ([schema](#))
- ❑ An ecosystem of tools for quality control and data publication
- ❑ A useful resource aiming for international visibility :
 - ❑ 81 datasets (as of early 2024)
 - ❑ 14 languages
 - ❑ from 9th century to nowadays



Website:

<https://htr-United.github.io/>

GitHub organization:

<https://github.com/HTR-United>

Creation of a multilingual recognition model

- ❑ Two **objectives**:
 - ❑ **Producing an efficient recognition model** for typescript documents, no matter the language
 - ❑ **Contributing to a PhD experiment** on the relation between the content of the ground truth's tokens and the recognition method of the model
- ❑ Creation of **training data** for the model
 - ❑ Dataset of **252 typescript documents**, from the four published editions, in **7 languages** (German, English, Danish, Hungarian, Polish, Slovak, and Czech)
 - ❑ Documents **segmented and transcribed** (copy/paste or manual)
 - ❑ Images, texts, and XML available at <https://github.com/FloChiff/ehri-dataset>
 - ❑ Presentation of the dataset at <https://flochiff.github.io/phd/dataset/ehri/dataset.html>

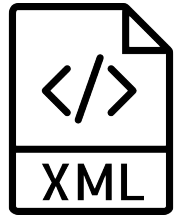
Creation of a multilingual recognition model

- ❑ Production of a model, with **97.20% of accuracy**, available [here](#)
- ❑ **Efficiency** of the model:
 - ❑ Working pretty well with the **languages it was trained on**, whether it had more or less training data on the language
 - ❑ Able to recognize the typescript content of documents in **languages it was not trained on** but with the same Latin script (tested with Italian and French)
 - ❑ Can have some trouble with the **diacritics and uppercase** as it learned many various ones (in the first case) or not enough (in the second)
- ❑ **Other models available**: production of single-language models for each language of the training data

HOMOGENIZING THE ENCODING

Homogenizing the encoding

- ❑ Creation of an **ODD for the EHRI editions**, constituted of a **specific XML schema**, instructing on what tags and attributes to use, supplied with an **integrated documentation**, explaining how to use the tags, in what occasion, what choice to make in some situation, etc.
 - ❑ The work has been done as part of an internship in 2023 by Sarah Beniere
 - ❑ The ODD is available online at <https://gitlab.inria.fr/dh-projects/workflow-ehri/-/tree/main/ODD>
 - ❑ This is to be presented more thoroughly at the EHRI Conference in June
- ❑ Following the indication of the ODD schema, the currently published **EHRI editions** have been **modified and homogenized**



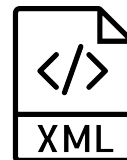
Text Encoding Initiative

PUBLISHING THE EDITIONS WITH TEI PUBLISHER



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



What can TEI Publisher do ...

- ❑ ... as the **main instance**?
 - ❑ Presentation of various **XML files, ODDs and templates** to observe the extent of the tool
 - ❑ **Playground area** to upload your own files and test the ODDs/templates to see what fits best
 - ❑ **Annotation area** to discover how to annotate (named entities or other) some already existing files or files you uploaded yourself
- ❑ ... as the **tool** that generates an application?
 - ❑ **Customized ODD and templates** to display your XML files exactly as you want
 - ❑ **Documentation or additional information** in HTML or Markdown format
 - ❑ PDF or EPUB **export** of the XML files

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

Filtrer



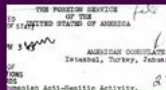
Bordered Escape

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).

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.



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 Titre Filtrer selon Titre Filtrer

< 1 > 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 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

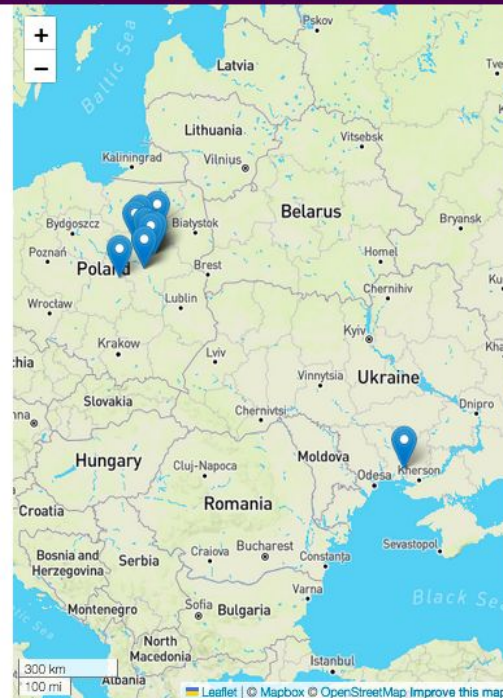
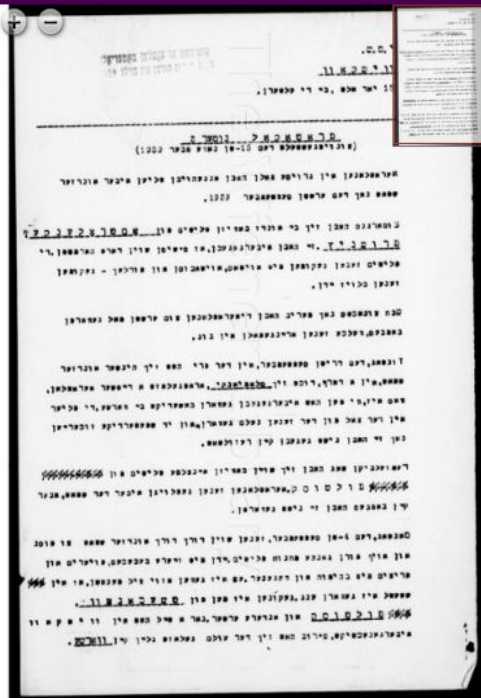
(recorded on the 15th of November, 1939)

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



View Original version

קאמיטעט צו זאמלען מאטעריאלן

וועגן יידישן חורבן אין פוילן 1939

י.מ.ש.

ווישקאָוו

יאָר אַלט, בײַ די עלטערן 18

פּראָטאָקאָל נומער 3

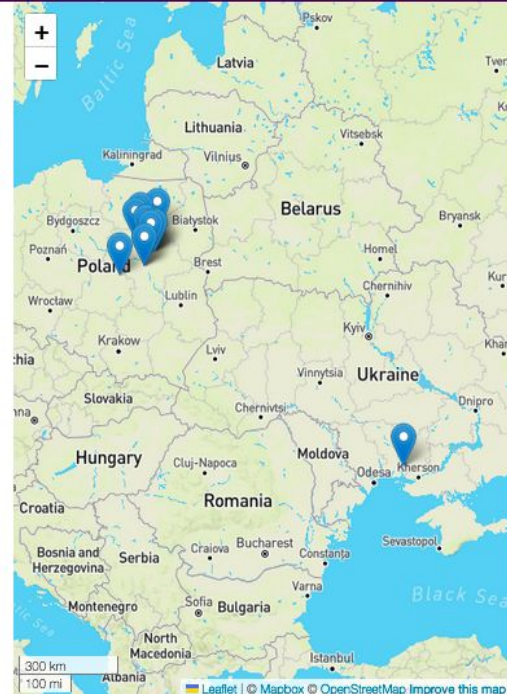
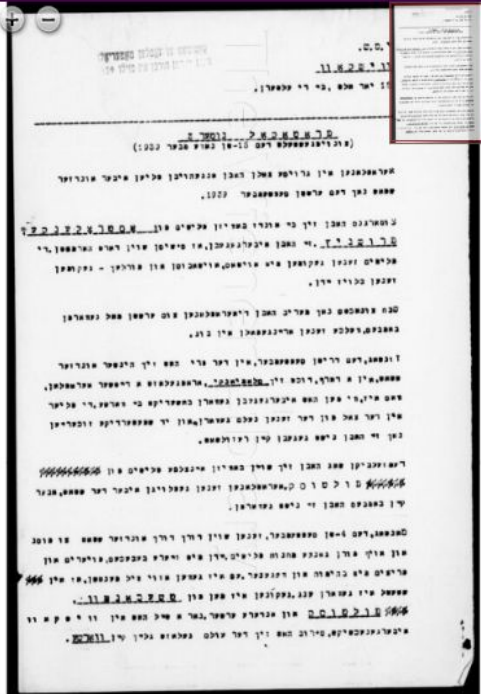
(צונויפגעשטעלט דעם 15טן נאָוועמבער 1939)

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

צומאָרגנס האָבן זיך בײַ אונדז באַוווּזן פּליטים פֿון אַסטראַלענקע, פּרוּשניץ. זײ האָבן איבערגעגעבן, אַז ס'שײַנט שוין דאָרט האַרמאַטן. די פּליטים זענען געקומען מיט אויטאָס, אויטאָבוסן און פֿורלעך – געקומען זענען בלויז יידן.

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

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



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

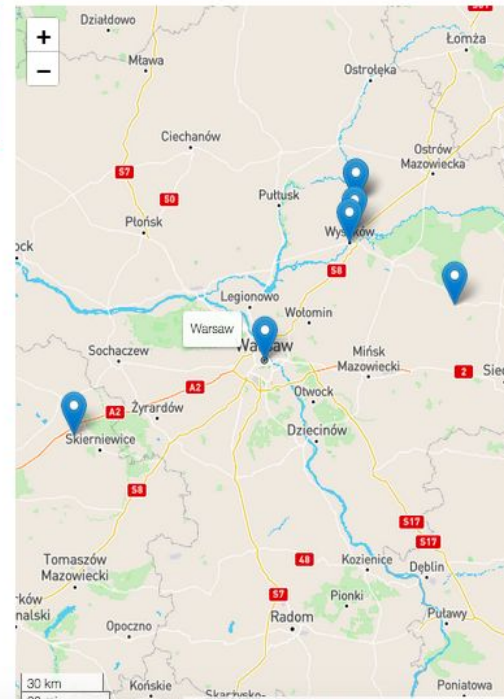
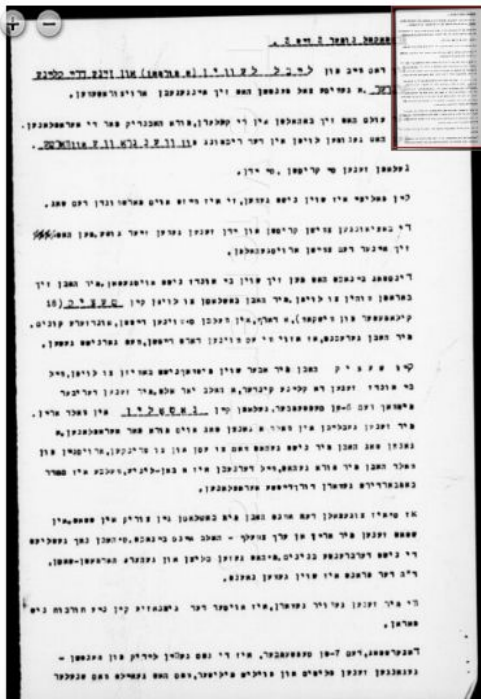
...undressed for the night. We

...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**.

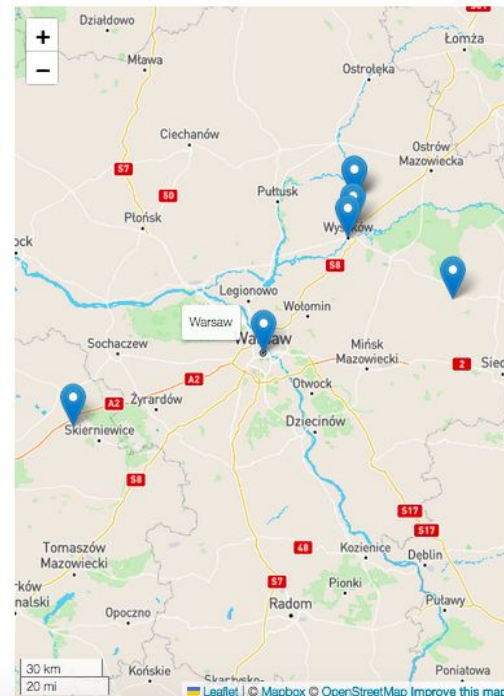
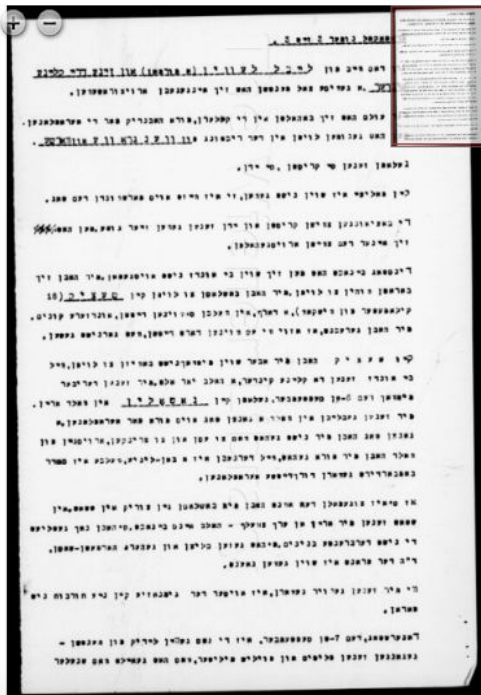
Both Christians and Jews were running away.

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 **Sieczchy** (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 **Sieczchy**, 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

- ❑ The EHRI online editions are a **great source for various NLP and DH tasks**
- ❑ For more information
 - ❑ **Break-out session** of the afternoon to work with TEI Publisher and discover the annotation tool
 - ❑ **EHRI Conference in June** where we will present the work we have done on the EHRI editions but in a more “pipeline-oriented” point of view

RESOURCES

Resources

- ❑ Dataset EHRI: <https://github.com/FloChiff/ehri-dataset/tree/main>
- ❑ Presentation of the EHRI dataset: <https://flochiff.github.io/phd/dataset/ehri/dataset.html>
- ❑ ODD EHRI: <https://gitlab.inria.fr/dh-projects/workflow-ehri/-/tree/main/ODD>
- ❑ TEI Publisher: <https://teipublisher.com/index.html>
- ❑ e-editiones: <https://www.e-editiones.org/>

Thank you for your attention

Any questions ?

Contact :
floriane.chiffoleau[at]inria.fr
hugo.scheithauer[at]inria.fr